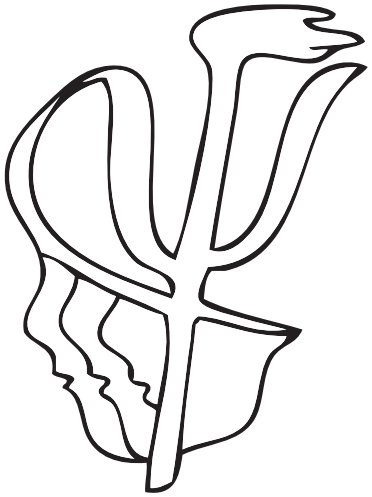


Words, Vectors, and Feelings: Advancing Psychological Emotion Research Through Natural Language Processing

MSc Hubert Wojciech Plisiecki
Graduate School for Social Reseach
Institute of Psychology
Polish Academy of Sciences

Doctoral Thesis prepared under the supervision of
dr hab. Artur Pokropek
and auxillliary supervision of dr. Grzegorz Pochwatko

Warsaw, Poland, 2025

Funding information: The articles published as a part of this thesis were supported by the SONATA BIS grant from the Polish National Science Center "Research Lab for the Digital Social Sciences" (2020/38/E/HS6/00302) led by dr hab. Artur Pokropek

Acknowledgements

Working on this dissertation for the last four years was one of the most formative experiences of my life. I am grateful to my supervisor, Professor Artur Pokropek, whose leadership combines steady guidance with a motivating stance. His calm and collected style of management, coupled with constant encouragement to do one's best, fostered both scientific rigor and a stress-free, collegial atmosphere. His advice was both thoughtful and constructive. In guiding me, he has not only been an exemplary research mentor but also the kind of principal investigator I aspire to become.

I want to also thank my family, who were deeply invested in my journey and encouraged me along the way. My parents, whose support allowed me to complete this journey without unnecessary stress — my mother, who I could always talk to about my troubles, and my father, who always offered an encouraging sense of direction. My grandparents from both sides, who rooted for me and cheered me on every step of the way - especially my grandfather, Jerzy Plisiecki, who nudged me towards becoming a scientist back when I was very young. Also my aunt, who was always curious about my work.

I also send my thanks to my friends, who inspired me and offered a great deal of positive emotions throughout this journey. I thank them for all the conversations we have had and all the things we did together, including but not limited to the publication of some of the articles included in this dissertation, and perhaps more importantly, the creation of the Polish Society for Open Science, an organization that helped me channel my scientific mission into something bigger.

Finally, my beloved girlfriend Kalina, who has been with me nearly since the beginning of this journey and who was both understanding of my engagement in my work and supportive of my goals on a daily basis.

To all of you, a heartfelt thank you.

Contents

Acknowledgements **2**

Table of Contents **3**

Abstract **5**

Streszczenie **7**

General Introduction **9**

Brief History of Text Analysis in Psychological Science 10

An Overview of the Main Tools in Natural Language Processing 12

Contemporary Research in Psychological Text Analysis 15

Natural Language Processing in the Study of Emotions 17

Scientific Articles Included in the Dissertation 19

Article Number One - Emotion Topology: Extracting Fundamental Components of Emotions from Text Using Word Embeddings 20

Article Number Two - Extrapolation of Affective Norms Using Transformer-Based Neural Networks and Its Application to Experimental Stimuli Selection 21

Article Number Three - High Risk of Political Bias in Black Box Emotion Inference Models 22

Article Number Four - Eradicating Social Biases in Sentiment Analysis using Semantic Blinding and Semantic Propagation Graph Neural Networks . . 23

General Discussion 24

Article Number One **33**

Article Number Two **45**

Article Number Three **61**

Article Number Four **72**

Abstract

To date, psychological research that relies on natural language processing (NLP) has borrowed largely unmodified tools from computer science, often overlooking psychological nuances that are crucial to realizing their full potential for the field. This dissertation aims to correct this pitfall by creating new NLP methods tailored to psychological inquiry, addressing domain-specific concerns such as ecological validity, construct bias, and measurement fidelity, thereby opening new avenues for theory, application, and measurement in affective science.

The first study shows that the fundamental structure of emotion can be recovered directly from everyday language. Using 58 000 Reddit posts tagged with 28 emotions, it builds unsupervised numerical emotion representations and submits them to principal-component analysis. By demonstrating that the circumplex model emerges organically from natural discourse, the study tackles the domain-specific concern of whether laboratory-derived structures generalise to "in-the-wild" expression and provides a blueprint for theorists to mine large corpora for latent psychological dimensions, expanding both theorical and exploratory potential of text-based studies.

The second study extends affective-norm databases with transformer models fine-tuned on existing ratings. The resulting models predict a range of word-level affective indices for unseen words with correlations up to $r = 0.95$ on the test set in English and comparably high scores in five other languages. In doing so, it addresses the practical bottleneck of sparse normative data - a core domain-specific limitation - and introduces automated extrapolation and stimuli-descent procedures that enable researchers to design better-controlled experiments.

The third study audits social bias in a novel Polish sentiment model. Regression analyses reveal that a politician's gender and party affiliation account for up to 66% of the variance in predicted valence—an effect rooted in the annotators' personal biases. By empirically exposing how supervised NLP pipelines can silently propagate psychologically relevant social and political biases, the work opens new methodological avenues for diagnosing and correcting such biases, thereby safeguarding the integrity of psychological measurement and scientific conclusions.

The fourth study introduces the Semantic Propagation Graph Neural Network, an explainable sentiment model designed to retain accuracy while curbing social bias. By "blinding" the network to word identity and letting the emotional information of singular words flow only through syntactic links, the model approaches transformer-level performance and significantly reduces prediction bias, demonstrating that high accuracy and fairness can coexist. This architecture directly addresses the domain-specific concern of

balancing validity with ethical neutrality and offers a potential method of tracking how emotional meaning propagates through syntax, broadening future applications of graph-based modelling in psychology.

Collectively, these four studies contribute a vetted battery of exploratory, predictive, diagnostic, and corrective methods that enable psychologists to investigate mind and behaviour through text while guarding against the methodological and ethical pitfalls of off-the-shelf NLP tools. Together they respond to pressing domain-specific challenges - ecological validity, normative coverage, bias mitigation, and interpretability - while opening fresh trajectories for future psychological theory, experimental application, and large-scale, text-based measurement.

***Keywords*:** Emotion Research, Natural Language Processing, Affective Norm Extrapolation, Social Bias in Sentiment Analysis, Psychological Methods

Streszczenie

Dotychczasowe badania psychologiczne wykorzystujące przetwarzanie języka naturalnego (NLP) w dużej mierze zapożyczały gotowe narzędzia z informatyki, często pomijając psychologiczne niuanse, które są kluczowe dla pełnego wykorzystania ich potencjału w tej dziedzinie. Niniejsza dysertacja ma na celu skorygowanie tego niedopatrzenia, tworząc nowe metody NLP dostosowane do potrzeb psychologii i uwzględniające specyficzne dla niej kwestie, takie jak trafność ekologiczna, stronniczość konstruktu i wierność pomiaru, otwierając tym samym nowe możliwości teoretyczne, praktyczne i pomiarowe w psychologii emocji.

Pierwsze badanie pokazuje, że podstawową strukturę emocji można odtworzyć bezpośrednio z codziennego języka. Korzystając z 58 000 postów z Reddita oznaczonych 28 emocjami, stworzono nienadzorowane, numeryczne reprezentacje emocji i poddano je analizie głównych komponentów składowych. Dwa pierwsze komponenty odtwarzają klasyczne wymiary walencji i pobudzenia, co dowodzi, że kołowy model emocji Russella wyłania się organicznie z naturalnego dyskursu; badanie to dotyka problemu czy struktury wywiedzione z laboratorium generalizują się na język „in-the-wild", dostarczając jednocześnie metody do wydobywania ukrytych wymiarów psychologicznych z dużych korpusów i poszerzając zarówno potencjał teoretyczny, jak i eksploracyjny psychologicznych badań z użyciem tekstu.

Drugie badanie rozszerza bazy norm afektywnych przy użyciu modeli transformerów dostrojonych na podstawie istniejących ocen. Stworzone modele przewidują szereg wskaźników afektywnych na poziomie słów dla niewidzianych wcześniej wyrazów, osiągając korelacje do r = 0,95 z zestawem testowym w języku angielskim i porównywalnie wysokie wyniki w pięciu innych językach. Badanie to rozwiązuje praktyczny problem niedoboru danych normatywnych – częsty problem dziedziny – i wprowadza zautomatyzowane procedury ekstrapolacji oraz algorytm „stimuli-descent", które pozwalają badaczom dobierać semantycznie dopasowane pary słów różniące się wyłącznie na docelowych wymiarach, ułatwiając projektowanie lepiej kontrolowanych eksperymentów.

Trzecie badanie audytuje uprzedzenia społeczne w nowym polskim modelu sentymentu. Analizy regresyjne ujawniają, że płeć polityka i przynależność partyjna wyjaśniają do 66% wariancji przewidywanej walencji - efekt zakorzeniony w osobistych uprzedzeniach anotatorów. Empiryczne ukazanie, w jaki sposób nadzorowane modele NLP mogą propagować psychologicznie ważne uprzedzenia społeczne i polityczne, otwiera nowe ścieżki metodologiczne do ich diagnozowania i korygowania, ulepszając rzetelność pomiaru psychologicznego oraz wniosków naukowych.

Czwarte badanie przedstawia Semantic Propagation Graph Neural Network – wyjaś-

nialny model sentymentu zaprojektowany tak, aby zachować wysoką dokładność predykcji przy jednoczesnym ograniczeniu propagowanych uprzedzeń społecznych. „Oślepiając" sieć na tożsamość słów i pozwalając, by informacja emocjonalna pojedynczych wyrazów rozprzestrzeniała się wyłącznie przez powiązania syntaktyczne, model osiąga wydajność zbliżoną do transformerów, a jednocześnie znacząco redukuje uprzedzenia w predykcjach, dowodząc, że wysoka dokładność i bezstronność nie wykluczają się nawzajem. Architektura ta bezpośrednio odpowiada na problem równoważenia trafności i neutralności etycznej oraz oferuje narzędzie do śledzenia, jak znaczenie emocjonalne rozchodzi się w strukturze składniowej, poszerzając przyszłe możliwości zastosowania modelowania grafowego w psychologii.

Łącznie te cztery badania dostarczają zweryfikowanego zestawu metod eksploracyjnych, predykcyjnych, diagnostycznych i korekcyjnych, które umożliwiają psychologom badanie umysłu i zachowania poprzez tekst, jednocześnie unikając metodologicznych i etycznych pułapek gotowych narzędzi NLP. Razem odpowiadają na palące wyzwania specyficzne dla dyscypliny – trafność ekologiczna, pokrycie normatywne, ograniczanie stronniczości i interpretowalność – otwierając zarazem nowe ścieżki dla przyszłej teorii psychologicznej, zastosowań eksperymentalnych oraz szeroko zakrojonych, pomiarów bazujących na tekście.

General Introduction

This dissertation critically evaluates and expands the methodological toolkit for studying emotions in text by introducing advanced machine learning (ML) and natural language processing (NLP) methods tailored specifically for psychological research. Whereas previous work often relied on off-the-shelf sentiment-analysis tools, the studies presented here develop custom approaches that tackle three key psychological challenges: (1) mitigating biases introduced by annotators, (2) extrapolating affective word norms for more comprehensive coverage of emotional vocabulary, and (3) creating exploratory tools that reveal how psychological phenomena are reflected in everyday language. These innovations capitalize on the ecological validity of unprompted, naturalistic text—an important step beyond traditional self-report measures. Thanks to the use of unsupervised methods, they let the data "speak for itself," thereby reducing dependence on pre-existing theories when exploring emotional constructs. Further, because text is widely available online, these approaches help overcome issues of small sample sizes and low statistical power that often plague psychological studies. By blending techniques such as word embeddings, transformer-based architectures, and graph neural networks, this thesis offers a more data-driven mode of discovery, crafting new tools of scientific inquiry, refining the way we extrapolate affective norms, and mitigating the social biases that can distort sentiment analysis results. Ultimately, this work is a step towards a greater adoption of ML and NLP methods within the psychological paradigm.

NLP methods differ in many ways from the tools that psychologists are classically accustomed to. The standard distinction of quantitative vs. qualitative approaches breaks down when applied to them. They constitute a type of middle-step between the two, offering quantification while at the same time often requiring qualitative interpretation, and benefiting from manual qualitative validation. Given the often-wide mismatch between the paradigms adopted by the proponents of quantitative vs. qualitative research, NLP methods may offer a middle ground between the two, fostering holistic approaches. The bright promises are, however, well counterbalanced by the limitations of these novel tools. Many NLP models struggle with explainability, functioning as black boxes which can be too complex to effectively probe for the details of what exact computations they conduct underneath – a problem to a large extent unfamiliar to psychologists whose analytical toolbox often stops at the level of Structural Equation Modeling. Beyond that, text data also comes with a high inherent portion of noise, where only a part of the information it conveys is relevant to what psychologists want to study – something possibly unheard of to people working with carefully crafted psychometric tools. These idiosyncrasies, if NLP methods are to be widely adopted in psychology, researchers have to learn how to deal with, and I hope that my dissertation will also constitute a small step in that direction.

Given that NLP methods can be viewed as novel due to the significant developments that took place in this field over the last decade and a half, some psychologists might look at them as foreign to the psychological paradigm. In the introduction to this dissertation, I will try to argue to the contrary. Psychological science has long been interested in the study of language, and its endeavors aimed at analyzing it computationally have a long history, which I intend to summarize on the following pages in order to better contextualize my research. This brief historical insert will be followed by an overview of contemporary NLP methods as well as a section devoted to current psychological research conducted with their use. Finally, a section outlining NLP based psychological research focusing on emotions will precede a summary of the four articles constituting this thesis. The final discussion will integrate the findings of the articles included in this dissertation and attempt to situate the NLP methodology within the wider psychological paradigm, considering its advantages as well as limitations and tracing new research directions.

Brief History of Text Analysis in Psychological Science

While many of the significant advances in the field of Natural Language Processing are very recent and haven't yet been widely adopted by mainstream psychology (e.g., Vaswani et al., 2017), the idea of quantifying language in psychological science is relatively old (Allport, 1942; Allport et al., 1953; Baldwin, 1942). For example, as early as in 1942 Baldwin used a corpus of 301 personal letters to create a matrix of co-occurrences of different topical themes (e.g. money, health) and attitudes (e.g., favorable, unfavorable) by hand with the aim to identify the main "personal structures" of the author (Baldwin, 1942). The main problem with this and other similar studies was that the quantification of text had to be done by hand, leading to long hours of counting words and expressions. This was no longer the case by 1966 when a team at the Massachusetts Institute of Technology published an algorithm that performed the calculations automatically (Stone, 1966). This tool, named The General Inquirer (GI) used large lexicons with words annotated with regards to their emotional load, expressiveness, language type (e.g. Academic, Economic etc.), object types (e.g. food, tool etc.) and many other taxonomies to quantify the context of natural language by computing their frequencies. While the tool was dedicated for use by a broader range of social scientists, the original publication included whole chapters dedicated to personality psychology including an analysis of the structure of personality from letters; clinical psychology with inquiries into psychotic language and therapeutic transcripts; and social psychology with an analysis of suicide notes, reports written by African field work volunteers, and inquiries into the nature of the self-perceived identity of college students. The software itself, however, wasn't widely adopted, owing in large part to the difficulty of use. A review of the GI concludes "Unless the investigator is willing to make a long-term commitment to research with computers, the use of the

General Inquirer system is not to be recommended at this time." (Psathas, 1969, p. 174).

The next big breakthrough in computer-assisted text analysis was inspired by the research of Walter Weintraub, who documented that people with depression consistently use more first-person singular pronouns such as I, and me than nondepressed people (Weintraub, 1981, 1989). This idea of analyzing the frequencies of function words to infer psychological phenomena gained traction over the next 20 years when a team of scientists developed a well-known psychological tool called Linguistic Inquiry and Word Count (LIWC) (J. W. Pennebaker, 2001; Tausczik & Pennebaker, 2010). LIWC worked in a very similar way to its predecessor, the GI, by counting words from preassembled dictionaries. However, the process of dictionary creation, and their focus changed over time. Currently, after several revisions of the original software LIWC (Boyd et al., 2022; J. Pennebaker et al., 2007; J. W. Pennebaker, 2001; J. W. Pennebaker et al., 2015), the newest version provides metrics for over 100 categories, split roughly into summary variables (e.g., word count, analytic thinking related words, emotional tone, words per sentence), linguistic dimensions (e.g., personal pronouns, determiners, prepositions, negations), psychological processes (e.g., drives, affect, prosocial behavior, friends related words), and the expanded dictionary (e.g., words related to politics, work, economics, mental health, need fulfillment, time orientation). Each of these metrics has been assembled in a multiple-stage process from manual dictionary assembly to internal consistency tests on test corpora, providing a sense of their psychometric reliability (Boyd et al., 2022).

The LIWC psychometric tool gave rise to a multitude of studies, the most popular LIWC paper being cited more than seven thousand times (Boyd & Schwartz, 2021; Tausczik & Pennebaker, 2010). The authors of the tool explained this rise in popularity by two factors. One of them was the advent of personal computing, and the other the validity of their approach – the focus on earlier ignored function words such as "the", "he", "is" etc. Indeed, previous studies have failed to achieve satisfactory performance, with regards to manual coders ground truth, using General Inquirer based content words (words carrying lexical meaning such as nouns, verbs, adjectives, and most adverbs) (Smith, 1968). A potentially related setback of content word-based metrics is their relative sparsity in text, when compared to function word-based ones. While only some texts might contain words related to content-based categories such as work and economics, nearly all texts will contain a high degree of function words, providing greater granularity of analysis. LIWC therefore offered a performant, easy to use research tool with relatively high degree of objectivity – especially when compared to manual scoring.

Despite LIWC's popularity, from the perspective of currently available tools for language analysis, its dictionary-based approach might seem a bit dated. For example, when it comes to emotion detection, machine learning based systems vastly outperform purely

dictionary-based approaches (Widmann & Wich, 2023). The creators of LIWC were aware of this shift, having pointed out themselves that "LIWC represents only a transitional text analysis program in the shift from traditional language analysis to a new era of language analysis." (Tausczik & Pennebaker, 2010, p. 38). The decade and a half between 2010 and 2025 (the current date) brought about many significant inventions in the field of text analysis, now more often referred to as natural language processing. Starting from word embeddings which allowed researchers to numerically encode words and documents in a manner that allowed for high classification performance as well as the computation of similarity scores, going through convolutional neural networks, and the transformer revolution initiated by the famous paper "Attention Is All You Need" by Vaswani and associates 2017, accelerated by the publication of the ChatGPT chatbot by OpenAI in 2022 which began the Large Language Models era. This decade and the subsequent years were nothing short of revolutionary for the text analysis and its repercussions are not yet fully realized with a constant output of new publications discussing their responsible use within the academia (e.g., Sohail & Zhang, 2025). The spirit of the moment was well encapsulated in a recent review of the field of NLP in psychology in the following words "The secrets of language are being unlocked in new and exciting ways, and we sit at the cusp of an absolutely revolutionary shift in how we conduct social scientific research." (Boyd & Schwartz, 2021, p. 33).

An Overview of the Main Tools in Natural Language Processing

Before examining contemporary research in psychological text analysis, it is essential to understand the fundamental tools that have enabled the shift from traditional linguistic quantification to modern computational approaches. One of the most consequential developments in the last decade has been the increase in our ability to encode language into continuous, multidimensional numerical representations. While methods that achieved the same goal such as Latent Semantic Analysis (Landauer & Dumais, 1997), existed long before, their application was limited due to computational constraints. Contemporary methods, unlike earlier lexicon-based methods, which relied on predefined word lists and categorical mappings, capture latent semantic relationships between words, allowing for a more nuanced representation of meaning. Such representations enable researchers to compute semantic similarity, cluster texts based on thematic content and predict psychological attributes. They also form the foundation for more complex models, including deep learning architectures that now drive state-of-the-art NLP research. However, these methods did not emerge overnight. The first breakthroughs in numerical representations of language came with word embeddings – most notably Word2Vec (Mikolov et al., 2013) – ushering in a new era of text analysis by transforming words into dense vector spaces that encode meaning beyond simple frequency counts.

The Word2Vec algorithm (Mikolov et al., 2013) is based on the distributional hypothesis, which states that words appearing in similar contexts tend to have related meanings (Firth, 1957; Harris, 1954). For example, in a large corpus, the word *lemon* is more likely to appear near other citruses like *orange* and *grapefruit* than near unrelated words like *chair*. This statistical pattern encodes semantic relationships, but directly storing co-occurrence frequencies in a large matrix is impractical due to memory constraints and sparsity. To overcome this, Word2Vec learns dense vector representations of words by training a neural network to predict words based on their context. This model can then be used to input specific words and output their numerical representations, where words with similar meanings are positioned closer together in the embedding space (i.e. their vectors are similar). This property allows Word2Vec embeddings to capture not only direct word associations but also more abstract relationships, such as analogies ($king - man + woman \approx queen$) (Church, 2017). Furthermore, the technique can be extended to create embeddings for longer texts such as whole documents (Le & Mikolov, 2014). The effectiveness of the Word2Vec model in encoding useful semantic information is evidenced by its widespread adoption for various tasks that require a numerical representation of text, including sentiment analysis, topic modeling, information retrieval, and psychological text analysis (Johnson et al., 2024).

While Word2Vec is an ingenious method of creating useful numerical representations of natural language, much like LIWC, it constituted only a transitional stage in the development of natural language processing tools. Its main setback was the inability to model how words change their meanings based on the context in which they are used—this includes negations, polysemy, and nuanced shifts in meaning that arise from syntactic or semantic dependencies within a sentence (Widmann & Wich, 2023). For example, Word2Vec assigns the same vector representation to a word like bank regardless of whether it refers to a financial institution or the side of a river, and it fails to capture how negations like not happy differ from happy in sentiment. These limitations highlighted the need for more sophisticated models that could dynamically adjust word meanings based on their surroundings. It led to the development of transformer-based models, such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2018), which introduced mechanisms capable of analyzing words in relation to all other words in a sentence, paragraph, or even an entire document.

The key breakthrough that enabled these models was the introduction of attention mechanisms, particularly self-attention, which allowed NLP models to dynamically assess the importance of each word in a sequence relative to all others (Vaswani et al., 2017). Unlike previous approaches, which relied on fixed-length context windows or strictly linear word relationships, attention allows a model to recognize dependencies between words across long passages of text, capturing meaning that might be spread across multiple sentences.

This makes it possible for a model to understand that, for example, in the sentence *She didn't like the movie because it was too slow*, the word *slow* is what explains *didn't like*, ensuring a more contextual interpretation. The transformer architecture introduced by Vaswani et al. 2017 leveraged multiple layers of stacked attention mechanisms, allowing each layer to refine its understanding of word relationships at different levels of abstraction. This stacking process became the foundation of modern NLP models, with GPT-2, GPT-3, and GPT-4 progressively increasing the number of attention layers and training data, leading to a dramatic leap in language modeling capabilities. The culmination of this approach is seen in Large Language Models (LLMs) like GPT-4 and its successors, where massive datasets and deep architectures enable a model to generate, summarize, and analyze text with unprecedented accuracy and coherence (Kocoń et al., 2023). These advancements represent not just an improvement in computational text analysis but a fundamental shift in how artificial intelligence processes and understands human language.

The power of transformer-based models comes not just from their ability to model long-range dependencies but also from their architecture, which relies on encoder-decoder structures trained in parallel. In broad terms, an encoder processes input text, converting it into a compact numerical representation (embedding), while a decoder uses this information to generate output text, whether in translation, summarization, or text completion (Kocoń et al., 2023). During training, these two components are optimized simultaneously, learning to predict masked words or reconstruct input sequences based on context. Crucially, encoders – when trained independently – can be used in much the same way as Word2Vec, extracting fixed-length numerical embeddings of words, sentences, or entire documents. These embeddings can then serve as inputs for a range of downstream tasks, including clustering, semantic similarity measurement, and regression-based prediction using additional layers of weight matrices, often referred to as regression heads (Widmann & Wich, 2023). This approach allows researchers to train models that predict psychological attributes, detect emotions, and analyze text with greater quantitative rigor.

While this short overview has highlighted key milestones in NLP over the last decade and a half – from early Latent Semantic Analysis approaches, to word embeddings, attention, and large language models – it is by no means exhaustive. Other significant advancements, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their improved variants like long short-term memory networks (LSTMs) and gated recurrent units (GRUs), have also played a role in the evolution of natural language processing as we see it today (see Perumal et al., 2024, for a comprehensive review). However, rather than providing a comprehensive historical account, this overview is intended mainly to establish a computational methodological context for the research

presented in this dissertation by focusing on the most influential developments in contemporary NLP.

Contemporary Research in Psychological Text Analysis

While some researchers have focused on advancing text-processing methods, others have started using them to investigate psychological phenomena. For example, word embeddings have been shown to be extremely useful in identifying cultural biases in text. By assembling lists of words related to the two poles of the bias dimension that researchers want to study (e.g., men vs. women) they can then compare the distance of different words (e.g., specific occupations) to those poles probing the degree of bias in the meaning encoded by the model (see Durrheim et al., 2023, for a comprehensive review). One study showed that gender bias tested this way with regards to such words as *nurse*, *librarian*, and *housekeeper* had a significant correlation (r=0.5) with the percentage of women working in these occupations (Garg et al., 2018). A different study showcased the convergent validity of word embedding based measures for probing intergroup attitudes and biases by demonstrating that they capture intergroup associations in a manner consistent with Implicit Association Test (IAT) results (Kurdi et al., 2019). More generally, word embedding based word similarity measures have been shown to reliably reflect human ratings of word associations (Hofmann et al., 2018). They have also been successfully applied to the prediction of multiple psychological constructs from text. Some notable avenues of research here include the prediction of psychological disorders based on text from social media (Couto et al., 2025), personality dimensions (Alsini et al., 2024), brain activations (Oota et al., 2018), as well as high-level human judgments across diverse behavioral domains (Richie et al., 2019), and emotions (Widmann & Wich, 2023).

The development of transformer-based models and their derivatives – LLMs brought about not only a rise in the prediction accuracy of psychological constructs (Widmann & Wich, 2023), but also applications related to text generation. For example, transformers have been used to generate novel, psychometrically valid items for psychological questionnaires (Hommel et al., 2022). New research explores the possibility of integration of LLMs in psychotherapeutic work in a way that will minimize the obvious dangers associated with leaving the responsibility for the state of a psychotherapeutic client within the hands of a computer algorithm (Hodson & Williamson, 2024; Hommel et al., 2022). Other studies have focused on estimating to what extent the text generated by LLMs corresponds to that generated by a human on a range of different cognitive tasks. LLMs have been successful in completing various theory of mind tasks (Kosinski, 2024), emotion intelligence tests (X. Wang et al., 2023), and various others cognitive tasks (Binz & Schulz, 2023).

This seeming resemblance to the performance of humans prompted research projects trying to use the transformer architecture to create computational models of human cognition by finetuning an LLM on data from various psychological experiments (Binz et al., 2024). While the preliminary results showed that such a model generalizes well to unseen cognitive tasks and that its internal numerical representations are to a degree aligned to aggregated neural activity of human participants taking part in the same experiments (R2 of more than 0.1 with variable results depending on the exact layer from which the representations were extracted), these results should be viewed with a significant degree of caution. The main optimization criterion for these models depends on being able to correctly guess the next word in a line of text, not to realistically reconstruct the mechanisms through which the human brain produces language (Sobieszek & Price, 2022). This means that while LLMs produce language that is hard to distinguish from that of a human, it might very well be doing this using completely different computational structures than those utilized by the human brain. At the same time, even the unlikely prospect of being able to create authentic artificial models of human cognition brings with it immense research potential, at the same time motivating the valuable study of the differences and similarities between LLMs and humans.

While NLP tools offer exciting opportunities, enabling the modeling of text in ways that previous generations of psychologists could only dream of, this power comes at a cost. The task of numerically encoding text while preserving its meaning is computationally complex. Our brains also perform this task, yet we still lack a clear understanding of how human cognition processes and produces language (Roland, 2023). Likewise, we struggle to fully comprehend the numerical transformations that machine learning models apply to textual data. This opacity is known as the black box problem—we cannot fully trace or interpret every computation within a machine learning model, limiting the potential for forming explanatory theories. However, techniques of making machine learning models more transparent are constantly evolving as researchers are experimenting with different explainable machine learning architectures (Burkart & Huber, 2021). Still, even with limited explainability, this limitation does not preclude their usefulness in psychology. Similarly to how earlier tools like LIWC provided interpretable linguistic metrics despite performance problems, NLP-based models can be leveraged for psychological research by focusing on the outputs they generate rather than the underlying computations, however only after the algorithms that produced them have been appropriately vetted for validity and reliability. Where necessary, the black box nature of these models should be considered as a limitation during research.

Natural Language Processing in the Study of Emotions

One particularly salient field that intersects psychology and natural language processing is the study of human emotions as they are expressed in language. Researchers in this domain usually have three approaches to choose from: 1) lexicon-based emotion analysis, where words from text are checked against a lexicon that has predefined associations with emotional categories or dimensions, allowing for a rule-based classification of emotional content (Hills et al., 2019); 2) machine learning approaches, where models are being trained to predict emotion based on annotated datasets, learning statistical patterns in textual features to generalize emotion recognition across different contexts (Widmann & Wich, 2023); 3) large language models, where texts are inputted into LLMs to ask them about the emotions that are being expressed, utilizing their pre-trained knowledge to infer and describe emotional content (Kocoń et al., 2023). The main qualities distinguishing these three approaches are accuracy of emotion prediction, ease of application, and the degree of explainability. While the latter two are an undeniable asset of lexicon methods, machine learning models outcompete them significantly in terms of precision. Large Language Models, on the other hand, approach the accuracy levels of their less complex machine learning counterparts, and are relatively easy to use with the help of available API's (e.g. OpenAI API), but due to their increased complexity can be viewed as even more black box in their nature (Plisiecki et al., 2024).

These methods were used across various studies to analyze emotional indices of text. For example, lexicon approaches have been used to study shifts in historical wellbeing by analyzing the valence of millions of books published across various countries from the beginning of the 18th century (Hills et al., 2019). While the valence of words written in books does not directly translate to the wellbeing of people living in the countries where these books were published, researchers were able to show that their valence metric significantly correlated with the available country well-being data. A different study analyzed the variation in sentiment expressed across the world on social media during the outbreak of the COVID-19 pandemic using a machine learning based model. They were able to show that the outbreak resulted in a steep decline of expressed sentiment across different geographical areas (J. Wang et al., 2022). Text-inferred emotions have also been used to identify suicide risk from text messages showing that prior to suicide attempts people expressed an increase in expressed anger and a lowering of the expression of positive emotions (Glenn et al., 2020). Yet another study used emotion lexicons to analyze teachers' enthusiasm during lessons and showed that teacher's self-reported enthusiasm is significantly related to the enthusiasm that they have expressed verbally during lessons (Frenzel et al., 2025)).

Here it is appropriate to outline what is meant by emotional indices of text, as measuring

the expressed or conveyed sentiment in text is by no means equivalent to a self-report on the emotional state of the author. A link nonetheless exists, as shown by studies that experimentally evoked emotions in participants and then analyzed the texts written under the influence of those emotions using LIWC (Kahn et al., 2007), with more complex sentiment analysis systems most likely achieving higher emotional congruence with the actual internal state of the author. However, it is important to note that texts such as essays written during an experimental study might be more revealing of the internal state of the author, as compared to a post on the internet, because the participant is directly incentivized to convey their emotions, while the internet user freely chooses which part, if any, of his internal experience to express. Even though assuming that every internet user is a Machiavellian trickster is a clear exaggeration, researchers have to consider the social desirability bias, along with other information filters, as being more salient in real-life text datasets. Practically this usually means that real-life text-data is noisier than its laboratory counterpart (J. W. Pennebaker, 2022). Another important variable that influences the emotional indices of text is the perspective from which the text is viewed. Research showed that texts annotated from the author perspective i.e. "what emotion is expressed by the author", receive significantly different ratings than those annotated from the perspective of the user i.e. "what emotion is conveyed" or "how do you feel after reading" (Buechel & Hahn, 2017). While this finding might seem obvious, it has important consequences for how emotional annotations are collected, as research authors also conclude that authors perspectives achieve higher inter-rater reliability. An interesting question from the psychological perspective here is to what extent can people shed their individual perspective and produce emotion ratings that are not biased by their own attitudes.

While the emotional annotation of text is usually conducted with the goal of creating a training dataset for a machine learning model, words annotated with regards to affective dimensions have been widely used as stimuli in cognitive research. Lexicons such as those created by Imbir 2016, or Warriner and colleagues 2013 contain thousands of words annotated with regards to dimensions such as valence, arousal, dominance, and many others. These databases, also known as norms, were used repetitively to prime subjects with words of specific emotional load (K. Imbir et al., 2023; Scerrati et al., 2022), to explore the associations between affective dimensions (Warriner, 2014), and as lexicon bases for sentiment analysis (Ribeiro et al., 2016) and therefore are of great value to scientific research. They are, however, limited by the size of their lexicon, which contains only a sample of all the words available in a given language. The task of synthetically extending the available lexicons has been attempted multiple times, with each consecutive approach achieving slightly better results (Recchia & Louwerse, 2015; Snefjella & Blank, 2020). It has not, however, been attempted before using transformer-based models, as it

is done in the second article presented in the current dissertation, perhaps due to their emphasis on long-form text input.

So far, the presented research focused on producing emotion ratings for different texts and using these in downstream analysis. This method, while very useful, is clearly limited with regard to what scientific questions it can answer. In it, vectors that represent semantic meaning of texts, are distilled to produce ratings for specific emotions. The same vectors however can also be used as encodings of emotional expression, and analyzed in their more granular form, without direct distillation. For example, Calder 2001, while not working directly with text, took pictures of facial emotion expressions and represented them as vectors of pixels that composed them. After applying Principal Component Analysis (PCA) to them, he was able to reconstruct the structure of the circumplex model of affect (Russell, 1980). A parallel example used the similarity of the vectors created based on the text of items in a big five personality inventory to show that it reflects the associations between personality factors in human subjects (Casella et al., 2024). This type of analysis, concerned more with the semantic content of the texts than with the distilled emotional value is further developed in study number one of the current dissertation, and offers a new avenue for the study of emotions in text.

Scientific Articles Included in the Dissertation

The papers that constitute this thesis introduce advanced machine learning methods to the field of psychology of emotions. While to date, the previous studies have mostly used ready-made tools to conduct their research, this research creates custom methods for use specifically in psychological research. This uniquely predisposes the work presented in this thesis to tackle challenges that might not be noticed or paid attention to when looked at from a different perspective but are important from the psychological perspective, addressing domain-specific concerns such as ecological validity, construct bias, and measurement fidelity. These issues are tackled by exploring the problems of personal bias of the annotators seeping into machine learning models; using transformers to extrapolate affective word norms; and creating exploratory tools that can be used to analyze the reflections of psychological phenomena in text. The research questions guiding the studies presented in this dissertation are as follows:

1. Article number one: ”Emotion Topology: Extracting Fundamental Components of Emotions from Text Using Word Embeddings” – Can numerical representations of discrete emotions (e.g., happiness, anger) extracted from text be used to replicate previous studies exploring the fundamental structure of emotions?

2. Article number two: “Extrapolation of Affective Norms Using Transformer-Based

Neural Networks and Its Application to Experimental Stimuli Selection" – Can machine learning models help psycholinguists in the task of extending available affective norms for words as well as in applying them in psychological studies?

3. Article number three: "High Risk of Political Bias in Black Box Emotion Inference Models"– Are black box machine-learning based techniques of text emotion inference devoid of social biases which might interfere with drawing accurate scientific conclusions from the studies that employ them?

4. Article number four: "Eradicating Social Biases in Sentiment Analysis using Semantic Blinding and Semantic Propagation Graph Neural Networks" – Is it possible to design machine learning based sentiment analysis models that do not propagate the social biases of the annotators that labeled the corpus on which these models were trained on?

Ultimately, these four research questions tackle real scientific problems on the intersection of psychology and machine learning. Furthermore, they also effectively pave the way for future psychological research using this methodology by 1) creating new machine learning methods that can answer psychological research questions, 2) showing the effectiveness of advanced machine learning techniques in solving methodological problems related to psychological research, and 3) identifying and solving the problems related to measurement fidelity associated with the use of black box models. A detailed summary of each of the four studies follows.

Article Number One - Emotion Topology: Extracting Fundamental Components of Emotions from Text Using Word Embeddings

The first study explores whether the fundamental structure of emotions such as the valence and arousal as they are construed in the circumplex model (Russell, 1980) previously identified through questionnaires and controlled laboratory experiments, can be extracted directly from natural language using word embeddings (Plisiecki & Sobieszek, 2024). The study utilized the GoEmotions dataset, which consists of approximately 58,000 Reddit comments annotated with 28 distinct emotions by human raters. This dataset provided a rich foundation of naturalistic emotional expressions in text. Rather than relying on predefined theoretical structures, we employed an unsupervised machine learning approach using Doc2Vec algorithm to create numerical representations (emotion vectors) for each emotion category based on the semantic content of texts expressing those emotions.

These emotion vectors were then subjected to Principal Component Analysis (PCA) to identify the underlying dimensions along which emotion expressions naturally vary in language. The results demonstrate that the first four principal components extracted

from the text data aligned with the established dimensions from traditional emotion research. The first component clearly separated emotions along the valence dimension (positive-negative), with joy, admiration, and gratitude on one end, and disgust, fear, and sadness on the other. The second component corresponded to arousal, distinguishing high-arousal emotions like surprise and anger from low-arousal emotions like sadness and caring. The third component resembled the dominance dimension, separating high-dominance emotions (anger, annoyance) from low-dominance emotions (fear, confusion), but was significantly noisier than the first two. The fourth component was the hardest to interpret, having explained the least variance out of the four the patterns it revealed were very noisy, however it is possible that it identified the unpredictability dimension of affect. These findings were validated through correlation analysis with established emotion norms (showing significant correlation of $r = 0.31$ for valence), qualitative inspection of words scoring high and low on each dimension, t-SNE visualization showing clear clustering by valence, logistic regression confirming the alignment between the first component and sentiment, and by repeating the analysis on randomly drawn halves of the dataset.

The study showcases a new method of extracting vectors that relate to psychological phenomena. In terms of findings, while the study was exploratory, it revealed structures that were similar to classical research on the dimensionality of emotion while using texts that were written by humans in their "natural habitat". This finding can be seen as providing complementary evidence for existing emotion theories while employing a completely different methodological approach, however this argument would be further strengthened by a replication on a different dataset.

Article Number Two - Extrapolation of Affective Norms Using Transformer-Based Neural Networks and Its Application to Experimental Stimuli Selection

The second study addressed the challenge of extending affective norms databases through the use of machine learning techniques (Plisiecki & Sobieszek, 2023). As already mentioned, these databases, constituted by thousands of words annotated with regards to emotional dimensions like valence, arousal, dominance etc., are limited with respect to the words that they contain. To overcome this limitation, this study trained transformer-based neural networks to predict affective norms for novel words. By fine-tuning pre-trained transformer models (some specifically trained on emotion-recognition tasks), on available affective norms datasets the study created extrapolation networks capable of predicting multiple affective dimensions simultaneously. Our model achieved state-of-the-art results with correlations between predicted and human-rated values reaching $r = 0.95$ for valence, $r = 0.76$ for arousal, $r = 0.86$ for dominance, $r = 0.85$ for age of acquisition, and $r = 0.95$ for concreteness in English. This represented an improvement of approximately

$\Delta r = 0.1$ across metrics compared to previous methods. The model also performed excellently across other languages, including Polish, Spanish, Dutch, German, and French.

To examine the limitations and robustness of our approach we have employed targeted experiments to show that while the model performed well on most words, its accuracy decreased slightly (by about 11% on average) for words that deviated significantly from those in the training database. This finding highlights the importance of using extrapolated norms as heuristic tools rather than definitive measurements, particularly for uncommon words. Additionally, the study developed a "stimuli descent algorithm" – a novel method for selecting experimental stimuli that manipulates specific emotional dimensions while controlling others. This algorithm is able to provide semantically matched word pairs that differ primarily in the emotional dimension being studied, thereby reducing the risk that uncontrolled variables might confound experimental results – providing a useful tool for experimental stimuli selection.

This work showcases how advanced machine learning techniques can be adapted to serve psychological research – not by replacing human judgments but assisting them in a way that acknowledges both the potential and limitations of computational approaches. Moreover, by making these tools available through a web application, the article extends the availability of these tools to researchers without specialized technical backgrounds.

Article Number Three - High Risk of Political Bias in Black Box Emotion Inference Models

The third study addresses a critical yet underexplored dimension of bias in machine learning-based sentiment analysis systems: political bias. While previous research has documented various social biases in computational models – particularly concerning gender and race (Kiritchenko & Mohammad, 2018) – this study specifically examines how annotators' political orientations can systematically influence the performance of emotion inference models in an implicit way, with potentially far-reaching implications for research that employs these tools (Plisiecki et al., 2025).

Using a previously developed Polish sentiment analysis model, we conducted a comprehensive bias audit to assess whether the model's valence predictions exhibited systematic differences based on the political affiliations of mentioned politicians. The analysis focused on 24 well-known Polish political figures from across the political spectrum, examining how the model rated both their names in isolation and when embedded within neutral or politically charged sentences.

The results revealed compelling evidence of political bias. Regression analyses demonstrated that political affiliation explained approximately 49% of the variance in the

model's valence predictions for politicians' names. When controlling confounding variables such as gender, this explanatory power increased to 66.5%. These differences were not randomly distributed but showed systematic patterns aligned with specific political orientations, as confirmed by permutation tests ($p = 0.008$ for names, $p = 0.049$ for neutral sentences, and $p = 0.018$ for political sentences). Importantly, the observed bias could not be explained by general public opinion toward these politicians (as measured by trust surveys) or by inherent linguistic properties of the texts in which they appeared. Instead, the bias appeared to originate from the subjective perceptions of the annotation team, despite the annotators being explicitly asked to rate the emotions they see expressed in the texts, as opposed to those they feel when reading it. This conclusion was further supported by an experiment in which we pruned the training dataset of all texts mentioning these politicians and retrained the model. The modified model exhibited significantly reduced bias, although some residual bias persisted, suggesting deeper associative patterns may also contribute to the effect.

The study's findings have significant implications for the use of machine learning-based sentiment analysis in psychological and social science research. Unlike lexicon-based approaches, which rely on pre-defined word lists evaluated independently of context, black box supervised models trained on human annotations inherently propagate the subjective judgments of their annotators – including implicit political biases that may operate outside of conscious awareness. This propagation creates systematic distortions that can significantly impact research conclusions. We recommend researchers exercise caution when using machine learning-based sentiment analysis for psychological research, particularly in politically sensitive contexts.

Article Number Four - Eradicating Social Biases in Sentiment Analysis using Semantic Blinding and Semantic Propagation Graph Neural Networks

The fourth study addresses a critical challenge highlighted by the third study: the propagation of social biases from training data to model predictions. To address this problem, the study introduces the Semantic Propagation Graph Neural Network (SProp GNN), a novel explainable architecture designed to analyze emotions in text while mitigating social biases (Plisiecki, 2024). The key innovation is the concept of "semantic blinding"—deliberately limiting the model's access to specific semantic information that could introduce unwanted biases. Instead of processing the full semantic content of words, the model focuses exclusively on syntactic relationships between words and their emotional values at the individual word level. This approach constitutes a fundamental shift from traditional machine learning models. While conventional models can learn associations between specific words (like politicians' names or gender-specific terms) and emotional values, the SProp GNN cannot form these direct associations because it does not have

access to the specific words themselves. Instead, it analyzes how emotional information flows through the syntactic structure of sentences.

The model was evaluated across three datasets spanning two languages (English and Polish) and two different emotion prediction tasks (categorical and dimensional). The results demonstrated that the SProp GNN significantly outperformed lexicon-based alternatives while approaching the accuracy levels of transformer-based models. Most importantly, rigorous statistical testing confirmed that the SProp GNN substantially reduced bias compared to transformer models. When tested on political content, the transformer model showed clear political and gender biases, with these covariates explaining up to 66% of the variance in valence predictions. In contrast, the SProp GNN showed no significant association between political affiliation or gender and predicted emotions. Both direct regression analysis and comparative approaches confirmed this bias reduction.

This work represents an important step toward more ethical and unbiased computational methods in psychological research, demonstrating that advanced machine learning techniques can be adapted to address concerns about bias while maintaining high performance for emotion analysis tasks. Due to its explainability, it furthermore allows researchers to directly probe the pathways of propagation of emotional information through the syntactic structure of the text, potentially opening new avenues for psycholinguistic research.

General Discussion

The introduction of NLP, or computer-based language analysis to psychology has been repeatedly referred to as constituting a paradigmatic shift in how we conduct science (Boyd & Schwartz, 2021; J. W. Pennebaker, 2022). The research presented in this thesis ties well into this argument. Firstly, paper number one offers an new method of quantifying psychological phenomena, by extracting their numerical representations from text. By relying on text written in an unprompted way by thousands of Reddit users, it provides a new medium for testing psychological theories, which circumvents the ecological problems associated with classical psychological studies. This method, given appropriate data sources, can be applied to other psychological topics, such as personality, where numerical representations, akin to those created slightly less than one hundred years ago by Baldwin 1942 can be created and analyzed. This approach, however, has one significant limitation that psychologists will have to grapple with, namely the fact that people do not always accurately portray their inner experiences when writing text. While having direct access to people's thoughts and feelings would be fascinating, the fact that we don't does not preclude us from drawing conclusions from the texts they write as at the heart of the text analysis paradigm (Pennebaker, 2022) lies the assumption that even though a person might not want to express her real feelings, the words they choose will still reflect

the things they pay attention to, and similarly to how an eye tracker does, it will rear scientifically meaningful data.

The second paper shows that machine learning approaches can also be applied to extend existing psychological studies on the affective load of words, by generating synthetic data. The important caveat here is that the patterns that the norm extrapolation models have learned do not have to reflect the actual mechanisms that guide our affective responses to singular words and therefore should not be treated as equivalent to empirically collected data. However, they can serve as an exploratory tool for finding interesting patterns, which can be afterwards validated by empirical studies. By extending available datasets, and creating an algorithm for picking experimental stimuli, the fruits of this study can help accelerate psychological research.

The third and the fourth studies are deeply tied together. The third pinpoints a problem, which can derail the conclusions drawn by studies utilizing sentiment analysis models. It legitimates the worries associated with using black-box models, showing that they indeed exhibit biases that can easily go unnoticed when researchers only focus on their optimization criteria. After all, the performance of these models is measured by the extent to which they can replicate the emotion labels from the dataset that is used to test them. However, given the psychological insight that the annotators' attitudes can influence their judgement in implicit ways, the most performant models according to those standards will be exactly those that learn these individual attitudes. This finding called for the creation of machine learning architectures that take this insight into account, leading to the fourth study which showcased a novel architecture that is less sensitive to those individual attitudes. The SProp GNN circumvents social biases associated with specific entities and ideas by hiding them from the prediction model at training and inference. By doing this, it is able to learn the general emotional value of texts from the annotators, while turning a blind eye to their individual implicit attitudes, thereby enabling more valid psychological insights drawn with the use of sentiment analysis models.

In total, these four studies solve important scientific problems on the intersection of psychology and machine learning. They create new tools of inquiry, apply advanced machine learning techniques to solve methodological problems, identify barriers of entry of black box models into psychological research, and develop techniques that help circumvent those barriers. Together, this research constitutes a directed push for the wider adoption of ML and NLP methods in psychological research. The studies presented here would not be possible without the psychological core on which they were built. From a purely engineering point of view the exploration of the fundamental components of emotions using text-based vectors has close to little value. Similarly, the extent of bias shown in the third study would probably be of little consequence in most commercial applications

of sentiment analysis, but when amassed in a psychological research database can lead to false conclusions. These studies therefore also show that while there are applications for which psychology can just borrow methods from natural language processing, there is an essential need for machine learning and natural language processing research that is strictly psychological.

References

Allport, G. W. (1942). The use of personal documents in psychological science. *Social Science Research Council Bulletin*, *49*, xix + 210–xix + 210.

Allport, G. W., Bruner, J. S., & Jandorf, E. M. (1953). Personality under social catastrophe: Ninety life-histories of the Nazi revolution. *Personality in nature, society, and culture*, 436–455.

Alsini, R., Naz, A., Khan, H. U., Bukhari, A., Daud, A., & Ramzan, M. (2024). Using deep learning and word embeddings for predicting human agreeableness behavior. *Scientific Reports*, *14*(1), 29875. https://doi.org/10.1038/s41598-024-81506-8

Baldwin, A. L. (1942). Personal structure analysis: A statistical method for investigating the single personality [Place: US Publisher: American Psychological Association]. *The Journal of Abnormal and Social Psychology*, *37*(2), 163–183. https://doi.org/10.1037/h0061697

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., … Schulz, E. (2024). Centaur: A foundation model of human cognition [Version Number: 2]. https://doi.org/10.48550/ARXIV.2410.20268

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, *10*. Retrieved February 6, 2025, from https://www.researchgate.net/profile/Ryan-Boyd-8/publication/358725479_The_Development_and_Psychometric_Properties_of_LIWC-22/links/6210f62c4be28e145ca1e60b/The-Development-and-Psychometric-Properties-of-LIWC-22.pdf

Boyd, R. L., & Schwartz, H. A. (2021). Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. *Journal of Language and Social Psychology*, *40*(1), 21–41. https://doi.org/10.1177/0261927X20967028

Buechel, S., & Hahn, U. (2017). Readers vs. Writers vs. Texts: Coping with Different Perspectives of Text Understanding in Emotion Annotation. *Proceedings of the 11th Linguistic Annotation Workshop*, 1–12. https://doi.org/10.18653/v1/W17-0801

Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. https://doi.org/10.1613/jair.1.12228

Calder, A. J., Burton, A., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, *41*(9), 1179–1208. https://doi.org/10.1016/S0042-6989(01)00002-5

Casella, M., Luongo, M., Marocco, D., Milano, N., & Ponticorvo, M. (2024). LLM embeddings on test items predict post hoc loadings in personality tests. Retrieved April 8, 2025, from https://www.iris.unina.it/handle/11588/961475

Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155–162. https://doi.org/10.1017/S1351324916000334

Couto, M., Perez, A., Parapar, J., & Losada, D. E. (2025). Temporal Word Embeddings for Early Detection of Psychological Disorders on Social Media. *Journal of Healthcare Informatics Research*. https://doi.org/10.1007/s41666-025-00186-9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Version Number: 2]. https://doi.org/10.48550/ARXIV.1810.04805

Durrheim, K., Schuld, M., Mafunda, M., & Mazibuko, S. (2023). Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, *62*(1), 617–629. https://doi.org/10.1111/bjso.12560

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis, Special Volume/Blackwell*.

Frenzel, A. C., Kleen, H., Marx, A. K. G., Sachs, D. F., Baier-Mosch, F., & Kunter, M. (2025). Is it in their words? Teachers' enthusiasm and their natural language in class–A sentiment analysis approach. *British Journal of Educational Psychology*, bjep.12734. https://doi.org/10.1111/bjep.12734

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16). https://doi.org/10.1073/pnas.1720347115

Glenn, J. J., Nobles, A. L., Barnes, L. E., & Teachman, B. A. (2020). Can Text Messages Identify Suicide Risk in Real Time? A Within-Subjects Pilot Examination of Temporally Sensitive Markers of Suicide Risk. *Clinical Psychological Science*, *8*(4), 704–722. https://doi.org/10.1177/2167702620906146

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2-3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hills, T. T., Proto, E., Sgroi, D., & Seresinhe, C. I. (2019). Historical analysis of national subjective wellbeing using millions of digitized books [Publisher: Nature Publishing Group]. *Nature Human Behaviour*, *3*(12), 1271–1275. https://doi.org/10.1038/s41562-019-0750-z

Hodson, N., & Williamson, S. (2024). Can Large Language Models Replace Therapists? Evaluating Performance at Simple Cognitive Behavioral Therapy Tasks. *JMIR AI*, *3*, e52500. https://doi.org/10.2196/52500

Hofmann, M. J., Biemann, C., Westbury, C., Murusidze, M., Conrad, M., & Jacobs, A. M. (2018). Simple Co-Occurrence Statistics Reproducibly Predict Association Ratings. *Cognitive Science*, *42*(7), 2287–2312. https://doi.org/10.1111/cogs.12662

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-Based Deep Neural Language Modeling for Construct-Specific Automatic Item Generation. *Psychometrika*, *87*(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

Imbir, K., Pastwa, M., & Walkowiak, M. (2023). The Role of the Valence, Arousing Properties and Subjective Significance of Subliminally Presented Words in Affective Priming. *Journal of Psycholinguistic Research*, *52*(1), 33–56. https://doi.org/10.1007/s10936-021-09815-x

Imbir, K. K. (2016). Affective Norms for 4900 Polish Words Reload (ANPW_r): Assessments for Valence, Arousal, Dominance, Origin, Significance, Concreteness, Imageability and, Age of Acquisition. *Frontiers in Psychology*, *7*, 1081. https://doi.org/10.3389/fpsyg.2016.01081

Johnson, S. J., Murty, M. R., & Navakanth, I. (2024). A detailed review on word embedding techniques with emphasis on word2vec. *Multimedia Tools and Applications*, *83*(13), 37979–38007. https://doi.org/10.1007/s11042-023-17007-z

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring Emotional Expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology*, *120*(2), 263–286. https://doi.org/10.2307/20445398

Kiritchenko, S., & Mohammad, S. M. (2018, May). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems [arXiv:1805.04508 [cs]]. https://doi.org/10.48550/arXiv.1805.04508

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, *99*, 101861. https://doi.org/10.1016/j.inffus.2023.101861

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45), e2405460121. https://doi.org/10.1073/pnas.2405460121

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, *116*(13), 5862–5871. https://doi.org/10.1073/pnas.1820240116

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowl-

edge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents [ISSN: 1938-7228]. *Proceedings of the 31st International Conference on Machine Learning*, 1188–1196. Retrieved March 22, 2025, from https://proceedings.mlr.press/v32/le14.html

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space [Version Number: 3]. https://doi.org/10.48550/ARXIV.1301.3781

Oota, S. R., Manwani, N., & Bapi, R. S. (2018). fMRI Semantic Category Decoding Using Linguistic Encoding of Word Embeddings [Series Title: Lecture Notes in Computer Science]. In L. Cheng, A. C. S. Leung, & S. Ozawa (Eds.), *Neural Information Processing* (pp. 3–15, Vol. 11303). Springer International Publishing. https://doi.org/10.1007/978-3-030-04182-3_1

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., & Booth, R. (2007). The Development and Psychometric Properties of LIWC2007.

Pennebaker, J. W. (2001). Linguistic inquiry and word count: LIWC 2001.

Pennebaker, J. W. (2022). Computer-based language analysis as a paradigm shift. In *Handbook of language analysis in psychology* (pp. 576–587). The Guilford Press.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Retrieved February 6, 2025, from https://repositories.lib.utexas.edu/items/705e81ca-940d-4c46-94ec-a52ffdc3b51f

Perumal, T., Mustapha, N., Mohamed, R., & Shiri, F. M. (2024). A Comprehensive Overview and Comparative Analysis on Deep Learning Models. *Journal on Artificial Intelligence*, *6*(1), 301–360. https://doi.org/10.32604/jai.2024.054314

Plisiecki, H. (2024). Eradicating Social Biases in Sentiment Analysis using Semantic Blinding and Semantic Propagation Graph Neural Networks [Version Number: 3]. https://doi.org/10.48550/ARXIV.2411.12493

Plisiecki, H., Koc, P., Flakus, M., & Pokropek, A. (2024). Predicting Emotion Intensity in Polish Political Texts: Comparing Supervised Models and Large Language Models in a Resource-Poor Language [Version Number: 1]. https://doi.org/10.48550/ARXIV.2407.12141

Plisiecki, H., Lenartowicz, P., Flakus, M., & Pokropek, A. (2025). High risk of political bias in black box emotion inference models. *Scientific Reports*, *15*(1), 6028. https://doi.org/10.1038/s41598-025-86766-6

Plisiecki, H., & Sobieszek, A. (2023). Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection. *Behavior Research Methods*, *56*(5), 4716–4731. https://doi.org/10.3758/s13428-023-02212-3

Plisiecki, H., & Sobieszek, A. (2024). Emotion topology: Extracting fundamental components of emotions from text using word embeddings. *Frontiers in Psychology*, *15*, 1401084. https://doi.org/10.3389/fpsyg.2024.1401084

Psathas, G. (1969). The general inquirer: Useful or not? *Computers and the Humanities*, *3*(3), 163–174. https://doi.org/10.1007/BF02401609

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training [Publisher: San Francisco, CA, USA]. Retrieved May 25, 2025, from https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance [Place: United Kingdom Publisher: Taylor & Francis]. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1584–1598. https://doi.org/10.1080/17470218.2014.941296

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(1), 23. https://doi.org/10.1140/epjds/s13688-016-0085-1

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting High-Level Human Judgment Across Diverse Behavioral Domains (S. Vazire & S. Vazire, Eds.). *Collabra: Psychology*, *5*(1), 50. https://doi.org/10.1525/collabra.282

Roland, P. E. (2023). How far neuroscience is from understanding brains. *Frontiers in Systems Neuroscience*, *17*, 1147896. https://doi.org/10.3389/fnsys.2023.1147896

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Scerrati, E., D'Ascenzo, S., Nicoletti, R., Villani, C., & Lugli, L. (2022). Assessing Interpersonal Proximity Evaluation in the COVID-19 Era: Evidence From the Affective Priming Task. *Frontiers in Psychology*, *13*, 901730. https://doi.org/10.3389/fpsyg.2022.901730

Smith, M. S. (1968). The computer and the TAT. *Journal of School Psychology*, *6*(3), 206–214. https://doi.org/10.1016/0022-4405(68)90017-4

Snefjella, B., & Blank, I. (2020). Semantic norm extrapolation is a missing data problem [Publisher: PsyArXiv]. https://doi.org/https://doi.org/10.31234/osf.io/y2gav

Sobieszek, A., & Price, T. (2022). Playing Games with Ais: The Limits of GPT-3 and Similar Large Language Models. *Minds and Machines*, *32*(2), 341–364. https://doi.org/10.1007/s11023-022-09602-0

Sohail, A., & Zhang, L. (2025). Using large language models to facilitate academic work in the psychological sciences. *Current Psychology*. https://doi.org/10.1007/s12144-025-07438-2

Stone, P. J. (1966). The General Inquirer: A computer approach to content analysis.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need [Version Number: 7]. https://doi.org/10.48550/ARXIV.1706.03762

Wang, J., Fan, Y., Palacios, J., Chai, Y., Guetta-Jeanrenaud, N., Obradovich, N., Zhou, C., & Zheng, S. (2022). Global evidence of expressed sentiment alterations during the COVID-19 pandemic. *Nature Human Behaviour*, *6*(3), 349–358. https://doi.org/10.1038/s41562-022-01312-y

Wang, X., Li, X., Yin, Z., Wu, Y., & Liu, J. (2023). Emotional intelligence of Large Language Models. *Journal of Pacific Rim Psychology*, *17*, 18344909231213958. https://doi.org/10.1177/18344909231213958

Warriner, A. B. (2014, November). *The Interplay of Language and Emotion: Using Affective Norms to Explore Word Recognition, Motivation, and Lexicon* [Thesis] [Accepted: 2014-10-28T15:51:09Z]. Retrieved April 8, 2025, from https://macsphere.mcmaster.ca/handle/11375/16227

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Weintraub, W. (1989). Verbal behavior in everyday life.

Weintraub, W. (1981). *Verbal behavior: Adaptation and psychopathology*. Springer Publishing Company.

Widmann, T., & Wich, M. (2023). Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text. *Political Analysis*, *31*(4), 626–641. https://doi.org/10.1017/pan.2022.15

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 08 October 2024
DOI 10.3389/fpsyg.2024.1401084

OPEN ACCESS

EDITED BY
Cristian López Raventós,
National Autonomous University of Mexico, Mexico

REVIEWED BY
Peter Lewinski,
University of Oxford, United Kingdom
Juan Manuel Mayor Torres,
Montreal Institute for Learning Algorithm (MILA), Canada
Rajeev Ratna Vallabhuni,
Bayview Asset Management, LLC, United States

*CORRESPONDENCE
Hubert Plisiecki
✉ hplisiecki@gmail.com

RECEIVED 14 March 2024
ACCEPTED 03 September 2024
PUBLISHED 08 October 2024

CITATION
Plisiecki H and Sobieszek A (2024) Emotion topology: extracting fundamental components of emotions from text using word embeddings.
Front. Psychol. 15:1401084.
doi: 10.3389/fpsyg.2024.1401084

Emotion topology: extracting fundamental components of emotions from text using word embeddings

Hubert Plisiecki[1]* and Adam Sobieszek[2]

[1]Research Lab for the Digital Social Sciences, IFIS PAN, Warsaw, Poland, [2]Department of Psychology, University of Warsaw, Warsaw, Poland

This exploratory study examined the potential of word embeddings, an automated numerical representation of written text, as a novel method for emotion decomposition analysis. Drawing from a substantial dataset scraped from a Social Media site, we constructed emotion vectors to extract the dimensions of emotions, as annotated by the readers of the texts, directly from human language. Our findings demonstrated that word embeddings yield emotional components akin to those found in previous literature, offering an alternative perspective not bounded by theoretical presuppositions, as well as showing that the dimensional structure of emotions is reflected in the semantic structure of their text-based expressions. Our study highlights word embeddings as a promising tool for uncovering the nuances of human emotions and comments on the potential of this approach for other psychological domains, providing a basis for future studies. The exploratory nature of this research paves the way for further development and refinement of this method, promising to enrich our understanding of emotional constructs and psychological phenomena in a more ecologically valid and data-driven manner.

KEYWORDS

word embeddings, emotion decomposition, natural language processing, valence, arousal

1 Introduction

In the study of core components of emotions various methods have been used. A large number of studies focus on the core components of emotions by using controlled environments. Here, participants either annotate distinct stimuli, such as photos of facial expressions (Calder et al., 2001; Fontaine et al., 2002, 2007; Schlosberg, 1952; Shaver et al., 1987) or assess their emotional experiences through structured questionnaires (Nowlis and Nowlis, 1956; Feldman, 1995; Stanisławski et al., 2021). These studies have explored areas such as facial expressions, emotion terms, and self-reported emotional experiences. Except for self-reports, participants annotate stimuli based on their emotional resonance. For instance, a photo capturing a broad Duchenne smile might receive a maximum rating for inferred happiness (Calder et al., 2001; Ekman et al., 1990; Tseng et al., 2014). Other research, following the Multidimensional Scaling (MDS) approach, requires participants to gauge the emotional similarity among various stimuli, such as musical pieces (Dellacherie et al., 2011), emotion terms (Bliss-Moreau et al., 2020), and facial expressions (Woodard et al., 2022).

To analyze these core components, researchers frequently utilize Principal Component Analysis (PCA) (e.g., Calder et al., 2001; Feldman, 1995; Fontaine et al., 2007; Lampier et al., 2022). At its core, PCA condenses intricate datasets by converting correlated variables into a smaller set of uncorrelated ones, known as principal components. These components highlight the primary patterns within the data (Abdi and Williams, 2010). When applied to emotional experience studies, PCA effectively pinpoints foundational dimensions like valence. It does so by transforming extensive emotional descriptors (e.g., scores from an emotional experience questionnaire) into distinct, principal emotional axes (e.g., positive–negative). This method provides researchers with a refined lens to understand the complex landscape of human emotions.

Through statistical analysis, psychologists have proposed various models of the core structure of emotional experience. These models often suggest two primary dimensions: valence (e.g., happiness vs. sadness) and arousal (e.g., stressed vs. relaxed) (Russell, 1980; Stanisławski et al., 2021). Some models also introduce additional dimensions like potency/dominance, which gauges how in control individuals feel over their environment and others (e.g., anger—high dominance; fear—low dominance), and unpredictability, reflecting the consistency of one's surroundings in eliciting emotions (e.g., surprise—high unpredictability; calmness—low unpredictability) (Fontaine et al., 2007; Mehrabian, 1996; Russell and Mehrabian, 1977). Nonetheless, certain researchers continue to advocate for a strictly 2-dimensional perspective (Bliss-Moreau et al., 2020).

The dimensional framework, despite some disagreements about its structure, has gained substantial support in the psychological community. It's been incorporated into neuroscientific research, offering fresh perspectives on emotional processing in the brain (Posner et al., 2005) and the origins of depression (Barrett et al., 2016). This approach has proven effective in gauging affect in physical activities (for a comprehensive review, refer to Evmenenko and Teixeira, 2022), advertising (Wiles and Cornwell, 1991), various priming and linguistic investigations (Imbir, 2016; Imbir et al., 2020; Syssau et al., 2021; Yao et al., 2016), and in machine learning (Islam et al., 2019; Martínez-Tejada et al., 2020; Nicolaou et al., 2011). While an exhaustive discussion of the dimensional model's applications is beyond this article's scope, we want to emphasize its broad appeal, not only within psychology but also in other scientific disciplines.

Our paper introduces a data-driven method that utilizes word embeddings (a machine learning technique) to analyze emotional expression as communicated and perceived through the medium of text and extract its core dimensions from vast amounts of text that reflect real-world contexts. Innovations in word embeddings facilitate the quantitative examination of extensive text datasets (Mikolov et al., 2013a,b). By automating insight extraction from texts, these embeddings have the potential to replicate previous findings in a new medium—unprompted written text—garnering more objective evidence for their validity. Furthermore, they can process vast text volumes, expanding the impact of conclusions drawn (Jackson et al., 2022). In subsequent sections, we offer a comprehensive review of word embeddings and discuss their potential benefits. We then transition into the details of our current study. Prior to presenting the methodology, we also establish clear definitions for the concepts associated with word embeddings, ensuring they are well anchored in emotion research.

Word embeddings are a technique popularized by Mikolov et al. (2013a,b) which makes it possible to quantify natural language. It computes separate strings of numbers (usually between 100 and 500 long), known as vectors, for each unit of text that is to be analyzed. Most often the units are words (hence "word" embeddings), and so each unique word in a given text gets assigned a vector which encodes its relation to the other words and can therefore be used to analyze its properties (Gutiérrez and Keith, 2019). In the case where one wants to analyze whole documents, composed of multiple words, separate vectors can be created for each of them as well (Le and Mikolov, 2014).

Some of the popular traits of these vectors are that, given that they were derived from a large enough batch of text (the more the better), their similarity (calculated through a formula called cosine similarity) correlates with human judgements about the similarity of the words that they relate to (Jatnika et al., 2019). Their results are therefore similar to the results obtained through the MDS method, providing a similarity metric that replaces human judgments made in the laboratory.

Importantly, these word embeddings have been used repeatedly to predict (using simple techniques, such as linear regressions) different meanings of text snippets. These use cases included, among others, predicting diseases based on the International Classification of Diseases (ICD-10) and the Unified Medical language System (UMLS) (Khattak et al., 2019), identifying cultural biases (Charlesworth et al., 2021; Durrheim et al., 2023), human judgements (Richie et al., 2019), moral values (Lin et al., 2018), and emotions and sentiments (Al-Amin et al., 2017; Jia, 2021; Plisiecki and Sobieszek, 2023; Widmann and Wich, 2022). This last application of word embeddings is especially important for the current study as it shows that word embeddings encode information that correlates with emotional meanings. This case is further strengthened by van Loon and Freese's (2023) research, which has directly shown that affective meaning can be recovered from word embeddings by successfully predicting evaluation, potency, and activity profiles of words. Al-Amin and his team (2017) predicted positive vs. negative sentiment of texts collected from Bengalese blogging websites. Jia (2021) classified both basic emotions and overall polarity in Chinese texts. Plisiecki and Sobieszek (2023) showed that leveraging advanced word embeddings makes it possible to predict a range of emotional indices for singular words in different languages (English, German, French, Polish, Dutch). Widmann and Wich (2022) prepared a comparison of different ways of creating word embeddings on German texts for the prediction of basic emotions, comparing both newer and more classical approaches of constructing them and showed that all of them have significant predictive ability. These examples stand as evidence that word embeddings encode emotional information. They are therefore good sources of data for the current application.

Think of creating word embeddings as mapping words to a multidimensional space where the location of each word is determined by its context, or the words with which it often coexists. Imagine a large book, where every unique word is listed. The creation of word embeddings begins with each word starting at a random location in this space. As we move through the book, sentence by sentence, the algorithm adjusts the positions of the words in this space based on their context. For instance, if "cat" and "kitten" often appear in similar contexts, they gradually move closer together. Conversely, "cat" and "refrigerator", unlikely to share much context, would drift apart. This process is repeated multiple times (known as iterations) on the entire

book, refining the word positions each time. After sufficient iterations, the distances and angles between word vectors represent different types of semantic and syntactic similarities. For instance, words with similar meanings would be closer together, and the direction of specific relations (such as verb tense or gender) would be consistent. This way, word embeddings provide a rich, numeric interpretation of word relationships, useful in various language-related tasks (Mikolov et al., 2013a,b).

These word-level embeddings can be extended to document-level representations. Le and Mikolov (2014) introduced the Paragraph Vector, or Doc2Vec, an extension of word2vec that computes a vector for a sentence or document, not only for individual words. The technique involves training a model where the document vector, along with the word vectors, work together to predict the surrounding words in a document, thereby capturing the semantic essence of the entire text. Just like single words move closer or further in this numerical space based on their cooccurrences with other words, so too now whole documents get embedded in places where they fit best based on the similarities and differences in their overall content and context. This document-level vector enables researchers to compare and contrast entire documents, opening up further avenues in natural language processing tasks.

In this study we explore whether similar emotional components to those identified in previous literature (e.g., Fontaine et al., 2002), can be extracted from a large text dataset using word embeddings. We reverse the process of annotation and make use of a dataset in which the participants did not describe emotions using questionnaires, but rather spotted them in an already existing array of natural language expressions. While describing human emotions using questionnaires is not an everyday task for human beings, and therefore is not natural to them, potentially leading to issues of ecological validity, the action of inferring emotions from language is an everyday, nearly constant exercise that humans engage in. Furthermore, this specific type of judging others' emotions—through text written by a stranger—is a very common occurrence in today's digital world, and therefore is of high importance to the research community. Using word embeddings, we represent the annotated texts in an emergent numerical space.

In the following text, we will use a specific terminology for describing different concepts related to word embeddings, as applied to the study of emotion. This is done to enhance clarity and provide psychologists with a strong conceptual grasp of the following study. 1. To describe the multidimensional space, within which numerical vectors reside, we will use the term Emotional Space. 2. The vectors representing the emotional content of texts will be called *Emotion Vectors*. 3. When vectors do not correspond to specific emotions, but to words or single documents we will use either *Word Vectors* or *Document Vectors*, to designate them.

2 Method

2.1 Dataset

The GoEmotions dataset was developed by a team of researchers at Google to study human emotions within the realm of machine learning (Demszky et al., 2020). It includes 58,000 Reddit comments annotated with regard to 28 unique emotions, totaling over 210,000 annotations. The data came from a Reddit data dump, sourced from the reddit-data-tools project. The data dump included all comments from 2005 to January 2019. As the Reddit platform is composed of different communities of users, called Subreddits, all communities with at least 10 k comments were chosen for the analysis. The comments from different subreddits were then further balanced. First, the number of comments from the most popular subreddits was capped at the median Subreddit count. The comments were then randomly sampled for annotation.

Because the Reddit community does not reflect the globally diverse population, due to a skew towards offensive language, the toxic comments were removed from the dataset using a pre-defined list of offensive words and the help of manual annotators. This was done before the sampling process. According to best practices the researchers have modified the dataset by removing stop words and stemming the words in order to transform them into their base form (e.g., "fearsome" into "fear").

2.2 Emotion taxonomy

The emotion taxonomy for annotation was created as a result of three steps: 1. Manual annotation of a small subset of the data to ensure proper coverage of emotions expressed in the text. 2. Review of psychological literature on basic emotions (Plutchik, 1980; Cowen and Keltner, 2020; Cowen et al., 2019). 3. Removal of the emotions that were deemed to have a high overlap to limit the overall number of emotions.

The resulting list of emotions included: *admiration, approval, annoyance, gratitude, disapproval, amusement, curiosity, love, optimism, disappointment, joy, realization, anger, sadness, confusion, caring, excitement, surprise, disgust, desire, fear, remorse, embarrassment, nervousness, pride, relief, grief*.

2.3 Annotation

Three raters were assigned to each comment, and asked to select those emotions, which they believed were expressed in the text. All three raters were native English speakers from India. The authors here rely on the results of a cross-cultural study showing that the emotion judgments of Indian and US English speakers largely occupy the same dimensions (Cowen et al., 2019). In the case where the annotators judged the text to be especially difficult to rate, they were able to choose not to assign any emotion to it. Whenever there was no agreement between the raters on a specific example, additional raters were assigned to it until each document was annotated at least twice with regards to the same emotional label.

2.4 Analysis

The analysis aims to represent the natural expression of emotions contained in the GoEmotions dataset in the word-embedding-based emotion space. The breakdown of the analysis is presented in Figure 1.

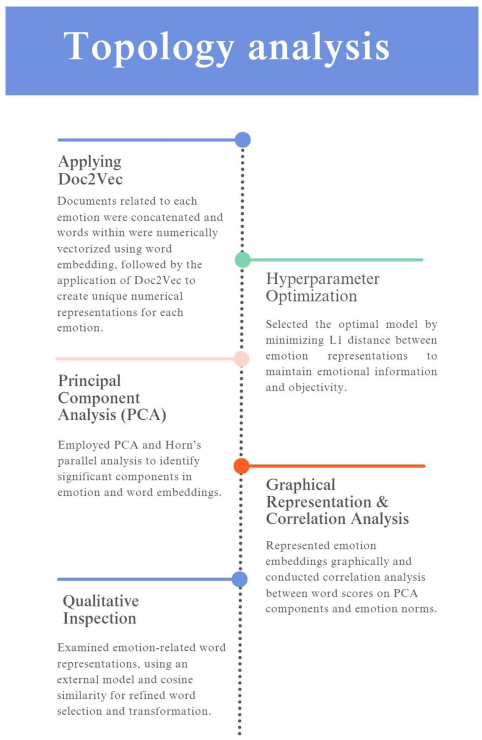


FIGURE 1
The steps of the analysis.

2.4.1 Applying Doc2Vec to create numerical representations of emotions

The Doc2Vec algorithm (Le and Mikolov, 2014) was used to create emotion vectors for each emotion in the dataset. Documents that corresponded to a given emotion were concatenated into long documents, and then, during training, singular emotion vectors were created for each of these long documents. For a document to be judged as corresponding to a given emotion it was enough for it to be classified as so once. So, if a document was classified by two raters into two different emotions, this document then complemented two different concatenated series. This approach was chosen because applying majority voting retains less information from the annotators, and judging emotions is a highly subjective task where the objective truth can be rarely established. First, the words in each document were transformed into word vectors via a word embedding method, capturing the information embedded in each word. Then, these word vectors were used to build an emotion vector using the Doc2Vec algorithm, which treats the document as another word in the sentence and assigns numerical representations to it (Le and Mikolov, 2014). This resulted in a distinct numerical representation for each emotion that encapsulated the underlying sentiment, and thematic nuances present in the corresponding documents. Supplementary analyses of the distribution of document vectors and their relation to label centroids, including top-k nearest centroid accuracy, conducted to explore the resultant document vector space are presented in the Supplementary Material for the interested reader.

2.4.2 Hyperparameter optimization

Because the Doc2Vec algorithm has a range of hyperparameters that had to be tuned in order to achieve the best representations, separate emotion spaces were created using different hyperparameter values. The hyperparameters that were taken into consideration were the collocation window size (5, 10, 20 words), minimum word count (10, 40, 60 words), embedding size (100, 200, 300, 400, 500, 600, 700, 800, 900 units). Every combination of the above parameters was tested. We chose the model that minimized the L1 distance between the emotion vectors to increase the likelihood that the emotion vectors represented meanings of emotions—as they would be more similar to each other if they truly belonged to the semantic space that describes emotions—while at the same time ensuring it did not impose any further predefined notions onto the contents of the vectors.

2.4.3 Principal component analysis (PCA)

The emotion vectors were then subjected to a Principal Component Analysis, in line with the previous literature on decomposing emotions (Fontaine et al., 2002, 2007), which finds the dimensions along which the emotional representations (emotion vectors) vary the most and situates the emotions along them. The PCA was applied to the emotion vectors. Horn's parallel analysis was used to determine the number of components that can be retained. This method compares the eigenvalues obtained from the factor analysis to those from a randomly generated dataset. If the eigenvalues from the factor analysis exceed those from the randomly generated dataset, the factors are considered significant and are retained.

2.4.4 Graphical representation and correlation analysis

Emotion vectors were then plotted on a graph, and the words corresponding to the word vectors were tested for correlation with a set of words annotated with regard to their emotional loads along the first components (stipulated to be related to the components reported in the previous literature, Gendron and Feldman Barrett, 2009). In order to inspect these components, the word vectors retrieved from the dataset were transformed to align with the components identified by the PCA.

2.4.5 Qualitative inspection

Because only some words present in the vocabulary were related to emotions, a qualitative inspection of only the highest and lowest-ranking words on each of the components could obscure the nature of the recovered dimensions, as it is the emotion related words that have the highest face validity when it comes to examining emotional dimensions. To circumvent this problem an external word embedding model with 300-dimensional vectors (Dadas, 2019) was used to sample the vocabulary for words related to the concept of emotions. The cosine similarity of word vectors was used to recover only 500 words most similar to the word vector for the word "emotion" based on the cosine similarity between the vectors that represented them. The resulting words were then subjected to the PCA transformation again, so that they could be evaluated qualitatively.

2.4.6 t-Distributed stochastic neighbor embedding (t-SNE) analysis

To complement the Principal Component Analysis (PCA) and further explore the structure of the emotion vectors, we used

t-Distributed Stochastic Neighbor Embedding (t-SNE). t-SNE is a nonlinear technique that helps visualize high-dimensional data by preserving local relationships, making it useful for identifying clusters and patterns that PCA might miss. For our analysis, we first standardized the emotion vectors to ensure that all features contributed equally. We applied t-SNE with the following settings: 2 components, a perplexity of 5, and a learning rate of 10. The random state was set to 22 to ensure that the results could be replicated. The perplexity was set to 5, the lower bound of the suggested values, due to the low number of emotion vectors. Perplexity, which balances attention between local and global aspects of the data, typically needs to be higher for larger datasets to capture broader relationships; however, for smaller datasets like ours, a lower perplexity is recommended as it helps maintain meaningful local structures (Van der Maaten and Hinton, 2008). The learning rate was set to 10, as this value provided a stable convergence during the embedding process, ensuring that the visualization accurately represented the underlying data patterns.

2.4.7 Logistic regression on documents

To confirm the alignment of the PCA components with the emotional dimension of Valence, we recoded the original GoEmotions dataset from 28 emotions into positive and negative labels. The emotions classified as positive were admiration, love, gratitude, amusement, realization, optimism, curiosity, excitement, caring, joy, approval, pride, desire, and relief. The emotions classified as negative were sadness, disapproval, disappointment, annoyance, confusion, disgust, remorse, anger, grief, embarrassment, surprise, fear, and nervousness. If a text was labeled with a different emotion it was dropped. Here again, all text labels were taken into consideration and so if two annotators annotated a given text as joy, these were treated as separate rows. This approach was chosen over majority voting to preserve as much information from the original annotations as possible, given the subjective nature of emotion labeling. The final dataset consisted of 155,663 text—label pairs. We then transformed the document vectors from the Doc2Vec model using the PCA model previously fit on the emotion vectors, resulting in a four-dimensional vector for each document. These vectors were subsequently used in a logistic regression with the positive/negative labels as the dependent variable.

3 Results

3.1 Horn's parallel analysis

The Horn's parallel analysis indicated that the first seven components were significant and should be retained (see Figure 2). Even though seven components were significant, we chose to only inspect the first four of them, as after that number, the percentage of explained variance drops sharply.

3.2 Visualizing the emotion vectors

To visualize the emotion vectors regarding the components recovered by the PCA, we plotted them on two 2-dimensional graphs. The visualizations can be found in Figures 3, 4.

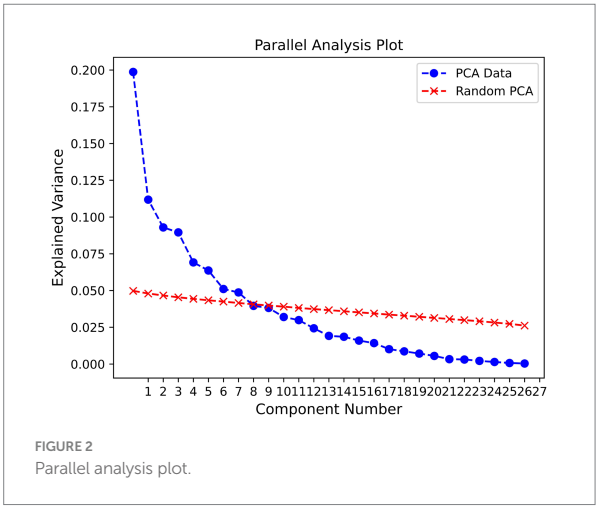


FIGURE 2
Parallel analysis plot.

3.3 Correlation results

Due to the issues with word norm availability, only the first three components were checked for correlations with the emotional norms. The vocabulary of words from the GoEmotions dataset was filtered to remove the words that occur fewer than 50 times and more than 1,000 times in the dataset. From among those, 364 words overlapped with the norm dataset (Bradley and Lang, 1999), which consists of 1,030 words. The scores from the first PCA component achieved a correlation of $r = 0.31$ for valence with $p = 2.48 \times 10^{-9}$. The correlation of the second component and the norms for arousal were found to be insignificant with $r = -0.13$, $p = 0.14$. The third component was also insignificant for its correlation with dominance at $r = -0.02$, $p = 0.68$. As the quality of word vectors is heavily dependent on the amount of text on which they were trained, this analysis was not replicated in the robustness analysis.

3.4 Qualitative words inspection

The external word embedding model (Dadas, 2019) was then used to pick 500 words from the vocabulary, which had the highest cosine similarity with the word "emotion". The numerical representations of words were then subjected to a PCA transformation. Finally, 30 highest and lowest words on each component were extracted (see Table 1). Again, as this analysis is word vector dependent, it was not replicated in the robustness analysis. For this check, we concentrated on the visual inspection of the emotion vectors. The overall positions of the emotion vectors on the PCA dimensions changed only slightly, which we attribute to the lower number of datapoints in the split datasets.

3.5 Robustness check

To analyze the robustness of our analysis we additionally randomly split the dataset into two equal halves and repeated the analysis described in the Method section on these two halves, to

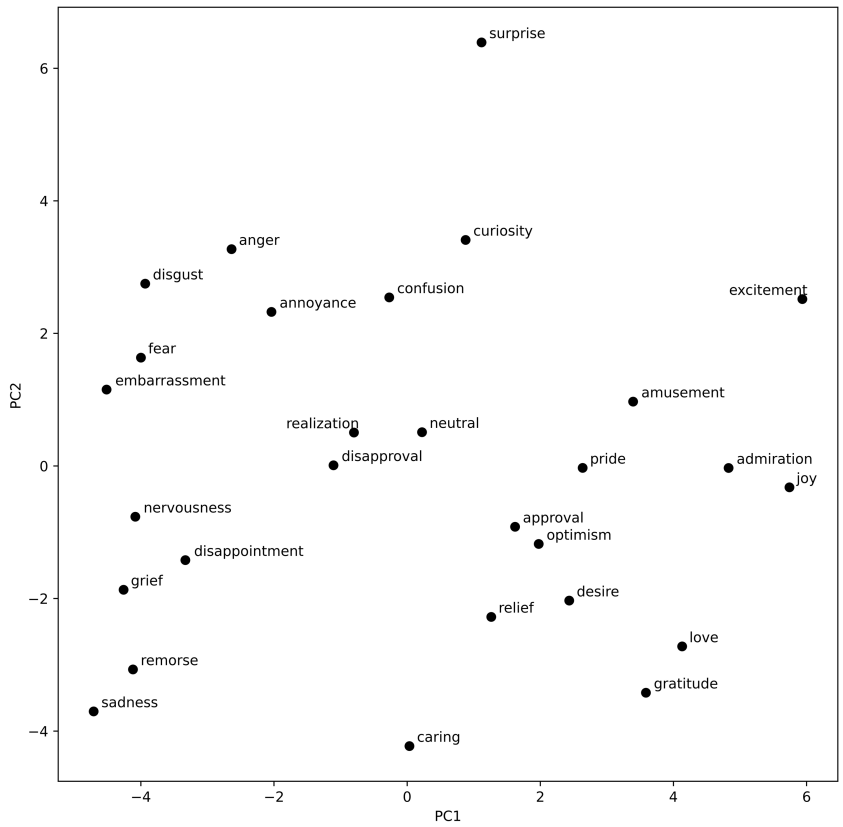


FIGURE 3
Emotional vectors plotted with regard to the first two PCA components.

ensure that similar distributions of emotion vectors are achieved. The overall positions of the emotion vectors on the PCA dimensions changed only slightly, which we attribute to the lower number of datapoints in the split datasets. The full report of the robustness check can be found in Supplementary materials.

3.6 t-SNE components visualization

The results of the t-SNE analysis were plotted in Figure 5.

3.7 The logistic regression

The only significant variable in the regression model was the first PCA component ($\beta = 1.60$; $p < 0.001$; see Table 2).

4 Discussion

The visualization of the emotion vectors (see Figure 3) along the first component complies with the valence negative–positive dichotomy. On the right, there are many high valence emotions such as joy, admiration, excitement, gratitude, love, and amusement. On the left, negative low-valence emotions can be found. These include disgust, fear, embarrassment, nervousness, disappointment, grief, remorse, and sadness. The second component seems to reflect the arousal dimension, with high scores assigned to such emotions as surprise, curiosity, anger, excitement, disgust, and annoyance; and low scores assigned to caring, gratitude, sadness, remorse, grief, and relief. Interestingly, love and desire are also classified among low arousal emotions. This could be an artifact of the nature of the dataset, and the fact that posts classified as love and desire could in many instances relate to those emotions being not satisfied, and thus including words that usually would be associated with sadness, and other low valence, low arousal emotions. Another possibility is that, purely due to the nature of the PCA, the first component does not fully capture the valence spectrum; however, the arrangement of the rest of the emotions enables a partial identification with the valence dimension. The third component (see Figure 4) is a lot less varied, with a lot of emotions clustered in the middle. Considering that it explains less than 10% of the variance in the word vectors, that is to be expected. This component, however, clearly separates such emotions as anger, and annoyance (high dominance) from emotions such as fear, curiosity, and confusion (low dominance). The fourth component, explaining the least amount of variance, could reflect the fourth dimension of emotional experience, namely unpredictability. This is evidenced in the strict separation of curiosity from amusement. However, the emotion of fear does not match this interpretation, and thus, it is not possible to state it with certainty. The distribution of emotion vectors was largely replicated during the robustness analysis for

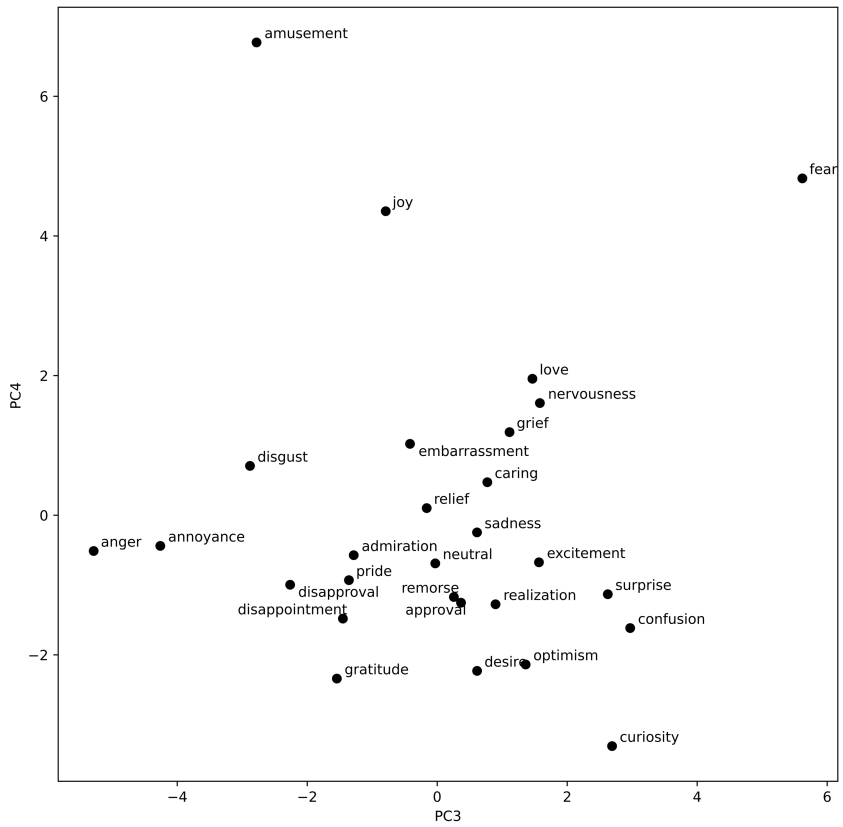


FIGURE 4
Emotional vectors plotted with regard to the third and fourth PCA components.

TABLE 1 Highest and lowest ranked words for each of the PCA dimensions.

PCA dimension	Words
PCA 1 high	together, fun, play, character, music, hate, story, interesting, amazing, album, especially, surprise, would, song, characters, interest, wish, unfortunately, perspective, love, stuff, one, happens, much, learned, ideas, quite, filled, sound, change
PCA 1 low	behavior, somehow, without, nobody, body, pain, almost, meant, happened, cause, clearly, completely, funny, away, humans, wrong, nothing, hurt, brain, others, trust, feel, saying, thinking, situation, someone, caused, truth, honestly, must
PCA 2 high	interesting, religion, seen, actually, basically, picture, crazy, different, even, clearly, wonder, literally, political, talking, beyond, individual, rather, actual, behavior, look, almost, quite, people, irony, pure, another, would, incredibly, power, exactly
PCA 2 low	pain, feel, feeling, appreciate, hear, alone, sometimes, hope, situation, life, better, good, feelings, felt, heart, true, always, everything, laugh, bad, able, thoughts, wonderful, choice, whatever, relationship, focus, anyway, loved, wish
PCA 3 high	story, happen, might, different, interesting, hope, scared, could, someone, something, weird, anyone, hear, totally, happened, never, afraid, crazy, bring, imagine, quite, would, moment, bit, alone, similar, true, surprise, nobody, together
PCA 3 low	give, good, say, literally, trying, opinion, people, words, saying, word, idea, understand, absolutely, mean, bad, play, incredibly, sound, everything, strong, either, power, behavior, move, point, every, reasons, everyone, telling, nothing
PCA 4 high	tell, imagine, moment, sad, sometimes, everyone, kinda, crying, loud, remember, little, fun, somehow, even, angry, someone, funny, feel, triggered, almost, thought, cry, still, tears, honestly, scene, seeing, hurt, scared, turn
PCA 4 low	interesting, change, give, need, opinion, anything, faith, unfortunately, means, different, situation, anyone, given, heard, deal, question, quite, rather, great, yet, something, individual, often, knowledge, hear, move, talent, nothing, however, another

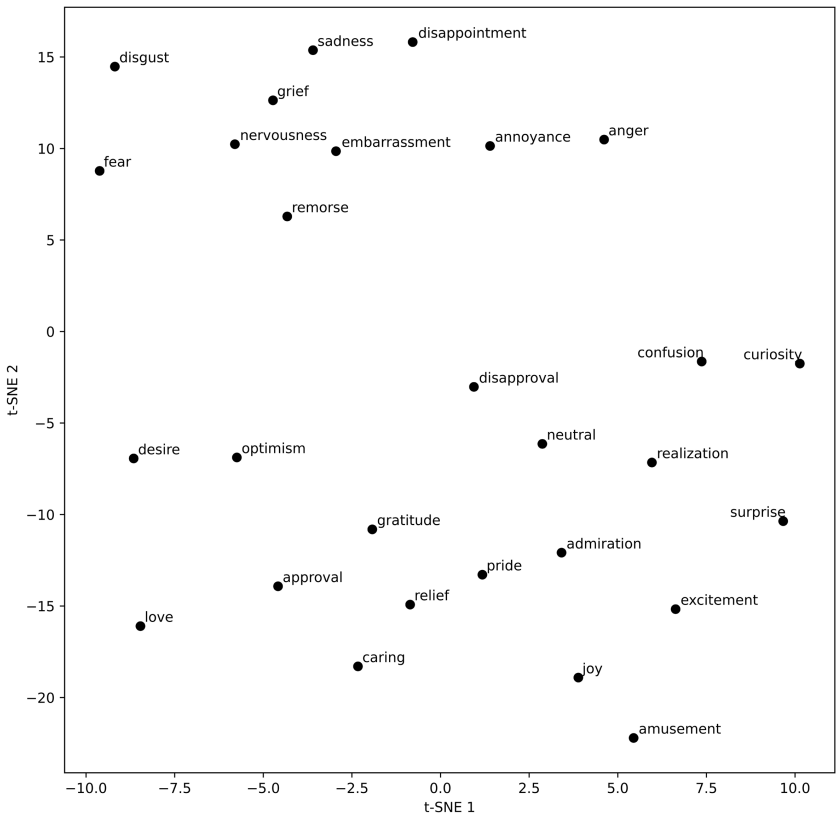


FIGURE 5
t-SNE components.

TABLE 2 Logistic regression results predicting sentiment for texts.

Variable	B	SE	z	*p*	95% CI
Constant	−0.401	0.259	−1.551	0.121	[−0.909, 0.106]
PCA 1	1.599	0.209	7.657	0.000	[1.189, 2.008]
PCA 2	−0.116	0.251	−0.462	0.644	[−0.609, 0.377]
PCA 3	−0.466	0.359	−1.297	0.195	[−1.170, 0.238]
PCA 4	−0.254	0.318	−0.801	0.423	[−0.877, 0.368]

Dependent variable: sentiment. Observations: 155,633. Pseudo R-squared: 0.0003071. Log-Likelihood: −103,800. LLR *p*-value: 4.684e−13.

the first two dimensions (valence and arousal). The last two dimensions were significantly less pronounced, which is most probably the effect of smaller datasets, as each of the two datasets contained only half of all the text available for the primary analysis (see Supplementary materials).

The results of the correlation tests of the placement of words alongside the different components have to be interpreted in the light of the fact that while emotion vectors synthesize the information from many documents, labeled with a given emotion, single word vectors only relate to a limited number of nearest words on each side of the specific token. This results in the word vectors carrying less information and thus not being a robust indicator of emotional expression. Still, even under these strict limitations, the first component achieves a robust correlation with its corresponding norm ($r = 0.31$; $p < 0.001$). It should be noted that the lack of significance of the other components does not prove that they are not related to their corresponding dimensions. This point is underlined by the scarcity of information embedded in their respective word vectors, as well as the fact that only a small portion of words from the GoEmotions dataset actually overlap with the available norms (Bradley and Lang, 1999).

Even though the qualitative inspection of words suffers from the same limitation of word vectors carrying less information than the emotion vectors, some interesting examples that corroborate the correlation between principal components and the dimensions of emotional experience can be found (see Table 1). Scored high on the first component (reflecting valence), are such words as together, fun, play, music, interesting, and love, all of which relate to high valence concepts. On the other side of the same component, there are words like pain, and hurt, both related to low valence concepts. For the

second component, high scoring are words like interesting, crazy, power, and incredibly, which relate to high arousal; low scoring words are feeling, and alone, reflecting low arousal. The high scoring words on the third component are among others: scared, afraid, crazy, and surprise corresponding to low dominance; lower-scoring words on this component are words like absolutely, strong, power, and incredibly, reflecting high dominance (keep in mind that in the case of the third component the factor loadings are stipulated to be negatively related to the dominance dimension; see Figure 4). Finally, the words presented for the fourth component do not seem directly related to the dimension of unpredictability.

The aforementioned words mostly confirm the relation of the PCA components to the emotional dimensions; however, as can be seen from Table 1, not all of the presented words fit this pattern. Examples such as hate for high valence, laugh for low valence, wonderful, and laugh for low arousal do not fit into the outlined interpretation. These outliers could exist both due to the aforementioned problem with low informative value of specific word vectors and due to the specific ways in which they were used in their corresponding posts. Because of the high volume of the dataset, a qualitative exploration of each and every post within which they were found is impossible.

The t-SNE analysis revealed two main clusters of emotion vectors (see Figure 5). One cluster comprises negative emotions such as anger, sadness, disappointment, and remorse. The other cluster includes mainly positive emotions such as admiration, pride, excitement, joy, and amusement, as well as neutral emotions. Interestingly, disapproval, an openly non-positive emotion, is also found in this cluster. This bipolar structure confirms a significant influence of the valence dimension on the semantic arrangement of the emotion vectors. Since t-SNE focuses on preserving pairwise distances between data points (Van der Maaten and Hinton, 2008), it primarily reflects the valence dimension, while the other dimensions identified by PCA are not visible in the t-SNE visualization, as expected. Consistent with t-SNE's objective, emotions with similar meanings and expressions (e.g., desire and optimism; confusion and curiosity; sadness and disappointment) are positioned close to each other. It is important to note that the t-SNE results are sensitive to the choice of hyperparameters. In this analysis, we selected parameters that clearly delineated clusters, but different settings could produce varying results. A comprehensive exploration of all possible hyperparameters is beyond the scope of this paper.

Finally, the logistic regression results indicated that the first PCA component has a significant relationship with the sentiment of the texts ($\beta = 1.60$, $p < 0.001$; see Table 2). This finding further corroborates the conclusion that the first component reflects the valence dimension. Although the amount of variance explained by the model is very low (Pseudo R-squared: 0.0003071), this is expected because the emotion vectors used to create the PCA components were derived from the compressed information of over 50,000 texts, making it impossible to retain all information about every single text. Similar to the situation with the word vectors, the individual texts were only small snippets of the long-concatenated series that generated the emotion vectors.

4.1 Limitations

Our methodology assumes that words surrounding a specific token are indicative of its emotional connotation. However, this assumption does not consider the complexity of language and semantics. The emotional connotation of a word can significantly change depending on its position and usage in the sentence. As a result, single-word vectors may carry less information and be less reliable indicators of emotional expression. This challenge is reflected in our correlation test results, which, while statistically significant, show a relatively low correlation coefficient ($r = 0.31$). While more advanced word embedding methods that consider distant relations between words exist, such as transformer models (Vaswani et al., 2017), they are limited in the length of the text that they can represent, and thus are not sufficient for the current task where long, concatenated texts were analyzed. One possibility of using them is to average the vectors representing texts related to specific emotions, however, due to the noise inherent in this averaging, this method was not chosen for the current study.

Additionally, the interpretations of the third and fourth components of the PCA analysis might not fully correspond to the emotional dimensions of dominance and unpredictability, respectively. The third component was less varied and mainly clustered around the middle, suggesting a limited variability in dominance among the emotions. The fourth component explained the least amount of variance and its link to the dimension of unpredictability was inconclusive, especially given the unexpected positioning of certain emotions such as fear. Furthermore, there were certain word examples that did not fit the expected emotional dimensions, such as 'hate' for high valence and 'wonderful' for low arousal. While we attribute these anomalies to discourse-related artifacts and noise, they may also point to the complexity and multidimensionality of emotions that a linear component analysis may not fully capture. Another possibility points back to the information issues related to analyzing single word vectors, as they carry significantly less information than their emotion vectors counterparts.

From the methodological perspective, the fact that the emotions were labeled by the readers of text, and not their authors, stands in disagreement with the methods of previous studies, which often probed the person who experienced the emotions directly for their descriptions. One cannot expect that in all possible cases the annotator will correctly judge the emotion of the writer, or that the writer will always honestly describe their internal affairs. While the question of whether the influence of these two confounders is strong enough to produce qualitatively different results is an open one, the problem of text-based emotion communication and understanding is important in itself. This is especially true in the current age, where a lot of communication is done through text.

The preset number of emotion labels can also be seen as a limitation in the sense that by using them, the results of the current study will be biased by previous literature that has produced them. On the other hand, if annotators had been asked to describe the emotions in an open-ended manner, their results would still have to be categorized into label-like groups just the same. This grouping would be necessary to bind enough different texts together to produce robust emotion vectors. Drawing from the knowledge generated by previous studies is therefore a defensible alternative.

Finally, it is worth mentioning, that while the research on emotional components has a long history (Gendron and Feldman Barrett, 2009), the current study is to our best knowledge the first attempt at recreating emotional components based on numerical representations of natural language and, as such, is to be viewed as

exploratory research. The findings of this study are best viewed as an invitation to use word embeddings to study psychological phenomena using newer, better-suited methods that allow researchers to analyze qualitative data in a quantitative manner.

4.2 Implications

Despite its exploratory nature, the current study shows that similar emotional components to the ones presented by the previous literature can be extracted from text using word embeddings. Specifically, these components were recovered by triangulating the semantic content of texts sourced from social media with peoples' judgements of what emotion the author of these texts wanted to express (limited to the 28 emotions picked for annotation). Considering the two confounders present—first the willingness of the author to honestly communicate their emotions, and second, the ability of the annotator to correctly gauge what the author wanted to communicate—it is difficult to claim that the topology reported in the current study perfectly reflects the topology of internal emotional experience. Furthermore, given that the annotators were limited in their responses to a preset list of 28 emotions based on psychological literature, this study cannot introduce novel emotional phenomena, as it is constrained to those studied by previous researchers.

However, what this study shows is that the defining dimensions of emotions, as studied through more direct, yet less ecologically valid means of questionnaires and self-reports, are reflected in the semantic structure of how they can be expressed in written language. This can be explained by the process through which our need to communicate our internal states through language shapes and creates language itself. This interpretation aligns with Chafe's work, which emphasizes that the structure and use of language are deeply influenced by the need to communicate conscious experiences and suggests that our expressions in written language naturally reflect the dimensions of internal emotional states (Chafe, 1996, 2013).

This method, when compared to the previous studies which mostly used specialized questionnaires, allows for a more ecologically valid analysis of the core dimensions of emotions. It ensures that the extracted components are grounded in the naturalistic expression of emotions and not artificially constrained by the assumptions of any particular theoretical model (Jackson et al., 2022). However, due to the indirect procurement of emotion labels (through readers and not directly from the authors), as well as the noise present in naturalistic expressions, it does not directly challenge existing methods, complementing them instead.

However, the presence of this noise could shed some light on the differences between the previous studies in the number of components that can be recovered (Bliss-Moreau et al., 2020; Fontaine et al., 2007; Mehrabian, 1996). This is evidenced by the clear dichotomy between fear and anger on the third component, supported in part by the qualitative word inspection, and by the vague sketch of unpredictability on the fourth of the recovered components. Perhaps with cleaner data and higher sample sizes, these components could be systematically recovered using classical methods. Another possibility is that laboratory studies obscure certain dimensions of emotional experience. This could be true especially for the dimension of dominance, the expression of which could be socially undesirable. Here the use of external annotators, rather than the authors of the text becomes an asset as it eradicates the influence of such social undesirability on the effects of the study.

As a last point, it is important to emphasize that the "emotion vectors" discussed in this study are purely mathematical representations derived from word embeddings, capturing the semantic and emotional content of text (Gutiérrez and Keith, 2019; Mikolov et al., 2013a,b). Unlike vectors of force in physics, which have a direction and magnitude related to physical movement, emotion vectors do not directly correspond to any physical or embodied experiences. They are abstract, numerical constructs designed to encapsulate the relationships between words in a multidimensional space, reflecting the latent structure of emotional content in language. This distinction is crucial to avoid conflating these computational representations with the physiological or psychological processes involved in action readiness, which pertains to the body's preparation for specific actions in response to emotions (Frijda, 2010). Nonetheless, this separation does not diminish the potential value of exploring how these numerical representations might correlate with or illuminate aspects of embodied emotions. Future research could delve deeper into this intersection, investigating how emotion vectors could be used to study the embodiment of emotions, perhaps by correlating these computational measures with physiological data or by incorporating word embedding techniques into previous studies that tested the influence of text data on participants' action-readiness (Lewinski et al., 2016). Such explorations could provide a richer, more integrated understanding of how emotions are represented and experienced.

4.3 Future directions

Future studies could try to recreate the current study on additional datasets of comparable quality. This would require researchers to assemble datasets of adequate length and content variance. The task of systematizing such endeavors has not been undertaken yet; however the great work done by Google (Demszky et al., 2020) can offer some directions in that regard. To our knowledge, as of yet, no dataset of comparable quality exists in open access. However, the data itself is available on the Internet, and its size is constantly growing, due to the popularity of social media sites.

Alternatively, recreating this study on a dataset with emotions annotated by the text authors instead of readers, could provide valuable information on the nature of the difference between these two emotional planes. This kind of research could shed more light on the problems related to communicating emotional information over the internet and through other text-based media, with an emphasis on the different sources of noise that partake in this process and can in many cases result in misunderstandings. The method itself can also be extended to different domains of psychology. For example, it could be well applied to the task of reconstructing the components of personality, assuming that the data are found to support this endeavor. Word embeddings can also be used in a completely data-driven way to analyze the results of qualitative interviews and create completely new psychological constructs. Furthermore, the method bypasses the difficulties in analyzing the emotional experience of individuals associated with such limitations as memory bias in answering questionnaires. The possibility of analyzing the text written by a specific individual over a span of time could therefore allow researchers to get a glimpse of what so far has been hidden behind

population-wide studies—the way people express and experience emotions individually.

From a technical perspective, there is a possibility that the method of creating emotion vectors and applying PCA to them with the aim to extract emotional dimension components could be repurposed as a feature extraction method for emotion prediction. Future studies could try to apply similar techniques to this and other datasets and see whether the addition of these extracted features to more advanced machine learning models, such as deep learning architectures, XGBoost, SVM with non-linear kernels, and artificial neural networks (ANNs) leads to improved model accuracy and robustness.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/hplisiecki/emotion_topology.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the [patients/ participants OR patients/participants legal guardian/next of kin] was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

HP: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AS: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Polish National Center of Sciences; under Grant number [2020/38/E/HS6/00302].

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1401084/full#supplementary-material

References

Abdi, H., and Williams, L. J. (2010). Principal component analysis. *WIREs Comput. Stat.* 2, 433–459. doi: 10.1002/wics.101

Al-Amin, M., Islam, M. S., and Das Uzzal, S. (2017). Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words. *2017 international conference on electrical, computer and communication engineering (ECCE)*, 186–190. doi: 10.1109/ECACE.2017.7912903

Barrett, L. F., Quigley, K. S., and Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philos. Trans. R. Soc. B, Biol. Sci.* 371:20160011. doi: 10.1098/rstb.2016.0011

Bliss-Moreau, E., Williams, L. A., and Santistevan, A. C. (2020). The immutability of valence and arousal in the foundation of emotion. *Emotion* 20, 993–1004. doi: 10.1037/emo0000606

Bradley, M. M., and Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Technical report C-1). The Center for Research in Psychophysiology, University of Florida.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., and Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vis. Res.* 41, 1179–1208. doi: 10.1016/S0042-6989(01)00002-5

Chafe, W. (1996). How consciousness shapes language. *Pragmat. Cogn.* 4, 35–54. doi: 10.1075/pc.4.1.04cha

Chafe, W. (2013). "Toward a thought-based linguistics" in Functional approaches to language. eds. S. Bischoff and C. Jany (Berlin, Germany: De Gruyter), 107–130.

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., and Banaji, M. R. (2021). Gender stereotypes in natural language: word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychol. Sci.* 32, 218–240. doi: 10.1177/0956797620963619

Cowen, A. S., and Keltner, D. (2020). What the face displays: mapping 28 emotions conveyed by naturalistic expression. *Am. Psychol.* 75, 349–364. doi: 10.1037/amp0000488

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., and Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nat. Hum. Behav.* 3, 369–382. doi: 10.1038/s41562-019-0533-6

Dadas, S. (2019). *Polish NLP resources* (1.0) [computer software]. Available at: https://github.com/sdadas/polish-nlp-resources (Original work published 2018)

Dellacherie, D., Bigand, E., Molin, P., Baulac, M., and Samson, S. (2011). Multidimensional scaling of emotional responses to music in patients with temporal lobe resection. *Cortex* 47, 1107–1115. doi: 10.1016/j.cortex.2011.05.007

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). *GoEmotions: a dataset of fine-grained emotions* (arXiv:2005.00547). arXiv. Available at: https://doi.org/10.48550/arXiv.2005.00547

Durrheim, K., Schuld, M., Mafunda, M., and Mazibuko, S. (2023). Using word embeddings to investigate cultural biases. *Br. J. Soc. Psychol.* 62, 617–629. doi: 10.1111/bjso.12560

Ekman, P., Davidson, R. J., and Friesen, W. V. (1990). The Duchenne smile: emotional expression and brain physiology: II. *J. Pers. Soc. Psychol.* 58, 342–353. doi: 10.1037/0022-3514.58.2.342

Evmenenko, A., and Teixeira, D. S. (2022). The circumplex model of affect in physical activity contexts: a systematic review. *Int. J. Sport Exerc. Psychol.* 20, 168–201. doi: 10.1080/1612197X.2020.1854818

Feldman, L. A. (1995). Valence focus and arousal focus: individual differences in the structure of affective experience. *J. Pers. Soc. Psychol.* 69, 153–166. doi: 10.1037/0022-3514.69.1.153

Fontaine, J. R. J., Poortinga, Y. H., Setiadi, B., and Markam, S. S. (2002). Cognitive structure of emotion terms in Indonesia and the Netherlands. *Cognit. Emot.* 16, 61–86. doi: 10.1080/02699930143000130

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychol. Sci.* 18, 1050–1057. doi: 10.1111/j.1467-9280.2007.02024.x

Frijda, N. H. (2010). Impulsive action and motivation. *Biol. Psychol.* 84, 570–579. doi: 10.1016/j.biopsycho.2010.01.005

Gendron, M., and Feldman Barrett, L. (2009). Reconstructing the past: a century of ideas about emotion in psychology. *Emot. Rev.* 1, 316–339. doi: 10.1177/1754073909338877

Gutiérrez, L., and Keith, B. (2019). "A systematic literature review on word embeddings" in Trends and applications in software engineering. eds. J. Mejia, M. Muñoz, Á. Rocha, A. Peña and M. Pérez-Cisneros (New York, United States: Springer International Publishing), 132–141.

Imbir, K. K. (2016). Affective norms for 4900 polish words reload (ANPW_R): assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Front. Psychol.* 7:1081. doi: 10.3389/fpsyg.2016.01081

Imbir, K. K., Duda-Goławska, J., Pastwa, M., Jankowska, M., Modzelewska, A., Sobieszek, A., et al. (2020). Electrophysiological and behavioral correlates of valence, arousal and subjective significance in the lexical decision task. *Front. Hum. Neurosci.* 14. doi: 10.3389/fnhum.2020.567220

Islam, M. R., Ahmmed, M. K., and Zibran, M. F. (2019). MarValous: machine learning based detection of emotions in the valence-arousal space in software engineering text. *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 1786–1793. doi: 10.1145/3297280.3297455

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., and Lindquist, K. A. (2022). From text to thought: how analyzing language can advance psychological science. *Perspect. Psychol. Sci.* 17, 805–826. doi: 10.1177/17456916211004899

Jatnika, D., Bijaksana, M. A., and Suryani, A. A. (2019). Word2Vec model analysis for semantic similarities in English words. *Procedia Comput. Sci.* 157, 160–167. doi: 10.1016/j.procs.2019.08.153

Jia, K. (2021). Chinese sentiment classification based on Word2vec and vector arithmetic in human–robot conversation. *Comput. Electr. Eng.* 95:107423. doi: 10.1016/j.compeleceng.2021.107423

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., and Rudzicz, F. (2019). A survey of word embeddings for clinical text. *J. Biomed. Inform.* 100:100057. doi: 10.1016/j.yjbinx.2019.100057

Lampier, L. C., Caldeira, E., Delisle-Rodriguez, D., Floriano, A., and Bastos-Filho, T. F. (2022). A preliminary approach to identify arousal and valence using remote Photoplethysmography. In T. F. Bastos-Filho, Oliveira CaldeiraE. M. De and A. Frizera-Neto (Eds.), XXVII Brazilian congress on biomedical engineering (Vol. *83*, pp. 1659–1664). New York, United States: Springer International Publishing

Le, Q., and Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st international conference on machine learning*, 1188–1196. Available at: https://proceedings.mlr.press/v32/le14.html (Accessed March 10, 2024).

Lewinski, P., Fransen, M. L., and Tan, E. S. (2016). Embodied resistance to persuasion in advertising. *Front. Psychol.* 7. doi: 10.3389/fpsyg.2016.01202

Lin, Y., Hoover, J., Portillo-Wightman, G., Park, C., Dehghani, M., and Ji, H. (2018). Acquiring background knowledge to improve moral value prediction. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 552–559. doi: 10.1109/ASONAM.2018.8508244

Martínez-Tejada, L. A., Maruyama, Y., Yoshimura, N., and Koike, Y. (2020). Analysis of personality and EEG features in emotion recognition using machine learning techniques to classify arousal and valence labels. *Mach. Learn. Knowl. Extr.* 2, 99–124. doi: 10.3390/make2020007

Mehrabian, A. (1996). Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Curr. Psychol.* 14, 261–292. doi: 10.1007/BF02686918

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv* [Preprint]. *arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. Available at: https://doi.org/10.48550/ARXIV.1310.4546

Nicolaou, M. A., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. Affect. Comput.* 2, 92–105. doi: 10.1109/T-AFFC.2011.9

Nowlis, V., and Nowlis, H. H. (1956). The description and analysis of mood. *Ann. N. Y. Acad. Sci.* 65, 345–355. doi: 10.1111/j.1749-6632.1956.tb49644.x

Plisiecki, H., and Sobieszek, A. (2023). Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection. *Behav. Res. Methods* 56, 4716–4731. doi: 10.3758/s13428-023-02212-3

Plutchik, R. (1980). "Chapter 1—A general psychoevolutionary theory of emotion" in Theories of emotion. eds. R. Plutchik and H. Kellerman. (Amsterdam, Netherlands: Academic Press), 3–33.

Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* 17, 715–734. doi: 10.1017/S0954579405050340

Richie, R., Zou, W., and Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology* 5:50. doi: 10.1525/collabra.282

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi: 10.1037/h0077714

Russell, J., and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *J. Res. Pers.* 11, 273–294. doi: 10.1016/0092-6566(77)90037-X

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *J. Exp. Psychol.* 44, 229–237. doi: 10.1037/h0055778

Shaver, P., Schwartz, J., Kirson, D., and O'Connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *J. Pers. Soc. Psychol.* 52, 1061–1086. doi: 10.1037/0022-3514.52.6.1061

Stanisławski, K., Cieciuch, J., and Strus, W. (2021). Ellipse rather than a circumplex: a systematic test of various circumplexes of emotions. *Personal. Individ. Differ.* 181:111052. doi: 10.1016/j.paid.2021.111052

Syssau, A., Yakhloufi, A., Giudicelli, E., Monnier, C., and Anders, R. (2021). FANCat: French affective norms for ten emotional categories. *Behav. Res. Methods* 53, 447–465. doi: 10.3758/s13428-020-01450-z

Tseng, A., Bansal, R., Liu, J., Gerber, A. J., Goh, S., Posner, J., et al. (2014). Using the circumplex model of affect to study valence and arousal ratings of emotional faces by children and adults with autism spectrum disorders. *J. Autism Dev. Disord.* 44, 1332–1346. doi: 10.1007/s10803-013-1993-6

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9:2579–2605.

Van Loon, A., and Freese, J. (2023). Word embeddings reveal how fundamental sentiments structure natural language. *Am. Behav. Sci.* 67, 175–200. doi: 10.1177/00027642211066046

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention Is All You Need. *arXiv* preprint.

Widmann, T., and Wich, M. (2022). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Polit. Anal.* 31, 626–641. doi: 10.1017/pan.2022.15

Wiles, J. A., and Cornwell, T. B. (1991). A review of methods utilized in measuring affect, feelings, and emotion in advertising. *Curr. Issues Res. Advert.* 13, 241–275. doi: 10.1080/01633392.1991.10504968

Woodard, K., Zettersten, M., and Pollak, S. D. (2022). The representation of emotion knowledge across development. *Child Dev.* 93, e237–e250. doi: 10.1111/cdev.13716

Yao, Z., Yu, D., Wang, L., Zhu, X., Guo, J., and Wang, Z. (2016). Effects of valence and arousal on emotional word processing are modulated by concreteness: behavioral and ERP evidence from a lexical decision task. *Int. J. Psychophysiol.* 110, 231–242. doi: 10.1016/j.ijpsycho.2016.07.499

Behavior Research Methods (2024) 56:4716–4731
https://doi.org/10.3758/s13428-023-02212-3

Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection

Hubert Plisiecki[1] · **Adam Sobieszek**[2]

Accepted: 30 July 2023 / Published online: 25 September 2023

Abstract
Data on the emotionality of words is important for the selection of experimental stimuli and sentiment analysis on large bodies of text. While norms for valence and arousal have been thoroughly collected in English, most languages do not have access to such large datasets. Moreover, theoretical developments lead to new dimensions being proposed, the norms for which are only partially available. In this paper, we propose a transformer-based neural network architecture for semantic and emotional norms extrapolation that predicts a whole ensemble of norms at once while achieving state-of-the-art correlations with human judgements on each. We improve on the previous approaches with regards to the correlations with human judgments by $\Delta r = 0.1$ on average. We precisely discuss the limitations of norm extrapolation as a whole, with a special focus on the introduced model. Further, we propose a unique practical application of our model by proposing a method of stimuli selection which performs unsupervised control by picking words that match in their semantic content. As the proposed model can easily be applied to different languages, we provide norm extrapolations for English, Polish, Dutch, German, French, and Spanish. To aid researchers, we also provide access to the extrapolation networks through an accessible web application.

Keywords Affective norms · Transformer-based neural network · Semantic extrapolation · Emotional norms extrapolation · Experimental stimuli selection · Valence and arousal

Affective norms of words have various applications across psychology, linguistics, and machine learning. Their importance is evidenced by the large number of use cases they enjoy. They have been used to select stimuli for experiments in social and affective psychology to investigate behavior (Crossfield and Damian, 2021), to study clinical populations (Williamson et al., 1991; Sloan et al., 2001), and as correlates of brain activity (Citron, 2012; Imbir et al., 2022; Kanske & Kotz, 2007; Yao et al., 2016). Together with norms of semantic dimensions they serve as tools for the study of lexical semantics, concerned with how concepts may be represented in the brain (Binder et al., 2016). Recently, semantic and affective norms have seen a surge in popularity with the growing interest in machine learning, where they have been used to train automatic classifiers of, for example, the sentiment expressed in a given piece of text (Nielsen, 2011). All such uses rely on databases of word – norm pairs, where norms are calculated based on human ratings of the word on a particular dimension of interest (e.g., how positive, or negative a given word is, a technique dating back to the work of Osgood et al., 1957). To this end, measurement scales for various lexical affective constructs have been developed, starting with the simple Likert scale, and continuing with the popular self-assessment manikin of Bradley and Lang (1994).

The two most popular approaches in creating emotional norms include either rating words on emotional dimensions or their association with discrete emotion categories. The most popular of these dimensions in the first approach include valence, arousal, and dominance (Bradley & Lang, 1999; Imbir, 2015; Sianipar et al., 2016; Söderholm et al., 2013; Stadthagen-Gonzalez et al., 2017; Verheyen et al., 2020; Warriner et al., 2013; Yao et al., 2017) and to a lesser extent other dimensions which may modulate emotional processing, such as concreteness, age of acquisition, subjective significance, or origin of emotional load (Brysbaert et al., 2014a, b; Imbir 2016; Kuperman et al., 2012). As for discrete emotional categories, norms are usually concerned

✉ Hubert Plisiecki
hplisiecki@gmail.com

[1] Institute of Psychology, Polish Academy of Sciences, SWPS University of Warsaw, Warsaw, Poland

[2] Faculty of Psychology, University of Warsaw, Warsaw, Poland

with subsets of the six basic emotions: fear, anger, joy, sadness, disgust, and surprise (Mohammad, 2018; Stevenson et al. 2007). The affective norms have been published for many languages other than English (Bradley & Lang, 1999): French (Syssau et al., 2021), German (Võ et al., 2009), Spanish (Redondo et al., 2007), Dutch (Moors et al., 2013), Polish (Imbir, 2016), Turkish (Kapucu et al., 2021), Italian (Montefinese et al., 2014), Portuguese (Soares et al., 2012), Greek (Vaiouli et al., 2023), and Chinese (Yao et al., 2017).

Some applications of affective norms, such as complex experimental designs, demand very large datasets, the creation of which can be prohibitively expensive. This demand has been partially satisfied for English words, with the expansion of the classic Affective Norms for English Words database (ANEW; Bradley & Lang, 1999) from 1035 to 13,915 by Warriner et al. (2013). However, such large dataset expansions are still unavailable for many languages. Moreover, even in English, the existing norms may not be enough for certain types of studies, which could require norms for *all* English words. Interesting examples of such include analyses of large-scale trends and shifts in the use of language across thousands – if not millions of texts (Kim & Klinger, 2019), where an accurate assessment may require a rating for each word to avoid bias (Snefjella & Blank, 2020). It is in this context that *lexical norm extrapolation* techniques start to be developed, as they allow researchers to use existing norms to expand the database lexicon by predicting the norms of previously unrated words.

Affective norms extrapolation

How does one go about determining the emotionality of words without any human judgment information? A first intuition may be to say that the emotional load of a word can be approximated with the emotional load of another, *similar*, word for which we possess affective norms. This indeed turns out to be the basis for most published norm extrapolation techniques, the difference being mostly in the level of sophistication with which the similarity metrics are defined and the introduction of ways to decrease noise by averaging across many similar words. A popular source of similarity metrics comes from the linguistic distributional hypothesis, which states that words that occur in similar contexts tend to have similar meanings (Boleda, 2020). Thus, the dominant approach in affective norms extrapolation is usually to average a norm of interest across a word's k-nearest neighbors based on a co-occurrence metric, which the review by Mandera et al. (2015) deemed the most effective method as of 2015. An early example of such an approach includes Bestgen and Vincze's (2012) use of latent semantic analysis to derive similarities between words, where co-occurrence is calculated from paragraphs of large language corpora. While varying the number of neighbors to average across, they found the highest correlations between human ratings and their estimates to be $r = 0.71$ for valence, $r = 0.56$ for arousal, and $r = 0.60$ for dominance.

More recent approaches used in machine learning for sentiment analysis employ similarity metrics based on distances in a vector space, where words are represented as points (e.g., Munikar et al., 2019). These spaces, called *word embeddings*, are intended to be lower-dimensional representations of the relationships between the words in language corpora and are created in various ways, which include dimensionality-reduction techniques on the co-occurrence matrix (e.g., multidimensional scaling) and the use of neural networks (e.g., in word2vec; Mikolov et al., 2013).

A different approach has been employed (to great effect) by Vankrunkelsven et al. (2015). Their method involves using a vast dataset of word associations (De Deyne et al., 2013), which are based on 70,000 participants reporting their three associations with one of 12,000 cue words. Since free associations are often based on semantic relationships with the cue word, these data can be used to construct a great similarity metric. Vankrunkelsven et al. used multi-dimensional scaling to construct word embeddings based on this data and achieved correlations of $r = 0.89$, $r = 0.76$, $r = 0.77$, $r = 0.67$, and $r = 0.81$, for valence, arousal, dominance, age of acquisition, and concreteness, respectively. As semantic norm extrapolation is most useful for languages where access to such data is limited, the need to collect vast word association data to perform norm extrapolation seems, while elegant, to be of limited practical utility. Still, a comparison by Vankrunkelsven et al. (2018) of their association-based method with previous methods based on co-occurrence shows that their method achieved state-of-the-art results for the time.

A challenge for all such extrapolation methods, recently presented by Snefjella and Blank (2020), posits that researchers may, however, be overestimating the accuracy of their methods of norm extrapolation by relying on cross-validation to evaluate performance. This is because the words that are missing from the norm databases (in which we are ultimately interested in extrapolation) are not missing at random from all possible words. They suggest considering norm extrapolation as missing data imputation.

Advances in neural network-based language models

In recent years, the field of computational linguistics has been taken by storm with rapid developments in neural network-based models. especially large language models, one of the most notable ones is called GPT-3 and was trained on a corpus of text comprising nearly 500 billion words (Brown et al., 2020). The performance of this model varies, as it has been shown to perform with near human ability on many high-level

tasks like imitating an author or waxing philosophical while failing at several very simple tasks like multiplying large numbers (Elkins & Chun, 2020; Sobieszek & Price, 2022). It is, however, worth stressing that on many occasions its performance is indistinguishable from that of a human and that this high performance is not merely the product of the sheer size of the training dataset. GPT-3, as well as its predecessors GPT-2 and GPT, utilize a specific machine learning architecture called *attention*, which allows them to attend to many distant words at once, thus being able to grasp complicated contextual information when analyzing text (Brown et al., 2020).

These developments lie in contrast to the previous attempts at text classification, translation and production in natural language processing, as previous models were very limited in their scope when it comes to attending to distant words. The earliest approaches utilized the already-mentioned word embeddings, which were heavily dependent on closely co-occurring words, and thus were unable to capture relations between distant signs (Almeida & Xexéo, 2019). The situation improved with the rise of recurrent neural networks such as LSTMs (Long Short-term memory modules) which tried to retain significant textual information as they gradually advanced through the lines of text (Yu et al., 2019). Unfortunately, these approaches were plagued by the problem of vanishing gradients, resulting in the retained information being lost over time as the activation progressed through the network (Hochreiter, 1998). They were therefore short-sighted. Afterwards came convolutional networks (CCN), which compressed contextual information using sliding windows, thus capturing their contents (Yin et al., 2017). These had a different flaw, however, as to capture complex contextual information one had to apply very large sliding windows (buffers for word compression), and many of them – which was incredibly costly in terms of computational power (Vaswani et al., 2017).

Finally, the concept of "attention" was introduced, which is in simple terms the process of weighing the inputs to a neural layer with the use of trainable weights. This method was first applied in recurrent neural network-based sequence-to-sequence models, which iteratively passed the generated sequence through a generator module to obtain the next word, appended the word to the generated sequence and repeated the process until the whole sequence of interest was generated. The real breakthrough, however, came when the recurrent network architecture was replaced by attention. This feat, achieved by Vaswani et al. (2017), morphed into a family of models called transformers, some among which are BERT, RoBERTa, and XLM (Devlin et al., 2018; Conneau & Lample, 2019; Liu et al., 2019).

Transformer models consist of two modules, an encoder, and a decoder. As this architecture was primarily designed for the task of language translation, we will use the task of language translation as a reference when explaining its mechanism. In simple terms, a sentence in the 1st language is given to the encoder, which transforms it into a numerical representation using attention and feedforward layers. At the same time, a similar thing happens in the decoder, where a corresponding sentence in the 2nd language (with blank "masks" instead of the words that the architecture is meant to predict) is also transformed into a numerical representation, using similar transformations. Then, the output of the encoder is passed to the decoder, where it is concatenated with the numerical representation of the L2 sentence and together they are passed through additional attention and feedforward layers. The final output is compared with the intended output, and the weights on each of the layers are updated to minimize the error between the two (Devlin et al., 2018). Once the model is trained, the encoder can be extracted from the model and be used to obtain rich, contextual text embeddings that can be used for further training (Munikar et al., 2019). The usefulness of such text embeddings stems from their ability to quantitatively describe the embedded text on meaningful dimensions the model discovered during training. A common example is the ability to use the vector in the embedding space corresponding to a given word as a direction in which one may manipulate the embedding of another word, e.g., find the representation of the word 'man' by subtracting the representation of 'royal' from that of the word 'king' (Ethayarajh, 2019).

In summary, the evolution of computational linguistics has led to the development of attention-based transformer models, such as GPT-3, which outperform their predecessors such as LSTMS and CCNS in processing distant words in text, with their high performance attributed not only to the vast training datasets but also to their ability to retain complex contextual information, a characteristic lacking in earlier models due to issues like vanishing gradients and computational costs.

Word stimuli selection

We know from various studies of word processing that a multitude of factors by which we can describe words influence neural and behavioral responses in experiments using words as experimental stimuli. These include accessible features like length and frequency in the language (Hauk & Pulvermüller, 2004; Kuchinke, et al., 2007; Méndez-Bértolo et al., 2011), but also features which are not easily accessible, such as differences in semantic features and emotive content (e.g., abstractness, valence, arousal; see Citron, 2012 for a review). Here, emotional databases are an important asset, as they enable word stimuli selection for experimental manipulation and control of such factors. There however remain challenges to valid stimuli selection based on available datasets. The first stems from recent striking results that even newly discovered emotional dimensions can influence behavior with effect sizes comparable to those previously reported in the literature for valence and arousal (e.g., origin

and subjective significance in experiments of Imbir et al., 2020, 2021, 2022), as well as known factors which were until recently not controlled in such studies (e.g., concreteness in experiments of Kanske & Kotz, 2007). The existence of such unaccounted-for dimensions may explain the disparity of reported results on the influence of emotional factors on behavior (such as those in the review by Citron, 2012) especially when stimuli lists are short and selected from a limited dataset. For this reason, a method of stimuli selection that would more likely produce valid stimuli lists may need to somehow reduce the influence of these unknown factors.

When constructing a manipulation of, for example, valence using emotional norms, we pick sets of negative and positive words that do not differ on some control dimensions (e.g., length, frequency, and arousal). To add an element of unsupervised control (control of unspecified factors), we propose to additionally perform semantic matching between these conditions, which involves selecting words for these groups containing words paired on their semantic features while differing in the manipulated factor. Words similar in meaning and used in similar contexts are more likely to have similar values on dimensions we did not explicitly control, such as imageability, than a random word pairing. An example of such a pairing may be the positive "peaceful", with the negative "boring". Both have low arousal, but even more importantly, both are approximately matched in their semantic content and connotations, while differing in the sign of the emotion attributed to the situation. This is indeed an example of a pairing found by our *stimuli descent* algorithm, which is introduced in Study 3.

Study 1: Transformer-based norm extrapolation

We hypothesize that the use of highly contextual representations of words as input to a model trained to predict the emotional norms will be able to outperform the previous approaches in norm extrapolation. While some of the previous attempts at this task also relied on machine learning to extend affective norms, they all relied on word embeddings (e.g., Mandera et al., 2015) and thus were unable to capture the sophisticated contextual relations among distant words. Furthermore, contrary to word embeddings, the numerical representations obtained from transformers are flexible with regard to the task that they are trained on. For example, a transformer can be first trained to simply generate sentences but then retrained on a different, more specialized task such as emotion recognition in Twitter posts. This retraining will lead to slight changes in the numerical representations generated by the transformer, as the emotional information present in the relation between Twitter posts and their training labels (e.g., happy, sad etc.) will seep into the weights of the model, crystallizing specialized affective knowledge. A good example of such a model is ERNIE, which was trained on several different tasks and achieved state-of-the-art results on several NLP benchmarks at the time of publication (Yu et al., 2019). The use of models pre-trained on emotion recognition-related tasks should therefore further increase the performance of our approach. Additionally, since the transformer models have been trained for many different languages, our models will help researchers from different countries to extrapolate their norm datasets cheaply and accurately.

In the manuscript, we first provide a detailed description of the norm datasets that will be used to train the model. Afterwards, we detail the model's architecture explaining the intuition behind choosing the right transformer module for the task. Then we describe the training regimes and present results comparing the outputs of the models to a specially designated part of the original datasets, ending with a discussion.

Method

Linguistic materials and data curation

In the current study, we make use of the ANEW corpus (Bradley & Lang, 1999), and a corpus collected by Warriner et al. (2013) to train and test our model. The former consists of 1030 words with accompanying rater-based metrics for valence and arousal. The latter has 13,915 words with both previous metrics and, additionally, the age of acquisition and concreteness metrics. All the metrics have been normalized to range from 0 to 1. In line with the argumentation from previous work on these corpora, we used the ANEW words for the test set, subtracting them from the training set composed of all the words present in Warriner's database. The test set, therefore, consists of 983 words. The rest of Warriner's corpus (12,885 words) was divided into training and validation sets at a 9-to-1 ratio Table 1 and 2.

To expand our model to other languages, we make use of five additional datasets. For the Polish language, we employ a norm repository of 4900 words (Imbir, 2016). For Spanish, the dataset contains 1400 words (Redondo et al., 2007). For German we use the BAWL-R dataset with 2902 words (Võ et al., 2009). For French we use the FANCat dataset with 1031 words (Syssau et al., 2021) Finally, we employ norms for 4299 words in the Dutch language (Moors et al., 2013). Unfortunately, all of the same metrics were not available for all of the different languages. While fewer dimensions were available for Spanish and Dutch, the contrary was true for Polish, where we were able to use an eight-metric database. The availability of the metrics in each of the languages can be checked in Table 3. The metrics were normalized to the 0 to 1 range, and the datasets were split for training, validation and testing according to the 8:1:1 ratio.

Table 1 The specification details of the different models

Specifications	English	Polish	Spanish	Dutch	German	French
Transformer encoders	ERNIE 2.0 (Yu et al., 2019)	RoBERTa-Polish (Dadas, 2020)	"bert-base-spanish-wwm-cased" (Perezrojas et al., 2020)	"bert-base-dutch-cased" (de Vries et al., 2019)	"bert-base-german-uncased" (von Platen, 2021)	"french_toxicity_classifier_plus_v2" (Stakovskii, n.d.)
Mean number of raters	28	50	21	63	Not reported	36
Number of words	13,915	4900	1400	4299	2902	1031
Learning rate	5e-5	5e-4	5e-4	5e-4	5e-4	5e-4
Dropout	0.1	0.2	0.2	0.2	0.1	0.1

Model architecture

The proposed model maintains the same architecture across all languages and norms, with the only variation being the transformer embedding model used for each language. As transformer models are usually trained for singular languages at a time, we cannot use a model that uses all languages at the same time. Beyond issues of accuracy, this could pose troubles related to the differences in norms between different languages (Pires et al., 2019). To facilitate the use of our architecture in new languages, the selection of the embedding model is explained in Appendix 1. We only use the encoder from the transformer model as the base encoding layer for our model, and we build additional layers on top of it. The added layers consisted of a single fully connected layer with layer normalization and another layer with one number as its output. On top of that, we have applied a sigmoid activation function, which ensures that the output of our model yields a normalized value between 0 and 1. This type of regression head was added for each of the predicted metrics.

Because models of similar infrastructure can be trained for different languages given enough training data, they can be fully substituted for each other, and similar models can be trained using them. This makes the proposed architecture versatile and open to being implemented in different languages. Transformer models have already been trained in many different languages and are freely available online (Hugging Face, n.d.). Therefore, for most of the languages, the only thing needed to prepare a similar model is a dataset with affective measures.

However, choosing the right transformer for the base of the model is not as straightforward. Wherever possible, we have opted for models that either had more parameters and could therefore model language more accurately, or were pre-trained on emotion recognition tasks. However, the scope of our search was limited by both hardware constraints and model availability. Since different transformer models are pre-trained on various tasks, their performance on a specific task like ours may vary. If researchers want to train a similar model for a language that is not covered in our article, we advise them to run tests using all the different transformer models available in their language, until they find the one that rears the best predictions (see Appendix 1 for more information).

The specifications for each of the models are shown in Table 1. The hyperparameters for our machine learning model were chosen according to common practices, wherein we used a combination of domain knowledge, model complexity considerations, and computational efficiency to guide the selection, minimizing the risk of overfitting and ensuring optimal performance.

Table 2 Correlation results for the past extrapolation models

Study	Valence	Arousal	Dominance	Concreteness	Age of acquisition
Current Study	**0.95**	**0.76**	**0.86**	**0.95**	**0.85**
Vankrunkelsven et al., 2018	0.86	0.69	0.75	0.87	0.59
Vankrunkelsven et al., 2015	0.89	0.76	0.77	0.81	0.67
Mandera et al., 2015	0.69	0.60	0.48	0.80	0.72
Recchia and Louwerse, 2015	0.74	0.75	0.62	-	-
Bestgen and Vincze, 2012	0.71	0.56	0.60	-	-

The best results for a certain metric are in *bold*. Lack of prediction for a certain metric is signified by a *dash*

Behavior Research Methods (2024) 56:4716–4731 4721

Each of the models was trained for 1000 epochs with early stopping (stopping the training before the model starts overfitting the training data). This was implemented by saving the model that had the best correlations with the validation metrics. We used the AdamW optimizer algorithm with an epsilon value of 1e-8, a weight decay of 0.3, amsgrad, and betas equal to (0.9, 0.999). Additionally, we implemented a warmup algorithm, which gradually elevated the learning rate for 600 learning steps, when it reached the maximum number, and then slowly lowered it, until the end of the training. The rest of the specifications like the learning rate and the value of the dropout can be found in Table 1 as it was language specific.

Results and discussion

The English model's predicted affective norms achieved the following Pearson correlations with human judgements on words from the test set: valence: $r = 0.95$, arousal: $r = 0.76$, dominance: $r = 0.86$, age of acquisition: $r = 0.85$, concreteness: $r = 0.95$, with an overall loss of 0.003. When compared to the previous methods (see Table 3), the present approach achieves the highest accuracy across all variables. This is true even compared to Vankrunkelsven et al. (2018) results after they have been adjusted for attenuation ($r = 0.91$, $r = 0.83$, and $r = 0.85$ for valence, arousal, and dominance respectively). The transformer-based model has therefore been shown to achieve higher accuracy when compared to the extrapolations reported based on LSA, and other word embedding methods (Bestgen & Vincze, 2012; Recchia and Louwerse, 2015), to methods based on human word association data (Vankrunkelsven et al., 2015), those based on simple machine learning methods (Mandera et al., 2015), as well as those combining the last two (Vankrunkelsven et al., 2018). It is worth pointing out that direct comparison was not always possible as the past models did not utilize the same high-quality validation set – the ANEW corpus. However, the improvement over those that did use it is so big that the change in the test set most probably would not change the overall conclusion.

Given the very high observed correlations we can compare their values to the theoretically highest correlation values we can expect for the norms of the test set. The uncertainty associated with a prediction may be broken down into epistemic and aleatoric uncertainty. The former concerns the model shortfall that may be amended with better models, the latter concerns the uncertainty inherent to the studied phenomena. For norms, we may estimate the aleatoric uncertainty from the reported standard error associated with the number of raters and the variance of their judgments. With this information we can calculate the limit on prediction performance, which is reported in Fig. 1. In the next section, we discuss other sources of aleatoric uncertainty that limit extrapolation performance. Given these results, along with improvements of around $\Delta r = 0.1$ on every metric, we can safely assume that the transformer model constitutes the current state of the art in norm extrapolation. Furthermore, due to the high popularity of transformers, the current architecture can be easily adapted to different languages. This is evidenced by the results achieved on the subsets of Polish, Spanish, German, French, and Dutch words, most of them being very high (see Table 2). The ease with which the current approach can be adapted to extrapolate word norms in different languages is an improvement on the previous methods, most of which relied on human-based word associations data (Vankrunkelsven et al., 2015, 2018), which is not freely available for most languages.

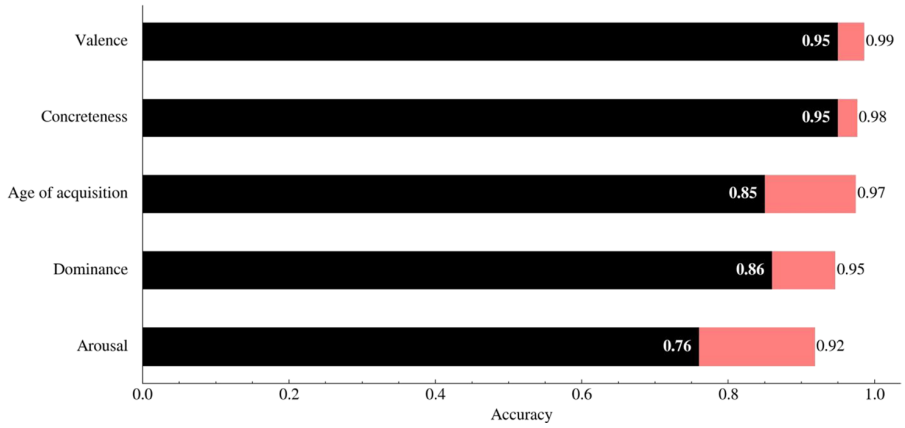
Thanks to its high accuracy, the current model can be used to provide approximations of human judgements in contexts where the actual norm values do not need to be known with precision. The extrapolated norms should not be used as independent variables in linguistic studies of large corpora. As we will investigate in the next section, the predictions could be biased towards uncommon words that

Table 3 Correlation results of affective metrics from the three additional languages on the test set

Affective metric	English	Polish	Spanish	Dutch	German	French
Valence	0.95***	0.93***	0.89***	0.87***	0.8***	0.8***
Arousal	0.76***	0.86***	0.80***	0.80***	0.7***	0.77***
Dominance	0.86***	0.92***	-	0.75***	-	-
Concreteness	0.95***	0.95***	0.89***	-	-	-
Age of Acquisition	0.85***	0.81***	-	0.82***	-	-
Origin	-	0.86***	-	-	-	-
Significance	-	0.88***	-	-	-	-
Imageability	-	0.88***	0.86***	-	0.82***	-
Familiarity	-	-	0.71***	-	-	-

* $p < 0.05$** $p < 0.01$*** $p < 0.001$

Lack of the prediction of a certain metric is signified by a *dash*



Note: Average cross-validated model performance (black bars) compared the expected correlation for a perfect estimator (indicated by the end of the red bar). The length of the red bar indicates the model shortfall.

Fig. 1 The comparison of the accuracy achieved by the model to the perfect estimator correlations. *Note:* Average cross-validated model performance (*black bars*) compared the expected correlation for a perfect estimator (indicated by the *end of the red bar*). The length of the red bar indicates the model shortfall

appear in corpora but are unlikely to be used in experimental research. In cases where precision is needed, the model can still serve as a useful heuristic tool to identify words with specific values, which can be in turn verified experimentally. For example, the model can help identify words whose specific combination of affective metrics are rare, like those with neutral valence and high arousal, to help populate specific stimuli sets, as such words are useful in studies which try to orthogonally manipulate emotional dimensions.

The possibility of using transformer-based extrapolation for the task of finding rare emotion-combination words is important, as k-nearest neighbor approaches often generalize worse in sparse regions, such as in these specific configurations of affective dimensions. If this generalization performance is confirmed it would establish a heuristic search for low-density words and extrapolation for unusual, low-frequency words as unique use cases of our method. However, we need to tackle a significant issue related to measuring extrapolation performance, highlighted by Snefjella and Blank (2020). This problem becomes known when we view norm extrapolation as a missing data problem. In short, the use of supervised learning to impute missing data, conditional on observed data is equivalent to single regression mean imputation, which is an imputation method known in the field of causal inference to produce biased estimates of accuracy via cross-validation (Van Buuren & Groothuis-Oudshoorn, 2011). What Snefjella and Blank (2020) rightly point out is that the set of words that do appear in norms databases is not random nor representative of all words in a language, creating a "missing not at random" problem in norm extrapolation. Words that are longer, less common, or more abstract all have a lower probability of appearing in norms databases, thus also the test set, resulting in accuracy bias. In Study 2 we use these insights to test how biased is the cross-validation estimate of accuracy, as well as how good our transformer-based method is at prediction generalization on different emotional dimensions. To this end, we conduct tests both with normative and new experimental data.

The results of this section, while showing promise in the ability of the model to generalize to unseen words, suggest it cannot overcome the issues with norm extrapolation to all words highlighted by Snefjella and Blank (2020). For a large number of words, the large aleatoric uncertainty in norms and the systematic bias from mean imputation largely coincide, causing an irreducible error that prohibits valid prediction with ad hoc methods. Additionally, the systematic bias in extrapolated norms can hide some complex relationships between the examined variables, which prohibits the use of extrapolated norms in corpora studies and demands researchers experimentally verify the norms in orthogonal designs.

Study 2: Evaluating robustness in out-of-distribution prediction

The goal of this section is to test how well the transformer-based extrapolation method generalizes under selection bias and assess which words' prediction is affected by the "missing not at random" problem. Recall that prediction uncertainty may be divided into epistemic and aleatoric uncertainty. The former is associated with the quality of the model, and the latter with the variability inherent to the studied phenomena. The use of norm extrapolation should be limited to words for which the prediction error is small – first to words that have

Behavior Research Methods (2024) 56:4716–4731 4723

low aleatoric uncertainty (for which prediction is possible), and next to words for which the epistemic uncertainty is low, which depends on the extrapolation method. Assuming cross-validation performance applies to all words runs into discounting both the missing not at random problem and the existence of words for which the concept captured by the norm does not apply the same way as to words in the dataset. To understand the latter point, take the norms for age of acquisition. To obtain the norms, Kuperman et al. (2012) asked participants to answer at what age they thought they had learned a given word. However, more than 50% of the words were not known by all respondents, which for this measure would imply the age of acquisition was larger than the age of the participant making the norm calculated only on respondents that knew the word biased downwards. Here, therefore, we encounter the first limitation of our method, as age of acquisition norm extrapolation should not be used for large values. We conduct three tests for other metrics to assess the two sources of additional prediction error: (a) stemming from the larger aleatoric uncertainty of words unlikely to appear in the test set (b) stemming from a larger epistemic uncertainty in out-of-distribution prediction.

We start by testing whether our method produces biased results under a meaningful selection bias. Abstract and concrete emotional words are processed differently by people, mediating the effects of valence and arousal on reaction times and the neural response (Kanske & Kotz, 2007). Concreteness is thus a good candidate for a factor that may bias results if selected not at random. We test the robustness of our method to additional systematic sampling bias by predicting norms of abstract words using a model trained only on concrete words. Next, we take advantage of English concreteness norms existing for a much larger set of words than the set of words in the test and training sets and establish how accuracy decreases for words known by fewer people. Lastly, we collect additional norms for words chosen entirely at random to find a lower bound for unbiased accuracy across the entire language.

Method

Design and linguistic materials

To test the generalization performance of our method, which the extrapolation methods need for accurate out-of-distribution prediction, we train a model within an artificially imposed selection bias. The English corpus used to train the original English model was resampled to include only words with a concreteness value above the center of the distribution, corresponding to a prediction of 0.5 in the original model (where low values are abstract), leaving 6307 words for training. We then constructed two test sets. The first included only highly abstract words (with concreteness < 0.5; 364 words), simulating the prediction of norms outside the database. The second contained a randomly selected test set from all words, matched on word length with the former. The models were trained in the exact same way as the English model in Study 1.

The norms for abstractness were taken from the dataset of Kuperman et al. (2012), which has two important features: (a) it contains an extremely large, compared to other databases, selection of 40,000 English lemmas, (b) the authors report the proportion of participants that knew the word. In the dataset, there were 27,000 single-word lemmas. We test whether there is a drop in accuracy compared to the accuracy calculated with cross-validation on words from Warriner et al. (2013). Second, we test how the accuracy decreases when decreasing the amount of people familiar with the word. General familiarity is strongly associated with the probability of being included in normative databases, as unknown words are not only hard to obtain norms for but also are not useful for experimental studies.

For the experimental validation, a set of completely random Polish words conditional on not appearing in normative norms database was created. First we gathered a list of all words that appeared at least five times in Polish language corpora (following Kazojć, 2011) and obtained a set of 31,967 words. From this set, we have randomly drawn 200 words. The first 150 words that did not appear in the polish norms database were chosen to be rated on five dimensions – valence, arousal, dominance, imageability, origin – 30 different words per dimension. The last dimension is unique to the Polish norm dataset and refers to the origin of emotional load from either more automatic or reflective processes. The choice to rate a different set of words for each dimension will bar us from inferring which dimension suffers the most from out-of-distribution prediction, but it will give us a more accurate estimate of the performance drop-off for all dimensions (as the same words for each dimension would make the drop-off more of a function of the particular word selection).

Participants

The validation study included 89 Polish-speaking participants (47 women, 42 men). The participants' mean age was 22.6 (SD = 4.1). The study was promoted on Facebook, specifically targeting college students, as the original ANEW Polish norms (Imbir, 2016) were rated by students from this demographic. To stimulate participation, we held a drawing for a 50 PLN reward, which participants were eligible for upon completion of the study. Ratings from 66 participants (50% male) entered into the analysis after removing participants whose answer's reliability was smaller than 0.8.

Procedure

The study was conducted through Qualtrics. We aimed to replicate the most relevant aspects of the procedure used in the

normative study by Imbir (2016). Each participant was given two emotional dimensions to assess. Before each dimension started, participants read a detailed description of the dimensions and saw a scale taken from the normative study. Participants rated words on a nine-point and answered a yes /no "Do you know this word?" question. Participants rated 30 words for each of the two dimensions, five words were repeated to assess rater reliability. This procedure differed from the normative study in that the normative study was done with pen and paper on a list of words, whereas the words appeared individually (on different questionnaire web pages) in our online study.

Results and discussion

We calculated accuracy metrics on out-of-distribution words (e.g., accuracy of valence prediction for highly abstract words, when the model was only shown concrete words during training), and the test set drawn from all words for the four affective metrics which were not manipulated. Below, the results are shown and compared to the accuracies of the original English model in Table 4. Again, all correlations were significant with $p < 0.001$. The result shows that the transformer-based predictions generalize completely over concreteness, only with regards to the metrics of arousal and age of acquisition, with a drop in correlation of results ranging from $\Delta r = 0.1$ in the case of arousal to $\Delta r = 0.15$ in the case of the age of acquisition. Valence and dominance were predicted with exactly the same accuracy in both conditions.

We estimate the correlation with human judgments on a large set of concreteness norms, none of which were included in the training set. We observe that the accuracy of our model decreases slightly on out-of-distribution words and when norms of words known by fewer people are included (shown in Fig. 2, all estimated correlations were significant with $p < 0.001$). First, estimating accuracy on all 6000 words, known by all participants among words not included in the training set, yields a correlation of $r = 0.91$, slightly lower than the original test set correlation of $r = 0.95$. The correlation decreases by around 0.03 to $r = 0.875$ for all words in the concreteness norms dataset, which includes words known by at least 85% of participants.

In the experimental validation, the words were chosen at random conditional on not being included in the emotional norms database. Mean ratings were calculated as an average of the mean ratings within each gender to control for the imbalanced participant sample within each rated dimension. Descriptive statistics for all words are available in the Supplementary Materials.

Every time new norms are collected, we expect to see a larger error caused by regression to the mean, a different participant sample, rating procedure, word selection, a finite number of participants, as well as the change of emotional load of words over time (e.g., the word "pandemic"). We can quantify the change in correlation due to additional noise as the percentage change (1-r2/r1), where r1 is the original correlation and can r2 is the experimental correlation. Averaging over the five variables (see Table 5), the average drop in accuracy on out-of-distribution words equaled $\Delta r = 11\%$ (95% CI [5%, 19%]). Even for the worst value inside the confidence interval, less than 20% of the accuracy is lost on out-of-distribution words. Note that one should not compare the drop in accuracy between dimensions reported here, as each experimental accuracy was obtained on a different set of words.

Study 3: Stimuli descent algorithm

To demonstrate the utility of neural norm extrapolation, we propose a method leveraging the differentiability of the neural network that predicts emotional norms to select words for experimental stimuli. Here, we aim to manipulate certain emotional factors while controlling others. The algorithm, which we call *stimuli descent*, has the ability not only to control specified factors but to control other, unspecified semantic properties of words in an unsupervised way. To understand how this may be achieved, recall how word embeddings describe words on meaningful dimensions the model discovered during training. This fact is utilized, for example, in the wide use of the distance in word embedding space as a measure of semantic similarity (Kenter & de Rijke, 2015; Sitikhu et al., 2019). Thus, words that are close together tend to have similar meanings, which is the

Table 4 The results of the concreteness dependent missingness robustness check on the test set

Condition	Valence	Arousal	Dominance	Age of acquisition
Original model accuracy	0.95***	0.76***	0.86***	0.85***
Random test set	0.94***	0.76***	0.86***	0.86***
Test set of abstract words	0.94***	0.67***	0.86***	0.71***

* $p < 0.05$** $p < 0.01$*** $p < 0.001$

The original model accuracy relates to the results presented in Study 1. The concreteness manipulation relates to the accuracies of the model trained on high concreteness words and tested on low concreteness words. The comparison dataset relates to a model trained on a random sample of original words, keeping the length of the words in each of the datasets the same as in the case of the concreteness manipulation

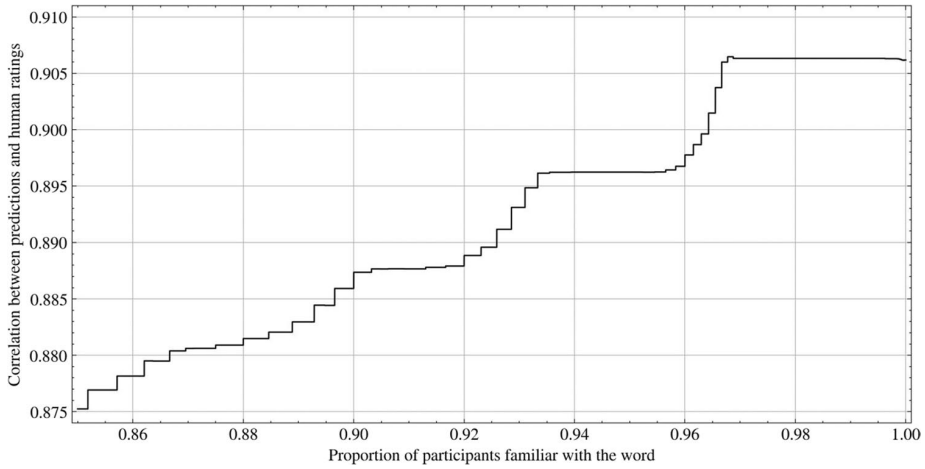


Fig. 2 The effect of word familiarity on prediction performance for concreteness. The *y*-axis indicates the correlation of prediction and ratings estimated from a selection of words that were known by at least the proportion of participants indicated on the *x*-axis

first way one can achieve control on semantic dimensions. The second way rests on the observation that the predictive models we have trained show a mapping from word embeddings to the emotional norms that is locally linear for all norms. This means that, locally, there is a single direction associated with a change in valence and this fact may be used to find a word whose embedding differs principally on this dimension, while being close to all others. By doing this, we increase the chance the matched word will be close in value on all unaccounted-for dimensions that these directions describe. This, in turn, decreases the chance that one of these outside dimensions may, for example, differentially affect reaction times in two conditions of an experiment, challenging its internal validity.

The stimuli descent algorithm takes advantage of the differentiability of our method to find semantically matched words by performing gradient descent in word embedding space with respect to the predicted emotional norm. As the position in word embedding space encodes information about how the word is used and its semantic connections, words that are close together in this space tend to have similar meanings (Stratos et al., 2015). Thus, stimuli descent moves down the function from word embeddings to norms to find the closest word that differs in this norm, allowing for the selection of semantically matched words. To do this, the algorithm at each step predicts the norms and calculates the gradient of the norm we wish to manipulate. As this gradient is the vector in word embedding space that points in the direction of the fastest increase in the predicted rating, the algorithm moves along this vector to find the closest word with the most divergent rating. https://colab.research.google.com/drive/1Cjcejg1AdDhsZWs4VioQ5tT8B496jgFZ

Method

We wish to find semantically matched words less or more pronounced on the manipulated dimension, for words in a set. To simplify the presentation of the algorithm, we will assume our manipulated dimension is valence, and we wish to match words less positive than those from a set of positive words. Apart from the word we must specify the minimum difference in ratings we wish to achieve (denoted $\Delta_{min}r$), so that the difference (denoted Δr) between a matched word's valence and the valence of the original word is at least as

Table 5 The results of the experimental validation test

Condition	Origin	Imageability	Dominance	Arousal	Valence
Original model accuracy	0.86***	0.88***	0.92***	0.86***	0.93***
Experimental accuracy	0.84*** [0.73, 0.92]	0.71*** [0.52, 0.82]	0.86*** [0.73, 0.93]	0.83*** [0.68, 0.92]	0.70*** [0.45, 0.89]

* $p < 0.05$** $p < 0.01$*** $p < 0.001$

Original model accuracy relates to the results presented in Study 1. 95% Confidence intervals are presented in square brackets. The 95% confidence interval for the original model is tighter than the precision with which correlations are reported. The "condition" column relates to the experimental condition, while the rest of the columns relate to the correlations between the respective emotional norms and their predictions

large as $\Delta_{min}r$. Now, we must define the loss function with respect to which we will perform gradient descent. In this instance, it will be the predicted valence rating (with either a plus or minus to increase or decrease the rating). We may also wish to control other emotional dimensions such as arousal. To this end, we may perform a kind of controlled gradient descent, where we want to remain at the same level of the controlled dimensions at each step. Thus, at each step we calculate the prediction of the controlled dimension. The gradient of this prediction shows the direction in which the prediction of controlled dimensions changes and we may use these "controlled gradients" to remove their components from the loss gradient, remaining approximately at the same level of controlled dimensions.

Last modifications to the algorithm involve accounting for the high dimensionality of the space we are moving in. First, as words are discrete points in this high-dimensional embedding space, there may not be a word that corresponds to the place this procedure has moved us to. Thus, to select new stimuli we need to check for approximate matches – words that are close in embedding space according to some distance metric. We chose cosine similarity, a metric typically used for comparing embedding similarity. Second, this high dimensionality makes it easier to move to regions of word embedding space where there are no words. This creates the issue that in such regions the network predictions are only extrapolations, which could easily be wrong. Thus, we add to the loss function a regularization term penalizing the algorithm for stepping outside of a multivariate gaussian distribution approximating the word occurrence distribution. This term is proportional to the logarithm of the Gaussian density function and is described in more detail in Appendix 1 along with other technical details on the objective function. The complete algorithm is described in Fig. 3.

Figure 3 The stimuli descent algorithm, which finds semantically matched words via controlled gradient descent in the word embedding space

Results and discussion

For matching words with the stimuli descent algorithm, we have trained another transformer model for English, but with just one word embedding space, instead of two (with embeddings from the BERT model "bertweet-base-emotion-analysis"; Pérez, 2021). The correlations with human judgments for this model were similar to the ones from Table 2 and evaluated to be $r = 0.95$, $r = 0.76$, $r = 0.86$, $r = 0.85$, and $r = 0.95$, respectively for valence, arousal, dominance, age of acquisition, and concreteness.

We have looked for semantically matched words for two commonly manipulated factors: valence and arousal. While looking for similar words that differ in these dimensions, we have also instituted controlled variables, such that the algorithm was either manipulating valence and controlling arousal or manipulating arousal while controlling valence and dominance. To start, the algorithm needs a word to match the next words to. For each dimension we selected approximately 250 words. Since we can choose to either decrease or increase the ratings, half of these words were high, and half were low on the dimension of interest.

To minimize the risk of bias, the selection of words was done based purely on ratings. For the valence manipulation we picked words whose valence ratings were closest to 1 standard deviation above or 1 standard deviation below the mean valence rating. Similarly, for arousal, we picked words close to 1.5 standard deviation above or below. Selected results of these analyses are presented in Table 4. All obtained results are available in the following OSF repository: https://osf.io/cug92/?view_only=6f246610bc0b43cc9e98d7c978f2f6fa Table 6.

Observing results, we see that successful matches sometimes also automatically match words also on more surface level features, such as length, or how the word sounds. One such example from the supplementary electronic material

Algorithm Stimuli Descent

Set current position to word w's position in embedding space
$\Delta r := 0$ ▷ Set rating difference to 0
while $\Delta r < \Delta_{min}r$ **do**
    Calculate gradient w.r.t. loss and controlled variables
    Remove loss gradient components parallel to controlled gradients
    Update position in the direction of loss gradient
    Find nearest word $\hat{w}$
    $\Delta r := r(w) - r(\hat{w})$ ▷ Compare ratings of the nearest and original words
    if $\Delta r > \Delta_{min}r$ **then**
        Add word $\hat{w}$ to stimuli set
    end if
end while

Fig. 3 The description of the stimuli descent algorithm

Behavior Research Methods (2024) 56:4716–4731 4727

is the match "trample" and "scramble". This is useful as such surface-level features have also been shown to influence behavior in experimental tasks (Hauk & Pulvermüller, 2004; Kuchinke, et al., 2007; Méndez-Bértolo et al., 2011). Matching stimuli with neural networks present an exciting direction for methods research that may lead to more robust experimental results with word stimuli.

Certain limitations of the method must be noted. First, certainly not every word has a word of different emotional value matching its semantic content. What this means for the method is that it can certainly fail to find matching words but will nonetheless propose candidate words. Thus, caution needs to be exercised when using the algorithm to generate stimuli to identify these cases. Some methods can be developed to identify mismatched words. First, the distance of the word at the matching stage can serve as an indicator of the appropriateness of the match. This leads to the second limitation of the validity of distance measurement in high-dimensional spaces. In higher dimensional spaces, the likelihood of finding points diminishes exponentially with the number of dimensions. Future work may try to address this limitation by picking several candidate words using the algorithm and performing the selection process with more sophisticated measures of similarity.

General discussion

The present study has shown that a transformers-based architecture can be useful in predicting a range of affective and other word metrics. While the usefulness of transformers for such tasks has been widely recognized in machine learning (Lin et al., 2022), the full extent of the benefits that these methods can bring to the study of human behavior is an open area for research. The psychologically interesting conclusions that can be drawn from the high correlation of transformer-based norm extrapolation with human judgments can provide support to the claim that the distributional properties of words in written language hold information about how word stimuli are judged in terms of their emotionality by participants of psychological studies (Sahlgren, 2008). In the end, the numerical representations of words, which we use in the training process of our models, are related to how words interact with each other in text (through the attention mechanism concept introduced by Vaswani et al., 2017). A statistical regularity therefore exists between emotion ratings and the structure of human language. The two cognitive mechanisms that can be hypothesized to support such statistical regularity are the influence of affective properties on the way language is used, and the converse mechanism of the patterns in which language is used influencing the affective reactions, either directly or through shaping the interpretation of semantic content (for a discussion of links between distributional co-occurrence and emotional dimensions see Snefjella & Kuperman, 2016). This is easy to imagine as the emotional meaning of words is well associated with their semantic meaning (Vankrunkelsven et al., 2015). This result, however, does not extend to all words of a language, as prediction is limited to words similar to the words used in emotional words databases (Snefjella & Blank, 2020).

The possibility of inferring vast amounts of information from distributional properties finds support in the natural language processing literature, tasked with extracting meaningful semantic information from text based on the co-occurrence of words in large corpora. A common example includes the algebraic treatment of vector word-meaning representations in word embeddings, using which it is possible to find the representation of the word 'man' by subtracting the representation of 'royal' from that of the word 'king' (Ethayarajh, 2019). More novel methods, such as the ones used in this paper, may represent an even wider array of semantic information, which is evidenced by the impressive semantic capabilities of the most advanced transformer-based large language models (Sobieszek & Price, 2022), which for GPT-3 included translation, summarization and

Table 6 Words generated using the stimuli descent algorithm

Manipulation of valence			Manipulation of arousal		
Low	Medium	High	Low	Medium	High
Skeleton	Fossil	Wishbone	**Pleasant**	Splendid	Glorious
Crass	Cheeky	Whimsical	**Villa**	Condo	Mansion
Unscheduled	X	Surprising	**Hanger**	Rake	Spikes
Pretentious	Contemporary	Philosophical	**Patron**	Promoter	Activist
Intimidate	Exceed	**Impress**	Willing	Attentive	**Eager**
Drain	X	**Fountain**	Bike	X	**Motorcycle**
Subdue	Neutralize	**Alleviate**	Mule	Possum	**Skunk**
Insurance	Consumption	**Income**	Unintentional	Overwhelming	**Uncontrollable**

Sample results of semantically matched words. *Bold font* indicates words that were put to the algorithm, the words in the next two columns have been matched by the algorithm. 'X' indicates the algorithm did not find any matching words for that level of the manipulated dimension

the execution of linguistic tasks purely from their description (Brown et al., 2020), which the introduction of GPT-4 expanded to a vast set of common sense task that seem to require some basic understanding of the world (for detailed tests, see Bubeck et al., 2023).

These developments point to an increasing role that machine learning may play in the conduct of psychological studies of language and emotions. A basic use case that the high correlations with human judgments could afford is the use of extrapolated norms for choosing experimental stimuli. While empirical verification of extrapolated norms is always advised, it does not render the extrapolation useless. Say one was designing a study with an orthogonal design that studies the influence of three emotional factors, for example, valence, arousal, and dominance, with affective word stimuli. As valence is highly correlated with dominance and has a quadratic relationship to arousal (Warriner et al., 2013), it is very rare for words to have an emotional load of positive valence, low dominance, and low arousal at the same time, but a group of such words would be required to construct such an orthogonal design. This means that not enough words may be present in the available affective norms dataset to construct such a design. The issue also arises in simpler designs when attempting to control correlated factors. The solution that precise norm extrapolation affords is to use its predictions as a heuristic tool for picking stimuli to put to human evaluation to balance the existing affective word databases. Using existing affective norms, one may predict which words are likely to have the unusual emotional load of positive valence, low dominance, and low arousal, and verify this prediction empirically. In this way, semantic norms extrapolation may be used as tools for picking experimental stimuli.

To support this application of the transformer-based norm extrapolation proposed in this paper, we developed an algorithm for the selection of stimuli. The algorithm leverages both the high correlations with human affective judgments and the semantic aspects of words learnt by the network to select words that are semantically similar, yet affectively different. The algorithm allows one to manipulate emotional factors while controlling others, but also employs the unsupervised control of word meaning that has not yet been explored in the literature on affective words. The novelty of the method stems from its leveraging of the differentiability of our extrapolation method. As the method does not use k-nearest neighbors for extrapolation we can find the gradients of all the predicted norms to find the nearest word with different affective ratings, where the distance is an auxiliary measure of semantic similarity.

An additional consideration is whether our model may be of use in the study of computational models of emotion (Marsella et al., 2010). Firstly, the norms modeled by our network are the average of the outcomes of individual emotional processes of the participants of the normative study. These population-level estimates, while useful, do not correspond to the experiences of any particular person and as such the ability to infer from norms to psychological mechanisms is severely limited. Currently the use of the model in such a context is additionally limited by the general limited understanding of how transformer networks make their predictions and as such drawing any specific scientific conclusions for cognitive science from the trained model should be discouraged.

To avoid scientifically dubious conclusions, the misuse of the model may bring, it is necessary to underscore the limitations highlighted in the robustness section, following the critique of Snefjella & Blank (2020). Our transformer-based extrapolation method, while versatile showing promise in dealing with selection bias, does not overcome the limitation posed by out-of-distribution prediction from a sample from which norms are missing not at random. To address this one should avoid using imputed norms in studies where a systematic bias on uncommon words may lead to false inferences, such studies which analyze stimuli corpora. If one wishes to select as stimuli uncommon words it is necessary to experimentally validate their norms, as there both epistemic and aleatoric uncertainties will impact the model's performance. As discussed previously, it is generally ill-advised to use extrapolated norms of age-of-acquisition and any norm of words known by a fraction of people. The conclusion of the experimental validation is that one should conservatively expect at least a 10% drop in accuracy for out-of-distribution words. Consequently, while the model may be a valuable tool, its applications should always take note of the fact it does not overcome the issues of missing data imputation highlighted by Snefjella & Blank (2020). The same issues apply to all the norm extrapolation methods reported previously as well (Snefjella & Blank, 2020).

Taking advantage of the advances in machine learning is important for the broader scientific community, especially considering the asymmetry in the availability of computational resources. Guided by this thought, we have chosen not only to share all our code and models, but to implement the methods from the paper in an online notebook that allows for their use with a simple graphical interface. In this way, the methods can be used without the need for coding or access to high-end GPUs. We hope that this will maximize the benefits that the methods can bring researchers in the psychology of emotion.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-023-02212-3.

References

Almeida, F., & Xexéo, G. (2019). *Word embeddings: A survey*. arXiv:1901.09069. Retrieved 20 January 2022 from https://doi.org/10.48550/ARXIV.1901.09069

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods, 44*(4), 998–1006. https://doi.org/10.3758/s13428-012-0195-z

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology, 33*(3-4), 130–174. https://doi.org/10.1080/02643294.2016.1147426

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics, 6*, 213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Technical report C-1, Vol. 30, No. 1, pp. 25−36). The Center for Research in Psychophysiology.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica, 150*, 80–84. https://doi.org/10.1016/j.actpsy.2014.04.010

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). *Sparks of artificial general intelligence: Early experiments with Gpt-4*. arXiv:2303.12712. Retrieved 20 January 2022 from https://doi.org/10.48550/arXiv.2303.12712

Citron, F. M. (2012). Neural correlates of written emotion word processing: A review of recent electrophysiological and hemodynamic neuroimaging studies. *Brain and Language, 122*(3), 211–226. https://doi.org/10.1016/j.bandl.2011.12.007

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems, 32*.

Crossfield, E., & Damian, M. F. (2021). The role of valence in word processing: Evidence from lexical decision and emotional Stroop tasks. *Acta Psychologica, 218*, 103359. https://doi.org/10.1016/j.actpsy.2021.103359

Dadas, S. (2020). *Sdadas/polish-roberta [Python]*. GitHub. Retrieved 10 January 2022 from https://github.com/sdadas/polishroberta

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods, 45*(2), 480–498. https://doi.org/10.3758/s13428-012-0260-7

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., Nissim, M. (2019). *BERTje: A Dutch BERT Model*. https://doi.org/10.48550/ARXIV.1912.09582

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018) *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805. Retrieved 20 January from https://doi.org/10.48550/ARXIV.1810.04805

Elkins, K., Chun, J. (2020). Can GPT-3 pass a writer's Turing test? *Journal of Cultural Analytics*. https://doi.org/10.22148/001c.17212

Ethayarajh, K. (2019). *Rotate king to get queen: Word relationships as orthogonal transformations in embedding space*. arXiv:1909.00504. https://doi.org/10.48550/arXiv.1909.00504

Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology, 115*(5), 1090–1103. https://doi.org/10.1016/j.clinph.2003.12.020

Hochreiter, S. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 06*(02), 107–116. https://doi.org/10.1142/S0218488598000094

Hugging Face. (n.d.). Models—hugging face. Retrieved March 29, 2022, from https://huggingface.co/models

Imbir, K. K. (2015). Affective norms for 1,586 polish words (ANPW): Duality-of-mind approach. *Behavior Research Methods, 47*(3), 860–870. https://doi.org/10.3758/s13428-014-0509-4

Imbir, K. K. (2016). Affective Norms for 4900 Polish Words Reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and age of acquisition. *Frontiers in Psychology, 7*, 1081. https://doi.org/10.3389/fpsyg.2016.01081

Imbir, K. K., Pastwa, M., Duda-Goławska, J., Sobieszek, A., Jankowska, M., Modzelewska, A., Wielgopolan, A., & Żygierewicz, J. (2021). Electrophysiological correlates of interference control in the modified emotional Stroop task with emotional stimuli differing in valence, arousal, and subjective significance. *Plos One, 16*(10), e0258177. https://doi.org/10.1371/journal.pone.0258177

Imbir, K. K., Duda-Goławska, J., Sobieszek, A., Wielgopolan, A., Pastwa, M., & Żygierewicz, J. (2022). Arousal, subjective significance and the origin of valence aligned words in the processing of an emotional categorisation task. *Plos One, 17*(3), e0265537. https://doi.org/10.1371/journal.pone.0265537

Imbir, K. K., Duda-Goławska, J., Pastwa, M., Jankowska, M., Modzelewska, A., Sobieszek, A., Żygierewicz, J. (2020). Electrophysiological and behavioral correlates of valence, arousal and subjective significance in the lexical decision task. *Frontiers in Human Neuroscience*, *427*. https://doi.org/10.3389/fnhum.2020.567220

Kanske, P., & Kotz, S. A. (2007). Concreteness in emotional words: ERP evidence from a hemifield study. *Brain Research, 1148*, 138–148. https://doi.org/10.1016/j.brainres.2007.02.044

Kapucu, A., Kılıç, A., Özkılıç, Y., & Sarıbaz, B. (2021). Turkish emotional word norms for arousal, valence, and discrete emotion categories. *Psychological Reports, 124*(1), 188–209. https://doi.org/10.1177/0033294118814722

Kazojć J (2011) Słownik frekwencyjny języka polskiego (Polish language dictionary of attendance). Available: http://www.slowniki.org.pl/i27html . Accessed 20 March 2014

Kenter, T., & De Rijke, M. (2015, October). Short text similarity with word embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1411–1420). https://doi.org/10.1145/2806416.2806475

Kim, E., & Klinger, R. (2019). A survey on sentiment and emotion analysis for computational literary studies. *Zeitschrift Für Digitale Geisteswissenschaften.* https://doi.org/10.17175/2019_008

Kuchinke, L., Võ, M. L. H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word

frequency but not emotional valence. *International Journal of Psychophysiology, 65*(2), 132–140. https://doi.org/10.1016/j.ijpsycho.2007.04.004

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). *A survey of transformers. AI Open, 3*, 111–132. https://doi.org/10.1016/j.aiopen.2022.10.001

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). *Roberta: A robustly optimized Bert pretraining approach*. arXiv:1907.11692. Retrieved 20 January from https://arxiv.org/abs/1907.11692v1

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *Quarterly Journal of Experimental Psychology, 68*(8), 1623–1642. https://doi.org/10.1080/17470218.2014.988735

Marsella, S., Gratch, J., & Petta, P. (2010). Computational models of emotion. *A Blueprint for Affective Computing-A Sourcebook and Manual, 11*(1), 21–46.

Méndez-Bértolo, C., Pozo, M. A., & Hinojosa, J. A. (2011). Word frequency modulates the processing of emotional words: Convergent behavioral and electrophysiological data. *Neuroscience Letters, 494*(3), 250–254. https://doi.org/10.1016/j.neulet.2011.03.026

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems, 35*, 17359–17372.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*. https://doi.org/10.48550/arXiv.1310.4546.

Mohammad, S. (2018, July). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 174–184). https://doi.org/10.18653/v1/P18-1017

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods, 46*(3), 887–903. https://doi.org/10.3758/s13428-013-0405-3

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A.-L., De Schryver, M., De Winne, J., & Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods, 45*(1), 169–177. https://doi.org/10.3758/s13428-012-0243-8

Munikar, M., Shakya, S., & Shrestha, A. (2019, November). Fine-grained sentiment classification using BERT. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)* (Vol. 1, pp. 1–5). IEEE. https://doi.org/10.1109/AITB48515.2019.8947435

Nielsen, F. Å. (2011). *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs*. arXiv:1103.2903. Retrieved 20 January from https://doi.org/10.48550/arXiv.1103.2903

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Pérez, J. M. (2021). *Bertweet base sentiment analysis [Model]*. Hugging Face. Retrieved 10 January 2022 from https://huggingface.co/finiteautomata/bertweet-base-sentiment-analysis

Perezrojas, J., Cañete, C., Chaperon, G., & Zúñiga, R. F. (2020). *BETO: Spanish BERT*. DCC U Chile. Retrieved 10 January 2022 from https://github.com/dccuchile/beto (Original work published 2019).

Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001. https://doi.org/10.18653/v1/P19-1493.

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology, 68*(8), 1584–1598. https://doi.org/10.1080/17470218.2014.941296

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods, 39*(3), 600–605. https://doi.org/10.3758/BF03193031

Sahlgren, M. (2008). The distributional hypothesis. *The Italian Journal of Linguistics*. Retrieved 20 January 2022 from https://www.italian-journal-linguistics.com/app/uploads/2021/05/Sahlgren-1.pdf

Sianipar, A., Van Groenestijn, P., & Dijkstra, T. (2016). Affective meaning, concreteness, and subjective frequency norms for Indonesian words. *Frontiers in Psychology, 7*, 1907. https://doi.org/10.3389/fpsyg.2016.01907

Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. (2019, November). A comparison of semantic similarity methods for maximum human interpretability. In *2019 artificial intelligence for transforming business and society (AITB)* (Vol. 1, pp. 1–4). IEEE. https://doi.org/10.1109/AITB48515.2019.8947433

Sloan, D. M., Strauss, M. E., & Wisner, K. L. (2001). Diminished response to pleasant stimuli by depressed women. *Journal of Abnormal Psychology, 110*(3), 488. https://doi.org/10.1037//0021-843x.110.3.488

Snefjella, B., Blank, I. (2020). *Semantic norm extrapolation is a missing data problem*. Retrieved 20 January from https://doi.org/10.31234/osf.io/y2gav.

Snefjella, B., & Kuperman, V. (2016). It's all in the delivery: Effects of context valence, arousal, and concreteness on visual word processing. *Cognition, 156*, 135–146. https://doi.org/10.1016/j.cognition.2016.07.010

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., & Frade, C. S. (2012). The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods, 44*(1), 256–269. https://doi.org/10.3758/s13428-011-0131-7

Sobieszek, A., & Price, T. (2022). Playing games with Ais: The limits of GPT-3 and similar large language models. *Minds and Machines, 32*(2), 341–364. https://doi.org/10.1007/s11023-022-09602-0

Söderholm, C., Häyry, E., Laine, M., & Karrasch, M. (2013). Valence and arousal ratings for 420 Finnish nouns by age and gender. *PloS One, 8*(8), e72859. https://doi.org/10.1371/journal.pone.0072859

Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods, 49*(1), 111–123. https://doi.org/10.3758/s13428-015-0700-2

Stakovskii, E. (n.d.). *French toxicity classifier plus v2 [Model]*. Hugging Face. Retrieved 20 January 2022, from https://huggingface.co/EIStakovskii/french_toxicity_classifier_plus_v2

Stevenson, R. A., Mikels, J. A., & James, T. W. (2007). Characterization of the affective norms for English words by discrete emotional categories. *Behavior Research Methods, 39*(4), 1020–1024. https://doi.org/10.3758/BF03192999

Stratos, K., Collins, M., & Hsu, D. (2015, July). Model-based word embeddings from decompositions of count matrices. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Volume 1: Long papers)* (pp. 1282–1291). https://doi.org/10.3115/v1/P15-1124

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2019). *ERNIE 2.0: A continual pre-training framework for language understanding*. arXiv:1907.12412. Retrieved 20 January 2022 from arXiv. https://doi.org/10.48550/arXiv.1907.12412

Syssau, A., Yakhloufi, A., Giudicelli, E., Monnier, C., & Anders, R. (2021). FANCat: French affective norms for ten emotional categories. *Behavior Research Methods, 53*(1), 447–465. https://doi.org/10.3758/s13428-020-01450-z

Vaiouli, P., Panteli, M., & Panayiotou, G. (2023). Affective and psycholinguistic norms of Greek words: Manipulating their affective or psycho-linguistic dimensions. *Current Psychology, 42*(12), 10299–10309. https://doi.org/10.1007/s12144-021-02329-8

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Vankrunkelsven, H., Verheyen, S., Storms, G., & De Deyne, S. (2018). Predicting lexical norms: A comparison between a word association model and text-based word co-occurrence models. *Journal of Cognition, 1*(1), 45. https://doi.org/10.5334/joc.50

Vankrunkelsven, H., Verheyen, S., De Deyne, S., Storms, G. (2015). Predicting lexical norms using a word association corpus.In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pp. 2463–2468. https://doi.org/10.5334/joc.50

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017). *Attention Is All You Need*. https://doi.org/10.48550/ARXIV.1706.03762.

Verheyen, S., De Deyne, S., Linsen, S., & Storms, G. (2020). Lexicosemantic, affective, and distributional norms for 1,000 Dutch adjectives. *Behavior Research Methods, 52*(3), 1108–1121. https://doi.org/10.3758/s13428-019-01303-4

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods, 41*(2), 534–538. https://doi.org/10.3758/BRM.41.2.534

Von Platen, P. (2021, May 19). *Bert base German uncased [Model]*. Hugging Face. Retrieved 20 January 2022, from https://huggingface.co/dbmdz/bert-base-german-uncased/tree/main

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Williamson, S., Harpur, T. J., & Hare, R. D. (1991). Abnormal processing of affective words by psychopaths. *Psychophysiology, 28*(3), 260–273. https://doi.org/10.1111/j.1469-8986.1991.tb02192.x

Yao, Z., Yu, D., Wang, L., Zhu, X., Guo, J., & Wang, Z. (2016). Effects of valence and arousal on emotional word processing are modulated by concreteness: Behavioral and ERP evidence from a lexical decision task. *International Journal of Psychophysiology, 110*, 231–242. https://doi.org/10.1016/j.ijpsycho.2016.07.499

Yao, Z., Wu, J., Zhang, Y., & Wang, Z. (2017). Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods, 49*(4), 1374–1385. https://doi.org/10.3758/s13428-016-0793-2

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). *Comparative study of CNN and RNN for natural language processing*. arXiv:1702.01923. Retrieved 20 January 2022, from https://doi.org/10.48550/arXiv.1702.01923

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation, 31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

Open practices statement All the data, code, and tools created as a part of the above research is made freely available under the following links:

GitHub repository:https://github.com/hplisiecki/affect_prediction

Google Colab project:https://colab.research.google.com/drive/1CjcejgIAdDhsZWs4VioQ5tT8B496jgFZ

Stimuli Descent data: https://osf.io/cug92/?view_only=6f246610bc0b43cc9e98d7c978f2f6fa

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

High risk of political bias in black box emotion inference models

Hubert Plisiecki[1,2,3]✉, Paweł Lenartowicz[2], Maria Flakus[3] & Artur Pokropek[3]

This paper investigates the presence of political bias in emotion inference models used for sentiment analysis (SA). Machine learning models often reflect biases in their training data, impacting the validity of their outcomes. While previous research has highlighted gender and race biases, our study focuses on political bias—an underexplored, pervasive issue that can skew the interpretation of text data across many studies. We audit a Polish sentiment analysis model developed in our lab for bias. By analyzing valence predictions for names and sentences involving Polish politicians, we uncovered systematic differences influenced by political affiliations. Our findings suggest that annotations by human raters propagate political biases into the model's predictions. To prove it, we pruned the training dataset of texts mentioning these politicians and observed a reduction in bias, though not its complete elimination. Given the significant implications of political bias in SA, our study emphasizes caution in employing these models for social science research. We recommend a critical examination of SA results and propose using lexicon-based systems as an ideologically neutral alternative. This paper underscores the necessity for ongoing scrutiny and methodological adjustments to ensure the reliability of the use of machine learning in academic and applied contexts.

Keywords Political Bias, Emotion inference models, Sentiment analysis, Social Science Research, Annotator Bias

"*The bias I'm most nervous about is the bias of the human feedback raters.*"

~ Sam Altman, OpenAI CEO.

It is a well-documented fact that machine learning models are prone to being biased by their training data. Studies have repeatedly shown the presence of biases against various social groups, including gender and race biases, in machine learning-based sentiment analysis (SA) systems—systems that predict the positivity of text snippets[1–3]. These types of biases are significant from a social justice perspective, as they can exacerbate the reporting of spurious differences between social groups and affect the interpretation and outcomes of studies across various domains.

In this paper, we highlight another critical dimension of bias in SA systems: political bias, or the propagation of the political orientation of the annotators through the annotated data to the predictions of the SA model. Political bias has the potential to skew the interpretation of data across a wide range of studies, affecting societal perceptions and policymaking at a systemic level. Given that nearly every text contains some level of political nuance[4], this kind of bias can potentially influence many studies that employ SA, especially in the social sciences.

The aim of this research is to show that political bias in SA systems is substantial and pervasive. This bias not only intersects with other biases reported so far, such as gender and race biases, but also extends beyond them, rendering many research conclusions less reliable. By addressing political bias, we seek to contribute to a more comprehensive understanding of biases in SA systems and to encourage the development of mitigation strategies that enhance the reliability and fairness of SA applications.

Emotion and sentiment analysis in Social sciences

In recent years, social scientists have increasingly recognized the profound influence of emotions across a broad range of disciplines. This interdisciplinary approach has illuminated the significant role emotions play in shaping human behavior and societal dynamics in fields such as political science[5], sociology[6,7], economics[8], anthropology[9], and organizational research[10], among others. The proliferation of text data sources—including social media, computer-based survey responses, political speeches, newspapers, online forums, customer reviews, blogs, and e-books—has provided unprecedented opportunities to examine emotions outside traditional psychological laboratory settings. Consequently, various tools have been developed to detect emotions[11,12]. As a result, research in this area has expanded rapidly.

[1]Institute of Psychology, Polish Academy of Sciences, Warsaw, Poland. [2]Stowarzyszenie na rzecz Otwartej Nauki (Society for Open Science), Warsaw, Poland. [3]Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw, Poland. ✉email: hplisiecki@gmail.com

To provide specific examples, in previous research SA been employed to predict election results[13], gauge public sentiment toward pressing social issues[14]and compare the emotional content between news sources from different ends of the political spectrum[15]. During the COVID-19 pandemic, numerous studies analyzed public sentiment based on online data[16,17]leading to conclusions such as describing the crisis communication styles on Twitter of different Indian political leaders[18]. Similar research examined the emotional tone of the Austrian 2016 presidential election candidates[19]. From a psychological perspective, SA and emotion prediction have been used to assess suicide risk[20], automate feedback in online cognitive behavioral therapy[21], predict the subjective well-being of social media users[22], and analyze the subjective well-being of people over the past centuries[23]. All of these studies relied on sentiment analysis written text to reach scientific conclusions, showcasing the importance of this technique in current social research. However, the exact implementation of SA can vary from study to study.

Overall, there exist three main categories of sentiment analysis (SA) systems: (A) dictionary-based approaches, (B) large language model (LLM) approaches, and (C) classical supervised predictive model approaches, referred to from now on as predictive model approaches for brevity. Dictionary-based approaches (A), also known as lexicon-based methods, rely on predefined lists of words associated with specific sentiments. These dictionaries, such as the AFINN, SentiWordNet, and LIWC[24–26], assign sentiment scores to words and phrases within a text to determine its overall sentiment. This method is straightforward and interpretable, but it can be limited by the coverage and accuracy of the dictionary, as well as by the inability to capture contextual information.

In contrast, large language model (LLM) approaches (B) leverage advanced neural networks trained on vast amounts of text data. Models such as GPT-4, LLAMA, and their derivatives can capture nuanced sentiment by understanding the context and relationships between words in a sentence and predict it in a zero-shot (without any examples to guide it), or multiple shot manner (with examples). However, their performance in emotion detection specifically falls short of the state-of-the-art (SOTA) predictive model approaches (C)[27].

The predictive model approach (C) involves training machine learning-based classifiers or regressions on labeled datasets. Techniques such as support vector machines, random forests, and deep learning models are used to predict sentiment based on features extracted from the text. These approaches are currently considered the best for analyzing emotion according to robust tests of prediction accuracy on political text datasets, as well as broader domain benchmarks[27,28]. However, this high accuracy comes at the cost of lower interpretability and, as this study will underline, a propensity for bias.

Bias in predictive models

Bias in predictive models originates from the training data, which in the case of sentiment analysis (SA), consists of annotated text datasets. These datasets are the result of the laborious work of annotators who read through provided materials and assign emotional labels. Annotators can differ on many accounts, including age, gender, socio-economic status, psychological individual differences, and political orientation. All these differences can impact the annotation process. Studies such as Milkowski and associates[29]have shown that individual differences among annotators can significantly affect emotion annotations in text. These individual differences introduce subjectivity into data assumed to be objective, leading to inconsistencies that can skew the training and evaluation of models designed to predict emotional reactions from text. Moreover, annotation bias can result from a mismatch between authors' and annotators' linguistic and social norms, as noted by Sap and colleagues[30]. This mismatch often reflects broader social and demographic differences that can manifest in critical research areas like hate speech and abuse detection. For instance, studies by Larimore and associates[31], and Waseem[32] show that the race and gender of annotators influence not only the annotation process but also the performance of NLP models, further compounding biases.

Particularly concerning is the influence of annotators' political and ideological biases. This type of bias not only includes biases against specific social groups reported in earlier studies, but its generality makes the specific extent of its influence on SA models difficult to determine, although we expect it to be significant[1–3]. Ennser-Jedenastik and Meyer[33]report that coders of political texts often incorporate their prior beliefs about political parties into their coding decisions. For example, annotators are more likely to perceive a sentence as supporting immigration if they believe it comes from a left-wing party, regardless of the actual content. Experimental studies by van der Velden[34]show that personal characteristics of annotators, like political ideology or knowledge, interfere with their judgment of political stances. It's important to note that this interference might not be fully realized by the annotator, as previous psychological studies have shown the influence of political orientation on implicit judgments[35,36]. Here of significant importance are the findings that show that people of different political orientations differ significantly in many annotation tasks related to political science, including emotion annotation of images[37]. This means that constructing an annotation strategy that eliminates the propagation of individual bias to SA models might be problematic. This problem parallels many similar ones in algorithm creation, where the human behavior information, on which the model is trained, falls short of the aim of the engineered algorithm. In such cases, Morewedge and associates[38] recommend auditing the models under suspicion by testing them for the presence of bias directly.

Current study

In this study, we conduct a bias audit of an existing Polish sentiment analysis model developed by our lab as a part of a different research endeavor[39] to determine whether its predicted valence ratings show systematic differences based on the party affiliation of a diverse group of politicians from different political parties. We predict the valence of the names of the politicians, as well as sentences in which their names are embedded to vary based on their political affiliation (the latter were included to analyze both the direct valence towards the politicians as well as take into account the usual settings in which such a model would be used, where the name

of the politicians would be a part of a specific sentence.) We regress the political affiliation of the politicians onto the sentiment readings of the model to see how much variance it can explain. To pinpoint the source of the bias, we prune the training set of any mentions of the aforementioned politicians, train a second model, and repeat the analysis.

Results

Regression models

The predictive model returns the valence metric as a continuous score scaled to a range from 0 to 100. When applied to the 24 names of politicians selected for analysis, the valence scores ranged from 42.3 to 56.6, with an average (not weighted) (M) of 49.5 and a standard deviation (not weighted) (SD) of 3.17. To examine potential bias in more natural contexts, we estimated valence for names embedded in both neutral and politically charged sentences. The mean valence was higher in neutral sentences (M = 54.4) compared to raw names (M = 49.5) and lower in politically charged sentences (M = 45.7). Interestingly, the differences in valence among politicians (measured by the standard deviation of valence) were larger for neutral sentences (SD = 4.35) compared to raw names (SD = 3.17), and smaller for politically charged sentences (SD = 1.29).

Along with the visualization of the differences in predicted valence scores of politicians' names (See Fig. 1), we regressed these valence scores, as well as the predicted valence scores for the aforementioned sentences, onto the independent variables of interest (Table 1). The fitted models with politicians' affiliation and gender (the reasoning for the inclusion of the gender confounder is driven by analyses explained in the later section Confounds) seem to describe the data well, and explain 66.5%, 52% and 66.2% of variance (R^2). All of the coefficients have the same direction and similar magnitudes in all of the three models (Model 1, 2, and 3). The hypothesis of exchangeability of scores[55] could be rejected due to low p-values: $p = 0.008$, $p = 0.049$ and $p = 0.018$, which implies that the differences in valence are not random.

Confounds

The association of political affiliation and valence was significantly stronger than between valence and the confounders. This is evidenced by comparing R^2 of regression on valence in raw names (Model 1, Table 2.) and political affiliation ($R^2 = 0.49$), and models with only confounds as independent variables. The model including only gender reared $R^2 = 0.109$ (statistics that relate to gender should be interpreted carefully since there are only 2 women in our sample), trust towards Zpolitician achieved $R^2 = 0.195$, and the mean valence of mentions in which a given politician appeared resulted in $R^2 = 0.175$. (These models are available in the Appendix)

To find the model that best describes the data, we compare adjusted R^2 with different sets of potential confounds. Since the model with affiliation and gender as independent variables has the highest adjusted R^2, this set of independent variables is used in other models. (See Table 2.)

The modified model

In the model modified (See Table 3.) by pruning texts containing mentions of our set of politicians, the relationships between affiliation and valence decreased significantly, but bias was still present in the model with raw names (Model 1., Table 3.) It should be noted that not all mentions affecting the model could be pruned, for example, the most mentioned politician is Jarosław Kaczyński, but in the dataset there are tweets mentioning his twin brother, Lech Kaczyński, the former president and member of the same party.

Discussion

In the current study we have shown that a supervised model trained on annotations created by expert annotators in their domain shows signs of political bias with regards to well-known politicians. The impact of this bias

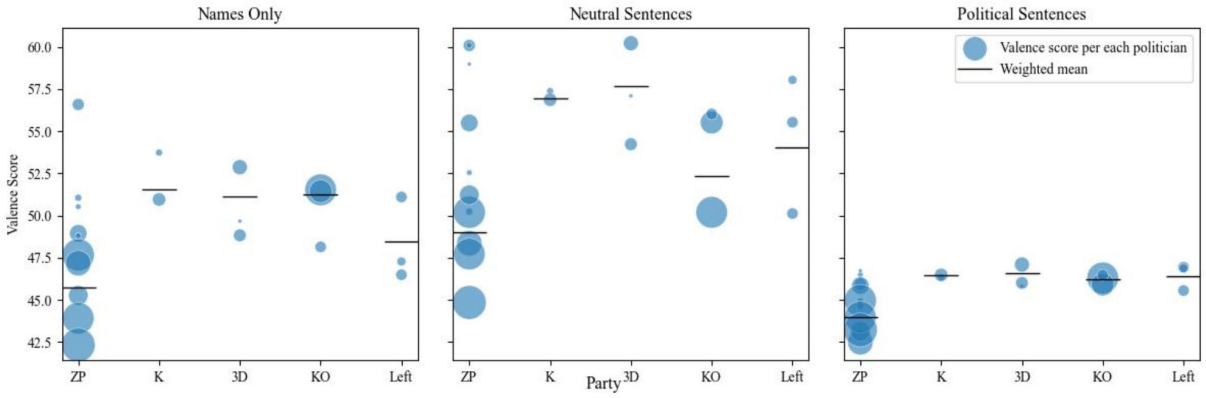


Fig. 1. Predicted valence scores of the names of the politicians. Note: the area of the dots corresponds to the weight (the amount of tweets containing the name of a politician). Abbreviations: ZP – Zjednoczona Prawica, K – Konfederacja, 3D – Trzecia Droga, KO – Koalicja Obywatelska, Left – Nowa Lewica.

	Dependent variable: Valence of:		
	Names only	Neutral Sentences	Political Sentences
	(1)	(2)	(3)
intercept	45.40***	48.61***	43.89***
	(0.67)	(0.92)	(0.26)
3D	5.73**	9.09**	2.72**
	(2.37)	(3.26)	(0.93)
K	6.15*	8.37*	2.56*
	(3.1)	(4.28)	(1.22)
KO	5.83***	3.71*	2.30**
	(1.31)	(1.8)	(0.51)
Left	3.03	5.46	2.51**
	(2.56)	(3.53)	(1.01)
gender	9.77**	10.10*	1.88
	(3.48)	(4.8)	(1.37)
Observations	22	22	22
R^2	0.665	0.52	0.662
Adjusted R^2	0.547	0.37	0.556
Residual Std. Error	3.38 (df = 16)	5.21 (df = 16)	1.29 (df = 16)
P-value (permutation)	0.008	0.049	0.018

Table 1. Regression models – differences in valence. Note: $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$ (*t*-test). Zjednoczona Prawica (ruling party) as intercept, gender: woman = 1, man = 0. K – Konfederacja, 3D – Trzecia Droga, KO – Koalicja Obywatelska, Left – Nowa Lewica.

depends on the analytical context. While using our original model, introducing a politically charged surname can alter the sentiment score of single text snippet by up to 6 points on a 1–100 scale or by 0.5 in terms of Cohen's d. Given the effect size, datasets with minimal politically charged content may experience only minor bias issues. However, when comparing political groups—as in regression analyses across different political parties—this bias becomes systematic. For instance, the difference between groups, when compared to the pooled standard deviation, is 5.47 versus 3.53 (Cohen's d = 1.55) for raw names, or 2.30 versus 1.23 (Cohen's d = 1.87) for modified names. This suggests that bias could be even more pervasive in datasets with greater political content and in inter-group analyses.

Interestingly, while the variance in sentiment predictions of raw names, and neutral sentences was higher than in the case of political sentences, which most probably relates to the higher level of emotional signal present in the latter as the words in political sentences contained more emotionally charged language such as "support", or "against", the systematicity of the bias present in political sentences counteracted this effect as the regression models conducted on each of these conditions explained similar levels of variance (0.665, 0.52, and 0.662 consecutively). This may suggest that the presence of additional political context in the predicted text could prime the model to focus on the political information, thereby making the bias more systematic.

Of note is the fact that this bias was not explained by the publics' trust towards politicians, indicating that it does not reflect the general society's political preferences, but rather those of a selected nonrepresentative group. Similarly, the mean valence of the training data tweets that included the names of the politicians did not explain this bias either. This rules out the hypothesis that the bias comes from a systematic difference in valence between how the politicians were portrayed in text, or other text-inherent reasons. Another argument against the influence of the linguistic bias is the moderate intraclass correlation coefficient (0.6) indicating limited agreement between the annotators which further undermines the possibility of the bias being inherent to the text of the training dataset, and not to the subjective perception of the text by the annotators.

The modified model, trained on a dataset pruned of texts containing politicians' names, exhibited significantly lower bias than the primary model suggesting that at least a substantial part of the bias can be attributed to the annotations made by the annotation team in a causal manner. It, however, does not indicate that pruning the names of the politicians eradicates all kinds of biases that political orientation might result in. Additionally, the model cannot be fully isolated from the influence of certain mentions that may affect its output. For instance, while identifying mentions of Jarosław Kaczyński, snippets related to his twin brother, Lech Kaczyński, might be included, potentially influencing the model's predictions. More indirect sources of bias might also be present. Moreover, certain word associations may be embedded in the model's initial architecture before training for emotion detection, and this pre-existing knowledge could interact with the annotations producing harder to eradicate bias. Given these limitations, along with broader challenges in practical application, we do not recommend pruning as a method for bias mitigation. Furthermore, the instructions given to the annotators, which prompted them to estimate the "positivity/negativity that they read in each text" rather than their emotional reactions to it, leads us to the conclusion that the bias propagated into the annotated dataset in an implicit manner. Instances of such implicit propagation of political orientation have been documented in previous psychological research[35,36].

	Dependent variable: Valence in raw names				
	(1)	(2)	(3)	(4)	(5)
intercept	45.76***	45.40***	45.52***	45.76***	46.06***
	(0.78)	(0.67)	(0.69)	(1.38)	(1.43)
3D	5.36*	5.73**	5.55**	5.14	4.67
	(2.8)	(2.37)	(2.39)	(3.14)	(3.19)
K	5.79	6.15*	5.86*	4.91	4.01
	(3.67)	(3.1)	(3.14)	(5.23)	(5.35)
KO	5.47***	5.83***	5.40***	5.69***	5.16***
	(1.54)	(1.31)	(1.41)	(1.43)	(1.55)
Left	2.67	3.03	3.01	2.71	2.54
	(3.03)	(2.56)	(2.86)	(2.85)	(2.87)
gender		9.77**	8.72*	9.61**	8.40**
		(3.48)	(3.71)	(3.63)	(3.88)
trust			0.71		0.76
			(0.8)		(0.83)
mentions				0.037	0.055
				(0.124)	(0.127)
Observations	22	22	22	22	22
R^2	0.485	0.655	0.672	0.657	0.676
Adjusted R^2	0.364	0.547	0.541	0.52	0.514
Residual Std. Error	3.88	3.38	3.39	3.35	3.31
	(df = 17)	(df = 16)	(df = 15)	(df = 15)	(df = 14)
P-value (permutation)	0.051	0.008	0.021	0.017	0.039

Table 2. Regression models – inspecting confounders. *p < 0.1; **p < 0.05; ***p < 0.01 (t-test). Zjednoczona Prawica (ruling party) as intercept, gender: woman = 1, man = 0, trust: normalized trust scores, mentions: mean valence of annotated text, in which politician was mentions in 0–100 scale. K – Konfederacja, 3D – Trzecia Droga, KO – Koalicja Obywatelska, Left – Nowa Lewica.

	Dependent variable: Valence of: (modified model)		
	raw names	neutral sentences	political sentences
	(1)	(2)	(3)
intercept	49.48***	53.85***	45.09***
	(0.51)	(1.14)	(0.34)
3D	3.70*	5.93	0.77
	(1.79)	(4.00)	(1.20)
K	2.69	4.64	0.41
	(2.35)	(5.26)	(1.58)
KO	1.55	0.87	0.75
	(0.99)	(2.21)	(0.66)
Left	0.76	3.38	0.9
	(1.94)	(4.34)	(1.30)
gender	6.47**	7.74	0.77
	(2.63)	(5.90)	(1.77)
Observations	22	22	22
R^2	0.421	0.224	0.104
Adjusted R^2	0.24	−0.019	−0.176
Residual Std. Error	3.02 (df = 16)	4.87 (df = 16)	1.10 (df = 16)
P-value (permutation)	0.076	0.101	0.202

Table 3. Regression models – modified model (text pruning). *p < 0.1; **p < 0.05; ***p < 0.01 (t-test). Zjednoczona Prawica (ruling party) as intercept, gender: woman = 1, man = 0. K – Konfederacja, 3D – Trzecia Droga, KO – Koalicja Obywatelska, Left – Nowa Lewica.

The most likely explanation for this effect is that when annotators saw a text mentioning a politician, they tended to label it in accordance to their own political orientation. During training, these biased labels were treated as ground truth. Consequently, the model learned to attribute any difference in valence between a text containing the politician's name and a similar text without it to the politician's name itself. Repeated exposure to this pattern reinforced a systematic bias. By removing texts containing such biased annotations, we therefore reduced this bias.

Because the model was originally created for a separate study, we lack detailed information about each annotator's political orientation, making it impossible to directly correlate their political views with the observed bias. Nonetheless, we contacted the annotators post hoc and invited them to complete an anonymous, voluntary survey on their political orientation, to which 15 out of the 20 responded. The results, while generally consistent with the observed bias, offer only tentative evidence for its propagation, and are therefore presented in the appendix.

The existence of political bias in the model has been clearly documented, and so has its causal link to the training data. Direct evidence that this bias aligns precisely with the political orientations of individual annotators is limited, as we lack complete information about their political preferences. However, this limitation does not weaken the conclusion that the model's bias was learned from the annotation process. First, the bias does not reflect society-wide patterns of trust toward these politicians. Second, pruning data that mentioned political figures significantly reduced the bias, supporting the idea that the skew originated in annotations. Finally, a post-hoc, voluntary survey of annotators—albeit incomplete—revealed trends consistent with the observed bias.

These findings highlight that annotator-based biases can readily transfer to trained models, even when instructions direct annotators to judge the text rather than their personal feelings about it. Although the post-hoc survey provides only preliminary insight into how annotators' political leanings might have shaped their labels, such information is not strictly necessary to conclude that the model is biased. The most plausible interpretation remains that model bias stems from the subjective political perceptions of a subset of annotators, whose labeling patterns the model then learned. This means not only that the annotations made by humans can lead to biased models, but also raises the very real possibility that their bias might have spread to more concepts in the dataset. If people implicitly propagate their political orientation towards social groups[1–3] as well as specific politicians as proven by the current study, the only thing standing in the way of abstract concepts being affected by the same type of bias is the ability of the model to pick up on it.

As language models become more advanced, their understanding of language becomes gradually less reliant on specific entities which they pick up from the text as in the case of for example Naive Bayes algorithms[40], and more reliant on relations between abstract concepts. This is evidenced by the distributed nature of the information that large language models and other transformer-reliant architectures use, through the mechanisms of attention, to generate their outputs[41], as well as by the recent LLM interpretability research showcasing the crystallization of abstract concepts within the inner layers of these models[42]. This means that as models improve, the propensity of the models being biased towards specific abstract concepts such as for example anarchism, or democracy might increase, given that such bias will be present in the training data, which is likely. Furthermore, the inspection of these kinds of distributed, conceptual biases will require new, more complex methods of bias detection.

Given the biases that have been already uncovered in SA models, as well as those more abstract that can lurk in the shadows, yet unidentified, we discourage the use of such models for research and advise caution in interpreting the results of those that have already used them. To stress the kinds of problems their use can lead to, let's go back to the examples of research performed with the use of SA systems. The analysis of the sentiment towards social issues might be biased towards the sentiment of the annotator's team[14]. Similarly, when comparing emotional content of news sources, the same propagation of bias can occur[15], directly biasing the conclusions. This problem of propagation of bias directly biases studies that apply their SA systems to compare different groups of texts in terms of emotionality. When trying to predict something using SA scores, like in the case of predicting election results, assessing suicide risk, or subjective wellbeing, the effectiveness of the predictive model can be influenced by the beliefs of the annotator group, leading to replication issues[13,20,22]. At the same time, when creating customer facing solutions such as automating feedback in online cognitive behavioral therapy one has to consider that annotator biases might lead to people with different political predispositions receiving different standards of care, however here the influence is not as clear cut as in earlier cases.

However, some of the studies mentioned in this paper may be less affected by this bias, as many of them have relied on lexicon-based SA systems, forgoing the increased accuracy of the predictive models in exchange for elevated transparency. As these approaches depend on lists of emotionally loaded words which are not ideologically relevant, annotated separately, and without any contexts, they are significantly less susceptible to propagating the bias of their annotators. Furthermore, any bias that they do propagate can be clearly read from the word annotations themselves, therefore researchers that want to buttress their analysis against specific biases can directly check for them within the lexicon and correct them there and then. The same task is orders of magnitude more complicated when using black box ML models and becomes even more complicated when the bias concerns concepts rather than entities. Lexicons, however, should not by any means be assumed to be bias-free, but rather less susceptible to carry it, and easier to buttress against it.

The higher accuracy of transformer-based, and other predictive models could be therefore traded in for the less accurate, but more bias- safe lexicon-based systems. However, given that the drop in performance when using lexicon approaches is quite severe this might not be a preferred solution for some researchers[28]. Additionally, lexicons might exhibit different types of biases – such as those related to lists of words that are not representative of their natural language use. They should, therefore, also be used with caution. Future research should therefore focus on creating emotion prediction ML models that are more robust to training bias. In the case where authors choose to use ML based SA systems anyway, we recommend them to take the possibility of

different types of potential biases into consideration when analyzing their results, and if possible, to corroborate their results using a lexicon-based system.

The alternative in the form of picking the annotation team so that it is balanced with regards to all of the individual differences such as political orientation and others that could influence their annotations is problematic as (1) it is as of yet not clear which differences could play a role in the annotation process (2) balancing a large number of them would require a very large annotation team which would be very resource intensive. Nonetheless in very specific applications where the nature of the bias relevant to a given experiment can be directly pinpointed such solutions might be viable.

In conclusion, the current paper shows that supervised models trained on datasets annotated by humans are susceptible to showing the same biases as annotators, despite the annotation instructions being phrased in a way that should avoid the propagation of such bias. This result should be taken into consideration when conducting and interpreting sentiment analysis research in the political science sphere and beyond. We therefore recommend the research community to perceive machine learning based sentiment analysis models as biased until proven otherwise and consider exploring alternative approaches.

The main limitation of the current study is its focus on a single sentiment analysis model and a specific dataset largely composed of political texts in Polish. While these conditions are ideal for exploring political bias within the context of Polish politics, the generalizability of the findings cannot be stated with certainty, although should be taken into consideration. We recommend the researchers that are in doubt about whether our results extend to their models to replicate our findings before using them. Additionally, the sample size of politicians and the specific sentences used to assess bias were relatively small, which may limit the robustness of our regression analyses. Future research should aim to replicate these findings across diverse datasets, expand the number of annotators and the range of their political orientations, and explore the interaction between different types of bias in sentiment analysis models. Explorations of the exact mechanisms through which the bias is propagated would also be insightful from a psychological perspective, and perhaps could bolster the development of bias-safe emotion prediction alternatives.

Methods

The prediction model

Model training data

The model has been trained on a training set sampled from a comprehensive database of Polish political texts from social media profiles (i.e., YouTube, Twitter, Facebook) of 25 journalists, 25 politicians, and 19 non-governmental organizations (NGOs). The complete list of the profiles is available in the Appendix. For each profile, all available posts from each platform were scraped (going back to the beginning of 2019). In addition, we also included texts written by "typical" social media users, i.e., non-professional commentators of social affairs. Our data consists of 1,246,337 text snippets (Twitter: 789490 tweets; YouTube: 42252 comments; Facebook: 414595 posts).

As transformer models have certain limits, i.e., their use imposes limits on length, we implemented two types of modification within the initial dataset. First, since texts retrieved from Facebook were longer than the others, we have split them into sentences. Second, we deleted all texts that were longer than 280 characters.

The texts were further cleaned from social media artifacts, such as dates scrapped alongside the texts. Next, the *langdetect*[43] software was used to filter out text snippets that were not written in Polish. Also, all online links and usernames in the texts were replaced with "_link_" and "_user_", respectively, so that the model does not overfit the sources of information nor specific social media users.

Because most texts in the initial dataset were emotionally neutral, we filtered out the neutral texts and included only those that had higher emotional content in the final dataset. To filter the neutral snippets, the texts were stemmed and subjected to a lexicon analysis[44] using lexical norms for valence, arousal, and dominance - the three basic components of emotions. The words in each text were summed up in terms of their emotional content extracted from the lexical database and averaged to create separate metrics for the three emotional dimensions. These metrics were then summed up and used as weights to choose 10,000 most emotionally loaded texts for the final training dataset. The proportions of the texts coming from different social media platforms reflected the initial proportions of these texts, resulting in 496 YouTube texts, 6105 Twitter texts, and 3399 Facebook texts, totaling 10,000 texts.

Annotators

The final dataset consisting of 10,000 texts was annotated by 20 expert annotators (age: $M = 23.89$, $SD = 4.10$; gender: 80% female) with regards to six emotions: happiness, sadness, disgust, fear, anger, and pride, as well as to two-dimensional emotional metrics of valence and arousal, using a 5-point Likert scale. All annotators were well-versed in Polish political discourse and were students of Psychology (70% of them were graduate students, which in the case of Polish academic education denotes people studying 4th and 5th year). Thus, they underwent at least elementary training in psychology. Each text was annotated by 5 randomly picked annotators. The inter annotator reliability as measured by the intraclass correlation coefficient (ICC(1)) for valence measured 0.60 indicating moderate reliability[45].

Since valence and arousal might not have been familiar to annotators, before the formal annotation process began, all annotators were informed about the characteristics of valence and arousal. General annotation guidelines were provided to ensure consistency and minimize subjectivity. For the purpose of annotating valence of texts, the annotators were given the following instruction:

English translation (An in-depth description of the annotation process is available in the Appendix):

Go back to the text you just read. Now think about the sign of emotion (positive / negative) and the arousal you read in a given text (no arousal / extreme arousal). Rate the text on these emotional dimensions.

Model training

For model training, we have considered two alternative base models: the Trelbert transformer model developed by a team at DeepSense[46], and the Polish Roberta model[47]. The encoders of both models were each equipped with an additional regression layer with a sigmoid activation function. The models have been trained to predict each of the six emotion intensities, as well as valence, and arousal. The maximum number of epochs in each training run was set to 100. At each step, we computed the mean correlation of the predicted metrics with their actual values on the evaluation batch, and the models with the highest correlations on the evaluation batch were saved to avoid overfitting. We used the MSE criterion to compute the loss alongside the AdamW optimizer with default hyperparameter values. Both of the base models were then subjected to a Bayesian grid search using the WandB platform[48] with the following values: dropout – 0; 0.2, 0.4, 0.6; learning rate – 5e-3, 5e-4, 5e-5; weight decay – 0.1, 0.3, 0.5; warmup steps – 300, 600, 900. The model which obtained the highest correlation relied on the Roberta transformer model and had the following hyperparameters: dropout = 0.6; learning rate = 5e-5; weight_decay = 0.3. Its average accuracy on the test set is $r = 0.80$, and $r = 0.87$ valence, which is the main metric analyzed in the current study as it shows the estimated general positivity of the analyzed text.

Bias testing

Stimuli

As stimuli for testing the bias hypothesis, to limit our arbitrary choice of stimuli, we used the names of 24 well-known Polish politicians who appeared in the November and October 2023 trust polls[49–51]. The politicians were assigned to 5 political parties/coalitions on the basis of their affiliation or because they were candidates of that party/coalition. These parties/coalitions are Zjednoczona Prawica, which is right-wing and was the ruling coalition, Trzecia Droga, Koalicja Obywatelska, Nowa Lewica, which were centre-right, centre and left opposition respectively. The fifth party was Konfederacja, which was a right to far right opposition. These coalitions cover 96.25% of the total votes in the November 2023 parliamentary elections[52].

The model was used to predict the valence for each of the aforementioned stimuli. While capable of estimating the intensity of other affective metrics, the choice of valence is both natural and self-evident: valence, by definition, reflects the positive or negative reaction to a stimulus. No other affective metric aligns as directly with the binary essence of approve/disapprove evaluations, making valence the most intuitive and robust indicator of bias in politically charged contexts. To further examine this, we predicted the valence of politicians' names in isolation, as well as in neutral and politically contextualized sentences, to estimate how their inclusion alters the model's predictions. Details of these stimuli are provided in the appendix.

Corpus modification

To identify the potential source of the model's bias, we locate the texts in the training set that contain the surnames of these politicians. We then manually review these texts to see if they are referring to a particular politician. There are 459 of these texts in total, with a range of 71 to 0 and a median of 8.5 per politician. We then prune the training set of these texts and train a second model with the same training parameters to estimate the degree to which their presence influences the model's bias. The training set contained 7999 texts before pruning, which means that the pruned texts constitute below 6% of its size.

Statistical analyses

To test for the presence of bias, we examine where there are noticeable differences in the valence of politicians' names and where they can be explained by the politicians' political affiliations. For this purpose, we build several regression models. As dependent variables, we use the valence score from the original model, the same score from the modified model (trained on the pruned corpus), and the differences in valence between the final and the modified model. The models return the valence score as continuous variables ranging from 0 to 1, which we chose to then recalculate on a 0–100 scale for better readability.

As independent variables we use the politicians' affiliation, and potential confounders: their gender, trust towards them (from the same trust surveys as the names of politicians) and mean annotated valence of texts in which these politicians appear, recalculated to 0–100. We included a trust "score" as a proxy to control for the general favourability of each politician, to separate it from nonrepresentative political bias. This allows us to account for positive or negative feelings about a politician that may stem from their popularity or personal traits rather than political alignment. Additionally, the mean valence of the training dataset snippets with the politicians' names was included to test the possibility that the bias of the model stems from text-inherent sources such as biased language. This could mean for example that certain politicians were accompanied by more negative language than others, translating to biased training. By incorporating both trust and mean annotated valence scores, we aimed to rule out alternative explanations for differences in valence beyond the annotators' political bias.

The trust surveys were decoded as 5-point Likert scales[51]or 3-point Likert scales[49,50]. Responses "I don't know" and "difficult to say" were recoded as neutral. For each survey, a normalized score was calculated, and the mean of these normalized scores was included in the analysis. Mean annotated valence scores were derived from texts that were later pruned, see 'Corpus Modification', and recalculated to 0–100 scale.

For the regression models we use the weighted least squares method[53], weighted by the number of mentions of a given politician. Due to the weighting process, two politicians without any mentions in training data were excluded. To test the null hypothesis of lack of correlation between political affiliation and bias in the models, we conducted the permutation tests on the observed valence (Manly, 1997) for each model, with 100,000 random

assignments. This method guarantees robustness and decent statistical power[54]. This method could be vulnerable to extreme outlier in dependent variables, which is not a problem in this study, due to the categorical or bounded character of dependent variables used in this study. The QQ-plots of the model residuals are included in the appendix. Due to small sample sizes for affiliations, parametric (assuming normal distributions) confidence intervals are calculated.

Data availability

The code and data used in the current study is available at the github repository https://github.com/hplisiecki/political-model-bias and the https://osf.io/q8bes/?view_only=6f246610bc0b43cc9e98d7c978f2f6fa . The base model used for the current study is available at https://huggingface.co/hplisiecki/polemo_intensity , while the modified model is available at the aforementioned OSF repository.

Received: 26 July 2024; Accepted: 14 January 2025
Published online: 19 February 2025

References

1. Diaz, M., Johnson, I., Lazar, A., Piper, A. M. & Gergle, D. Addressing Age-Related Bias in Sentiment Analysis. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–14. (2018). https://doi.org/10.1145/3173574.3173986
2. Kiritchenko, S. & Mohammad, S. M. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems (arXiv:1805.04508). arXiv. (2018). https://doi.org/10.48550/arXiv.1805.04508
3. Ungless, E. L., Ross, B. & Belle, V. Potential pitfalls with automatic sentiment analysis: the Example of Queerphobic Bias. *Social Sci. Comput. Rev.* **41** (6), 2211–2229. https://doi.org/10.1177/08944393231152946 (2023).
4. Fairclough, N. Language and Power (2nd ed.). Routledge. (2013). https://doi.org/10.4324/9781315838250
5. Mintz, A., Valentino, N. A. & Wayne, C. *Beyond Rationality: Behavioral Political Science in the 21st Century* (Cambridge University Press, 2022).
6. Bericat, E. The sociology of emotions: four decades of progress. *Curr. Sociol.* **64** (3), 491–513. https://doi.org/10.1177/0011392115588355 (2016).
7. Turner, J. H. & Stets, J. E. Sociological theories of human emotions. *Ann. Rev. Sociol.* **32** (1), 25–52. https://doi.org/10.1146/annurev.soc.32.061604.123130 (2006).
8. Loewenstein, G. Emotions in Economic Theory and Economic Behavior. *Am. Econ. Rev.* **90** (2), 426–432. https://doi.org/10.1257/aer.90.2.426 (2000).
9. Lutz, C. & White, G. M. The Anthropology of emotions. *Annu. Rev. Anthropol.* **15** (1), 405–436. https://doi.org/10.1146/annurev.an.15.100186.002201 (1986).
10. Diener, E., Thapa, S. & Tay, L. Positive emotions at work. *Annual Rev. Organizational Psychol. Organizational Behav.* **7** (1), 451–477. https://doi.org/10.1146/annurev-orgpsych-012119-044908 (2020).
11. Mohammad, S. M. Sentiment Analysis. In Emotion Measurement (pp. 201–237). Elsevier. (2016). https://doi.org/10.1016/B978-0-08-100508-8.00009-6
12. Üveges, I. & Ring, O. HunEmBERT: a fine-tuned BERT-Model for classifying sentiment and emotion in Political Communication. *IEEE Access.* **11**, 60267–60278. https://doi.org/10.1109/ACCESS.2023.3285536 (2023).
13. Ramteke, J., Shah, S., Godhia, D. & Shaikh, A. Election result prediction using Twitter sentiment analysis. 2016 International Conference on Inventive Computation Technologies (ICICT), 1, 1–5. (2016). https://doi.org/10.1109/INVENTIVE.2016.7823280
14. Kim, S. Y., Ganesan, K., Dickens, P. & Panda, S. Public sentiment toward Solar Energy—Opinion Mining of Twitter using a transformer-based. *Lang. Model. Sustain.* **13** (5). https://doi.org/10.3390/su13052673 (2021).
15. Rozado, D., Hughes, R. & Halberstadt, J. Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models. *PLOS ONE.* **17** (10), e0276367. https://doi.org/10.1371/journal.pone.0276367 (2022).
16. Alamoodi, A. H. et al. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: a systematic review. *Expert Syst. Appl.* **167**, 114155. https://doi.org/10.1016/j.eswa.2020.114155 (2021).
17. Wang, J. et al. Global evidence of expressed sentiment alterations during the COVID-19 pandemic. *Nat. Hum. Behav.* **6** (3), 349–358. https://doi.org/10.1038/s41562-022-01312-y (2022).
18. Kaur, M., Verma, R. & Otoo, F. N. K. Emotions in leader's crisis communication: Twitter sentiment analysis during COVID-19 outbreak. *J. Hum. Behav. Social Environ.* **31** (1–4), 362–372. https://doi.org/10.1080/10911359.2020.1829239 (2021).
19. Kušen, E. & Strembeck, M. Politics, sentiments, and misinformation: an analysis of the Twitter discussion on the 2016 Austrian Presidential elections. *Online Social Networks Media*. **5**, 37–50. https://doi.org/10.1016/j.osnem.2017.12.002 (2018).
20. Glenn, J. J., Nobles, A. L., Barnes, L. E. & Teachman, B. A. Can text messages identify suicide risk in Real Time? A within-subjects pilot examination of temporally sensitive markers of suicide risk. *Clin. Psychol. Sci.* **8** (4), 704–722. https://doi.org/10.1177/2167702620906146 (2020).
21. Provoost, S., Ruwaard, J., Van Breda, W., Riper, H. & Bosse, T. Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: an exploratory study. *Front. Psychol.* **10**, 1065 (2019).
22. Chen, L., Gong, T., Kosinski, M., Stillwell, D. & Davidson, R. L. Building a profile of subjective well-being for social media users. *PLOS ONE.* **12** (11), e0187278. https://doi.org/10.1371/journal.pone.0187278 (2017).
23. Hills, T. T., Proto, E., Sgroi, D. & Seresinhe, C. I. Historical analysis of national subjective wellbeing using millions of digitized books. *Nat. Hum. Behav.* **3** (12), 1271–1275. https://doi.org/10.1038/s41562-019-0750-z (2019).
24. Baccianella, S., Esuli, A. & Sebastiani, F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *Lrec* **10** (2010), 2200–2204 (2010). http://lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf
25. Boyd, R. L., Ashokkumar, A., Seraj, S. & Pennebaker, J. W. *The Development and Psychometric Properties of LIWC-22*1–47 (University of Texas at Austin, 2022).
26. Nielsen, F. A. Afinn project. DTU Compute Technical University of Denmark. (2017). http://www2.imm.dtu.dk/pubdb/edoc/imm6975.pdf
27. Kocoń, J. et al. ChatGPT: Jack of all trades, master of none. *Inform. Fusion*. **99**, 101861. https://doi.org/10.1016/j.inffus.2023.101861 (2023).
28. Widmann, T. & Wich, M. Creating and comparing Dictionary, Word Embedding, and transformer-based models to measure Discrete emotions in German Political text. *Political Anal.* **31** (4), 626–641. https://doi.org/10.1017/pan.2022.15 (2023).
29. Milkowski, P. et al. Personal Bias in Prediction of Emotions Elicited by Textual Opinions. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, 248–259. (2021). https://doi.org/10.18653/v1/2021.acl-srw.26

30. Sap, M., Card, D., Gabriel, S., Choi, Y. & Smith, N. A. The risk of racial Bias in hate Speech Detection. *Proc. 57th Annual Meeting Association Comput. Linguistics*. **1668-1678** https://doi.org/10.18653/v1/P19-1163 (2019).
31. Larimore, S., Kennedy, I., Haskett, B. & Arseniev-Koehler, A. Reconsidering annotator disagreement about Racist Language: noise or signal? *Proc. Ninth Int. Workshop Nat. Lang. Process. Social Media*. **81-90** https://doi.org/10.18653/v1/2021.socialnlp-1.7 (2021).
32. Waseem, Z. Are you a racist or am I seeing things? Annotator influence on hate Speech Detection on Twitter. *Proc. First Workshop NLP Comput. Social Sci.* **138-142** https://doi.org/10.18653/v1/W16-5618 (2016).
33. Ennser-Jedenastik, L. & Meyer, T. M. The impact of Party cues on Manual Coding of Political texts. *Political Sci. Res. Methods*. **6** (3), 625–633. https://doi.org/10.1017/psrm.2017.29 (2018).
34. Van Der Velden, M. A. C. G. et al. Whose truth is it anyway? An experiment on Annotation Bias in Times of factual opinion polarization. (2023). https://doi.org/10.31235/osf.io/nd6yr
35. Carraro, L., Negri, P., Castelli, L. & Pastore, M. Implicit and explicit illusory correlation as a function of political ideology. *PLoS ONE*. **9** (5), e96312. https://doi.org/10.1371/journal.pone.0096312 (2014).
36. Jost, J. T. The IAT is dead, long live the IAT: context-sensitive measures of implicit attitudes are indispensable to Social and political psychology. *Curr. Dir. Psychol. Sci.* **28** (1), 10–19. https://doi.org/10.1177/0963721418797309 (2019).
37. Webb Williams, N., Casas, A., Aslett, K. & WilkersonJ.D. When conservatives see Red but liberals feel Blue: why labeler-characteristic Bias matters for data annotation. *SSRN Electron. J.* https://doi.org/10.2139/ssrn.4540742 (2023).
38. Morewedge, C. K. et al. Human bias in algorithm design. *Nat. Hum. Behav.* **7** (11), 1822–1824. https://doi.org/10.1038/s41562-023-01724-4 (2023).
39. Plisiecki, H., Koc, P., Flakus, M. & Pokropek, A. Predicting Emotion Intensity in Polish Political Texts: Comparing Supervised Models and Large Language Models in a Resource-Poor Language (arXiv:2407.12141). arXiv. (2024). http://arxiv.org/abs/2407.12141
40. Webb, G. I., Keogh, E. & Miikkulainen, R. Naïve Bayes. *Encyclopedia Mach. Learn.* **15** (1), 713–714 (2010).
41. Vaswani, A. et al. Attention is all you need (Version 7). *arXiv* https://doi.org/10.48550/arXiv.1706.03762 (2017).
42. Templeton, A. et al. Scaling monosemanticity: extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread* (2024). https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html
43. Danilak, M. M. langdetect: Language detection library ported from Google's language-detection. (1.0.9) [Python; OS Independent]. (2021). https://github.com/Mimino666/langdetect
44. Imbir, K. K. Affective norms for 4900 Polish words reload (ANPW_R): assessments for Valence, Arousal, Dominance, Origin, significance, concreteness, Imageability and, Age of Acquisition. *Front. Psychol.* **7**, 1081. https://doi.org/10.3389/fpsyg.2016.01081 (2016).
45. Koo, T. K. & Li, M. Y. A Guideline of selecting and reporting Intraclass correlation coefficients for Reliability Research. *J. Chiropr. Med.* **15** (2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012 (2016).
46. Szmyd, W. et al. TrelBERT: A pre-trained encoder for Polish Twitter. In J. Piskorski, M. Marcińczuk, P. Nakov, M. Ogrodniczuk, S. Pollak, P. Přibáň, P. Rybak, J. Steinberger, & R. Yangarber (Eds.), Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023) (pp. 17–24). Association for Computational Linguistics. (2023). https://doi.org/10.18653/v1/2023.bsnlp-1.3
47. Dadas, S. Sdadas/polish-roberta [Python]. https://github.com/sdadas/polish-roberta (2020). (Original work published 2020).
48. Wandb/wandb. [Python]. Weights & Biases. (2024). https://github.com/wandb/wandb (Original work published 2017).
49. CBOS. Zaufanie do polityków w październiku. (2023b). https://www.cbos.pl/SPISKOM.POL/2023/K_134_23.PDF
50. CBOS. Powyborczy ranking zaufania do polityków. (2023c). https://cbos.pl/SPISKOM.POL/2023/K_144_23.PDF
51. Ibris. Sondaż IBRiS dla Onetu: Powyborcze trzęsienie ziemi. Wielki powrót na podium rankingu zaufania. (2023). Retrieved from https://wiadomosci.onet.pl/tylko-w-onecie/sondaz-zaufania-powyborcze-trzesienie-ziemi-wielki-powrot-na-podium/ewfn3vv
52. PKW. Wyniki głosowania w wyborach do Sejmu w 2023 r. Wybory do Sejmu i Senatu Rzeczypospolitej Polskiej 15 października 2023. (2023). https://sejmsenat2023.pkw.gov.pl/sejmsenat2023/pl/sejm/wynik/pl
53. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical modeling with Python. 92–96. (2010). https://doi.org/10.25080/Majora-92bf1922-011
54. Anderson, M. J., Robinson, J. & Australian Permutation Tests for Linear Models. *New. Z. J. Stat.*, **43**(1), 75–88. https://doi.org/10.1111/1467-842X.00156 (2001).
55. Manly, B. F. J. *Randomization, Bootstrap and Monte Carlo Methods in Biology* 0 edn (Chapman and Hall/CRC, 2018). https://doi.org/10.1201/9781315273075

Author contributions

H.P. made substantial contributions to the conception and design of the work, as well as the acquisition, analysis, interpretation of the data, work drafting, and revision.P.L. made substantial contributions to the design of the work, as well as the analysis, interpretation of the data, and revision.M.F. made substantial contributions to drafting and revision of the work.A.P. made substantial contributions to design, drafting and revision of the work, as well as funding acquisition.All of the authors have approved the submitted version and agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-86766-6.

Correspondence and requests for materials should be addressed to H.P.

Eradicating Social Biases in Sentiment Analysis using Semantic Blinding and Semantic Propagation Graph Neural Networks

Hubert Plisiecki

hplisiecki@gmail.com

Institute of Psychology, Polish Academy of Sciences

Society for Open Science

ORCID: 0000-0002-5273-1716

Abstract

This paper introduces the Semantic Propagation Graph Neural Network (SProp GNN), a machine learning sentiment analysis (SA) architecture that relies exclusively on syntactic structures and word-level emotional cues to predict emotions in text. By semantically blinding the model to information about specific words, it is robust to social biases such as political or gender bias that have been plaguing previous machine learning-based SA systems. The SProp GNN shows performance superior to lexicon-based alternatives such as VADER (Valence Aware Dictionary and Sentiment Reasoner) and EmoAtlas on two different prediction tasks, and across two languages. Additionally, it approaches the accuracy of transformer-based models while significantly reducing bias in emotion prediction tasks. By offering improved explainability and reducing bias, the SProp GNN bridges the methodological gap between interpretable lexicon approaches and powerful, yet often opaque, deep learning models, offering a robust tool for fair and effective emotion analysis in understanding human behavior through text.

DOI: https://doi.org/10.48550/arXiv.2411.12493

Eradicating Social Biases in Sentiment Analysis using Semantic Blinding and Semantic Propagation Graph Neural Networks

The automated assessment of emotional content in textual data, or Sentiment Analysis (SA) has revolutionized research across the social sciences, enabling applications such as suicide risk prediction from text messages (Glenn et al., 2020), analysis of historical well-being trends (Hills et al., 2019), political election forecasting (Ramteke et al., 2016), and monitoring global emotional responses during crises like the COVID-19 pandemic (Wang et al., 2022). Thanks to SA researchers gained access to an extensive array of authentic data on human emotions, as vast as the multitude of texts available on the internet.

However, current methods for emotion assessment have notable limitations. Transformer-based architectures, and other machine learning models — while being recommended for their high performance (Widmann & Wich, 2022), are at the same time susceptible to inheriting social biases from their training data, including gender, racial, ageist, and political biases (Kiritchenko & Mohammad, 2018; Díaz et al., 2018; Plisiecki et al., 2024). For instance, Kiritchenko & Mohammad surveyed 219 automatic sentiment analysis systems 75% of which showed signs of significant racial and/or gender bias. A more targeted investigation conducted found that a model trained on annotated political texts exhibited biases aligned with the political orientation of the annotators. Removing bias relevant items from the training data reduced these biases, implying that the annotations were their source of origin (Plisiecki et al., 2024). As these findings highlight, addressing bias in emotion modeling has become an essential challenge for sentiment analysis research.

Given that balancing the annotator group in terms of bias is problematic as, aside from the labor required to find the right people, there always exists a risk of the existence of a bias that was not accounted for. The alternative so far has been the use of simpler models, such as those based on lexicons (also called norms or dictionaries), which are long lists of manually selected words annotated for their emotional information (Plisiecki et al., 2024). The most basic lexicon approaches rely on simply looking up the emotional value of each available word in a text and averaging the results. Unfortunately, this approach works well only for very simple texts, as it does not consider syntactic information. An example here is negation, which can transform the meaning of a word in a sentence but goes unnoticed by simple dictionaries.

More complex alternatives rely on hard coded rules to handle syntactic dependencies. Examples of such approaches are the VADER (Valence Aware Dictionary and Sentiment Reasoner), and EmoAtlas (Hutto & Gilbert, 2014; Semeraro et al., 2023). Both of these techniques rely on dictionaries combined with hard-coded rules that were arrived at through examination of sentence structures. Rules such as “if the negation is three words away from an emotionally loaded term, flip the emotional loading of the term” allow those models to handle negations and other semantic structures beyond the reach of normal lexicons. Their performance however rarely

approaches that of pretrained transformers and the degree of generalization to languages different than English is questionable, due to different syntactic patterns being present in languages further away on the language phylogenetic tree.

The Proposed Solution

To provide a better solution, this paper presents the Semantic Propagation Graph Neural Network (SProp GNN), a supervised approach that bridges the methodological gap between simple lexicon-based methods and complex black-box models providing high performance that is robust to training data bias. This approach uses the syntactic relationships within sentences to create graphs enhanced with emotion information at the word level. The SProp GNN is then trained on these graphs, providing emotion predictions at text level. The risk of bias propagation is reduced by purposefully blinding the model to semantic information that it could otherwise overfit.

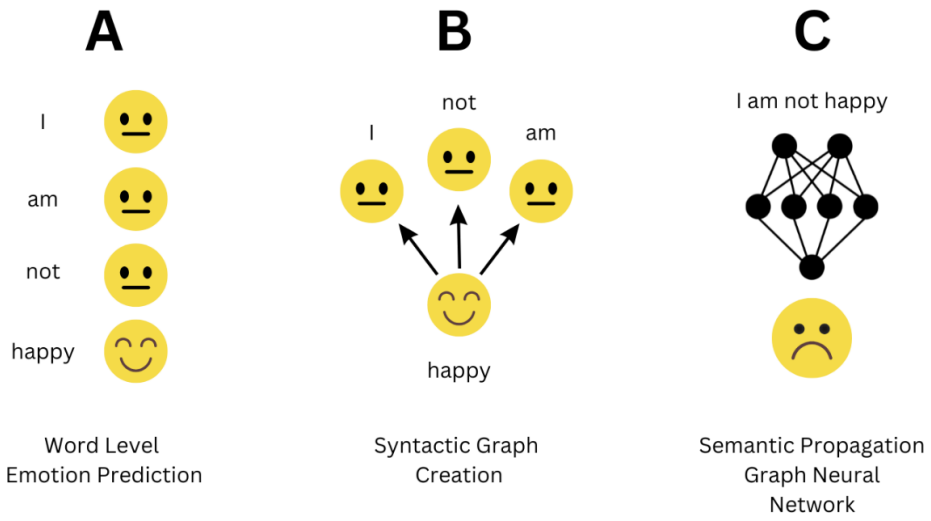
The SProp GNN emotion prediction pipeline can be split into three distinct stages (as seen on Figure 1.):

Stage A – Word Level Emotion Prediction - The emotional value of each word is identified.

Stage B – Syntactic Graph Creation – The sentence is transformed into a graph that reflects the syntactic connections between words.

Stage C – Semantic Propagation Graph Neural Network – A specialized neural network processes these graphs, along with the emotional information of singular words, to predict the overall emotional meaning of a text, without relying on the direct knowledge of the words that constitute it.

Figure 1. *Steps of the Emotion Prediction Pipeline*



The proposed approach of selectively withholding specific semantic information from the model can be termed *Semantic Blinding*. Deliberately limiting the model's access to particular semantic details prevents it from associating emotional predictions with specific words or concepts that could introduce unwanted biases. By focusing exclusively on syntactic relationships and word-level emotional cues, semantic blinding ensures that the model's emotional assessments are free from training data biases related to specific groups or subjects. This technique, therefore, enhances the model's capacity to generalize across varied text sources without inheriting unintended, potentially harmful associations, providing a robust and unbiased tool for emotion prediction.

Brief Introduction to Graph Neural Networks

Graphs are mathematical structures used to represent entities (nodes) and their relationships (edges), providing a powerful framework for analyzing structured data (Zhou et al., 2020). In the context of natural language processing (NLP), a sentence can be represented as a graph where words serve as nodes, and edges capture syntactic or semantic relationships, such as dependencies between words. Graph Neural Networks (GNNs) extend this framework by applying deep learning techniques to graphs, enabling models to learn from the relationships and structures inherent in the data. Unlike traditional neural networks, which operate on fixed-size inputs like vectors or grids, GNNs analyze the connectivity patterns and features of nodes and edges to perform tasks such as classification, prediction, or clustering (Zhou et al., 2020). For example, GNNs have been successfully used in applications ranging from molecule property prediction in chemistry to fraud detection in financial networks (Motie & Raahemi, 2024; Wieder et al., 2020). They have also gained recognition in sentiment analysis applications (for a comprehensive review see Rad et al., 2023), however since the main aim of previous methods was to maximize the predictive performance of their approaches, none of them limited the amount of information that the model received, as is done in the case of Semantic Blinding.

The Contents of the Paper

Through comparative experiments, this paper demonstrates that the SProp GNN outperforms traditional lexicon-based models as well as lexicon-based alternatives across both discrete and dimensional emotion prediction tasks in English and Polish. It closely approaches transformer-based model accuracy in both languages, and task types offering a compelling alternative to biased black-box models. Furthermore, the paper provides detailed statistical and theoretical evidence that the SProp GNN is robust to the biases shown in previous research.

Methods

The Emotion Prediction System

The method proposed by the current paper essentially combines the use of three different machine learning based approaches, which process the text sequentially.

Word Level Emotion Prediction

The task of the first model is to predict the emotional value of the words in the text. This task, also seen as norm-, or lexicon-extrapolation is currently best attempted using transformer-based models (Plisiecki & Sobieszek, 2023). For the purposes of the current paper, either existing pretrained transformers norm extrapolation models are used, or new ones are trained when no off-the-shelf solutions are available. This stage results in emotion estimates for each separate word in a given text.

Creation of the Syntactic Graph

The text is then divided into sentences, and these sentences are analyzed using the *spaCy* package (Ines Montani et al., 2023). This software uses machine learning algorithms and linguistic rules to parse text, creating detailed syntactic structures for each sentence. *SpaCy* generates dependency graphs, which represent the relationships between words, as well as dependency labels (e.g. negations) and part-of-speech (POS) tags (e.g. verb). If a text consists of multiple sentences, these are linked back together using dedicated sentence nodes. This procedure allows the framework to capture the structure of each text, providing a type of scaffolding for the SProp GNN to propagate word-level emotional information through. This stage results in the creation of a syntactic graph, enriched with the information about the part-of-speech categories, and emotions from the first stage, at the node level, and dependency labels at the edge level.

Semantic Propagation Graph Neural Network

The final part of the pipeline is the SProp GNN, a neural network model designed to propagate emotional information extracted in the first stage of the pipeline, through the graph generated in the second stage. SProp GNN can be split into three main components: a custom SProp (GNN) layer, an attention pooling mechanism, and linear output layers.

Custom Graph Neural Network Layer: At the heart of the model is the custom Semantic Propagation Layer. This layer operates on the dependency graphs generated by spaCy (Ines Montani et al., 2023), where each node represents a word with associated features, and edges represent syntactic relationships between words. The Semantic Propagation Layer integrates information from the word node features (earlier predicted emotional load, and part-of-speech tags) and edge features (such as dependency types) to compute a scaling factor for each of its edges. It then propagates the emotional information from word nodes along those edges, scaling them accordingly. The hope here is that by doing so, it can model the propagation of emotional information through the sentence.

Attention Pooling Mechanism: After the graph has been processed by the SProp GNN layer, the model employs an attention-based pooling mechanism. This component aggregates the information from all nodes in the graph to create a single, fixed-size vector representation of the entire text. The attention mechanism assigns different weights to word nodes based on their relevance, effectively allowing the model to focus on the most significant words and relationships when forming this overall representation.

Linear Output Layers: The aggregated text representation is then passed through multiple linear layers. These layers transform the high-dimensional embedding into a scalar value between 0 and 1, corresponding to the predicted score for each predicted emotional metric. By having separate output layers for each metric, the model can simultaneously make multiple predictions, each tailored to the unique aspects of the respective psychological construct. This is different for discrete classification, where the layers transform the embedding to a vector of size equal to the number of predicted classes. This vector, when transformed, becomes an array of class probabilities.

The SProp GNN model processes text by first constructing a rich representation of its syntactic and semantic structure using the SProp layer. It then distills this information into a concise and meaningful summary via attention pooling. Finally, it translates this summary into actionable predictions through the linear output layers. Before prediction, this model has to be trained on a dataset of texts, with annotated emotional metrics in the form of either emotion intensities, or discrete emotion classes. As the model does not have direct access to the words with regards to which people exhibit social bias (e.g. certain politicians, gender information etc.), it cannot learn the association between them and the biased emotion estimates. Therefore, the biased part of an emotion estimate is from its perspective indistinguishable from noise, as it has no systematic relationship with the input data. This renders the model blind to the socially sensitive features of the input, therefore rendering it agnostic with regards to social biases. For a more detailed description of the model architecture see the Technical Appendix.

Comparative Experiments

The SProp model has been tested on three separate datasets, the GoEmotions dataset, the EmoBank dataset, and the dataset used in the Plisiecki and colleagues political bias study, referred to from now on as the Polish Political Dataset (2024). These datasets cover two languages (Polish and English), and two different emotion prediction tasks (categorical, and continuous emotion prediction).

The GoEmotions Dataset

The GoEmotions dataset, developed by Google researchers (Demszky et al., 2020), consists of around 58,000 English Reddit comments annotated with 28 distinct emotions, totaling over 210,000 annotations. Sourced from a Reddit data dump spanning 2005 to early 2019, the dataset includes comments from diverse subreddits, balanced by capping comment counts from

the most popular communities and sampling evenly across others. Emotion categories were curated based on psychological research to represent a broad but non-overlapping range of emotions. Each comment received annotations from three English-speaking raters from India, with additional raters assigned when agreement was low.

The EmoBank Dataset

The EmoBank dataset, created by Buechel and Hahn (2022), consists of 10,062 English sentences from sources like news, blogs, fiction, and letters, annotated along three emotional dimensions: Valence, Arousal, and Dominance (VAD). Each sentence was rated by five annotators from the crowdsourcing platform CrowdFlower for both *writer* and *reader* perspectives on a 5-point scale, giving insights into both expressed and perceived emotions. In accordance with the recommendations of the researchers, the current paper uses the version of the dataset with the weighted average of the reader and writer perspective labels provided at their online repository (*JULIELab/EmoBank*, 2017/2024).

The Polish Political Dataset

The Polish Political dataset (Plisiecki et al., 2024) includes 1.25 million Polish social media posts from journalists, politicians, NGOs, and general users. The emotionally neutral texts were filtered out using lexical norms on valence, arousal, and dominance. The final 10,000 texts were annotated by 20 psychology-trained annotators on six emotions (happiness, sadness, anger, disgust, fear, and pride) and two dimensions (valence and arousal) using a 5-point scale. Each annotator completed five weekly sets of 100 randomly assigned texts, ensuring each text was labeled by five raters for reliable coverage and minimizing cognitive fatigue over the five-week process. The resulting scores were averaged to create an intensity score for each text – emotion pair.

Dataset Preparation

Each of the dataset was first prepared by either calculating the most voted emotion category in the case of GoEmotions or normalizing the intensity of annotations to 0 to 1 range in the case of the two continuous datasets. Each dataset was then split into the training, evaluation, and test subsets in a proportion of 8:1:1, with the exception of the Polish dataset, for which the split dataset was taken from the original paper (Plisiecki et al., 2024). For more information about the preparation of the datasets and the datasets themselves see the Technical Appendix.

Comparative Approaches

The aforementioned datasets are used to compare the SProp model's performance to four alternative methods. The first three methods rely on lexicons, and as such are resilient to annotator bias. In order for the proposed framework to become a preferred alternative to them, it has to outperform them on evaluation metrics. The fourth method relies on transformer base models to predict emotions. It is added for comparison with high performing, but bias prone, models to better

inform researchers' decision-making. Each of the methods' performances is calculated on the test sets of respective datasets.

The Lexicon Approach

The lexicon approach works by averaging the emotional intensity of words in a given text. In the case of the Emobank dataset and the Polish political dataset I average the word ratings of previously published transformer-based norm extrapolation models (Plisiecki et al., 2024). In the results section I only report the results of averaging after removing stop words, as this method attained better results. As the EmoAtlas approach has proven superior to the lexicon approaches (Semeraro et al., 2023) in the task of discrete emotion prediction on the GoEmotions dataset, I do not report the performance of the lexicon approach for that specific task.

The Vader Approach

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a rule-based model designed for sentiment analysis, particularly effective in capturing sentiment from social media and informal text. VADER combines a lexicon with rules that account for various intensifiers, negations, and punctuation, making it particularly adept at assessing the sentiment intensity conveyed in short online texts. VADER assigns polarity scores for positive, neutral, and negative sentiment, averaging these scores to produce an overall sentiment value for a given text. This approach is only capable of producing valence estimations (Hutto & Gilbert, 2014).

The EmoAtlas Approach

The EmoAtlas utilizes an extensive lexicon-based network to profile emotions by mapping syntactic and semantic relationships in text, effectively capturing nuanced emotional cues without extensive model training. Using validated emotional lexicons for Plutchik's eight core emotions in conjunction with a spaCy based (Ines Montani et al., 2023) syntactic analysis, it efficiently identifies emotional tones in multiple languages. Its rule-based structure enables it to run significantly faster than transformer-based models, providing researchers with interpretable insights into how emotions are conveyed in text associations (Semeraro et al., 2023).

The Transformer Approach

A base transformer can be finetuned to predict both continuous and categorical emotions. Here the *roberta-base* transformer model developed by Facebook (Liu et al., 2019) is finetuned on the two aforementioned English datasets. After a hyperparameter sweep for each dataset, the final models were trained on the parameter setup that led to the best performance. For the Polish political bias dataset, the performance of the GNN model is compared with a transformer model that was finetuned to predict emotions in the original dataset paper (Plisiecki et al., 2024). For a more detailed description of the implementation of each of the comparative approaches refer to the Technical Appendix.

Testing for Bias

This section evaluates whether the SProp GNN model mitigates overfitting on biases present in the training data. It follows the approach established by Plisiecki et al. (2024) [1], who analyzed a transformer-based language model's predictions of sentiment toward 24 prominent Polish politicians. The original study selected politicians based on a public trust survey and tested the model's sentiment responses to three types of text stimuli: (1) the politicians' names alone, (2) politically charged sentences containing these names, and (3) neutral sentences featuring the same names. The model's task was to classify each stimulus as having positive or negative valence.

To quantify political bias, Plisiecki et al. (2024) fit linear regression models in which the model's predicted valence was the dependent variable (Y), and the politician's political affiliation (a dummy-coded factor) and gender were predictors (X1: political affiliation, X2: gender). Thus, the model took the form:

$$Y = \beta_0 + \beta_1(\text{political affiliation}) + \beta_2(\text{gender}) + \epsilon$$

They found that these predictors explained a substantial proportion of the variance in valence predictions (52% for neutral sentences, 66% for political sentences, and 67% for names alone), indicating significant bias.

The current study replicates and extends this approach with the SProp GNN model using three complementary methods. The goal is to determine whether SProp GNN exhibits significantly less bias than the previous model.

Approach 1: Replicating the Original Regression Procedure

The first approach directly replicates the original regression methodology. The same linear regression model is applied:

$$Y_{\text{SProp}} = \beta_0 + \beta_1(\text{political affiliation}) + \beta_2(\text{gender}) + \epsilon$$

Here, Y_{SProp} represents the valence predictions made by the SProp GNN for each stimulus. The null hypothesis (H_0) states that the SProp GNN model does not exhibit bias (i.e., $\beta_1 = \beta_2 = 0$) while the alternative hypothesis (H_1) is that at least one of the bias coefficients is non-zero.

To test this, a permutation test on the observed valence predictions is employed. The correspondence between stimuli and the predictor values (political affiliation, gender) is randomly shuffled 100,000 times and the regression is re-estimated each time. This produces a null distribution of test statistics (e.g., F-statistics or sums of squared residuals) (Manly, 1997). If the observed statistic falls into the extreme tails of this distribution, the H_0 is rejected with the conclusion that the SProp GNN exhibits bias. Non-significant results should be interpreted with caution, as it is difficult to ascertain the test's power precisely.

[1] The author thanks Paweł Lenartowicz for help in coming up with the statistical tests required to test the SProp model's bias

Approach 2: Assessing Bias Reduction in the SProp Model

The second approach aims to determine whether the SProp GNN model reduces bias compared to the original model. Instead of testing if bias exists, the test related to whether the SProp model's bias is equivalent to (or less than) that of the original transformer model.

First, the SProp GNN predictions are adjusted by removing the estimated bias from the original model. To do this, the original model's estimated bias coefficients are used ($\widehat{\beta_1}$ and $\widehat{\beta_2}$) to create adjusted predictions:

$$Y_{\text{SProp, adjusted}} = Y_{\text{SProp}} - \left(\widehat{\beta_1} \cdot \text{political affiliation} + \widehat{\beta_2} \cdot \text{gender}\right)$$

Next, a regression is performed with the original bias factors as predictors on these adjusted scores. If the adjusted SProp predictions still show a significant relationship with the bias factors, it means the SProp model retained the same pattern of bias. If the bias factors do not predict the adjusted valence (or predict it inversely), it suggests bias has been reduced.

The null hypothesis (H_0) for this approach states that the SProp model's bias is the same as the original model's bias. The alternative hypothesis (H_1) suggests that the SProp model's bias is reduced. This is tested using a one-sided permutation test (100,000 random assignments). A statistically significant negative beta coefficient would indicate that the SProp model is inversely related to the original bias factor, signifying bias reduction.

Approach 3: Comparing Differences Between Models

The third approach examines the difference in predictions between the transformer model and the SProp model. Define the difference in predicted valence as:

$$\Delta Y = Y_{\text{SProp}} - Y_{\text{transformer}}$$

This difference is then regressed on the original bias factors:

$$\Delta Y = \gamma_0 + \gamma_1(\text{political affiliation}) + \gamma_2(\text{gender}) + \epsilon$$

The null hypothesis (H_0) is that the difference in predictions between models is unrelated to the bias factors ($\gamma_1 = \gamma_2 = 0$). The alternative hypothesis (H_1) is that these factors significantly predict the difference, confirming that bias is driving the disparities between models.

As before, a one-sided permutation test (100,000 random assignments) is conducted to determine whether the observed association differs from what would be expected by chance. A significant result would indicate that bias plays a key role in differentiating the two models.

Note on Interpreting the Results

While these methods help determine whether the SProp GNN model reduces bias, it is important to recognize that errors in the original model's estimated bias parameters may attenuate the observed relationships in the SProp model. Due to such estimation errors, the bias parameters

in the new tests are not expected to reach exactly 1 (or to show a perfect elimination of bias). Even if no bias were present in the SProp model, random measurement errors and attenuation effects may prevent the parameters from perfectly reflecting the removal of bias.

Explainability

For the sake of explainability, the SProp GNN saves its scaling factors as well as attention weights, allowing the user to understand the type of information on which the model based its decisions. As a full analysis of how the model reacts to a large array of diverse sentences is beyond the scope of this paper, the focus is shifted to explaining the basic mechanics using two sentences, employing an emotional word and a negation: “I am happy” and “I am not happy.” The activity of the model is then compared between these sentences.

This explainability approach is particularly important because it allows users to assess not only the model's outputs but also the reasoning behind them. By exposing the scaling factors and attention weights, it becomes possible to pinpoint how specific words and their relationships, such as the negation in “I am not happy,” influence the emotional predictions. This transparency is crucial in ensuring trust and interpretability in sentiment analysis models, especially for applications where ethical considerations or fairness are paramount.

Results

The GoEmotions Dataset

In the task of discrete emotion prediction conducted on the GoEmotions dataset, the Semantic Propagation GNN generally outperforms the EmoAtlas approach across the three key performance metrics: accuracy, precision, and recall (See Table 1.). Here, accuracy measures how often the model's predictions are correct overall, precision assesses the proportion of correct positive predictions among all positive predictions made, and recall evaluates the model's ability to identify all actual instances of each emotion. The only exceptions are in the precision metric for the emotions of anger and disgust, where Emo Atlas slightly exceeds the GNN.

While the SProp GNN shows better performance than Emo Atlas, both methods are generally surpassed by their transformer-based counterpart. The RoBERTa model, which leverages advanced language representations, outperforms both the Emo Atlas and the Semantic Propagation GNN across all emotions and metrics. However, the GNN is not far behind RoBERTa, achieving a mean accuracy difference of only 5.70 percentage points, compared to a difference of 20.73 percentage points between RoBERTa and the Emo Atlas.

Similarly, for precision, the average difference between RoBERTa and the GNN is 17.05 percentage points, whereas the difference between RoBERTa and the Emo Atlas is 30.33 percentage points. In terms of recall, the average difference between RoBERTa and the GNN is 20.81 percentage points, while the difference between RoBERTa and the Emo Atlas is 48.46

percentage points. These results indicate that while the GNN approaches the performance of RoBERTa, the Emo Atlas method lags significantly behind in all three metrics.

Table 1.

Performance results on the goemotion dataset

emotion	Accuracy score %			Precision score %			Recall score %		
	roberta	emoa	sprop	roberta	emoa	sprop	roberta	emoa	sprop
anger	91.3	70.0	**80.0**	86.7	**70.1**	65.9	87.8	33.8	**82.3**
disgust	93.9	66.8	**88.9**	78.2	**73.7**	50.7	63.0	19.9	**35.1**
fear	94.7	77.3	**93.2**	71.2	39.6	**72.1**	85.6	48.2	**59.6**
joy	97.3	73.7	**92.2**	93.0	70.2	**76.4**	90.7	47.5	**77.8**
sadness	94.8	71.6	**88.9**	81.4	52.3	**61.5**	80.2	35.5	**48.9**
surprise	96.1	84.3	**90.7**	85.1	7.7	**66.7**	86.4	18.0	**65.1**

Note. The emotion categories had to be limited to those presented in the table both due to lexicon availability and EmoAtlas emotion coverage. The results written in bold pinpoint the best performance in a given metric out of the two alternatives to transformers. The metric results for the EmoAtlas were taken from the original manuscript, which introduced the technique. While that means that they were tested on a wider test set, it still provides a good overview of the approach performance given that it does not require any finetuning. Model codes: roberta – finetuned transformer; emoatlas – the Emo Atlas approach; sprop – Semantic Propagation GNN model.

The EmoBank Dataset

The task of sentence level emotion prediction was run on the EmoBank dataset, on which the SProp GNN outperformed both the lexicon approach and the Vader approach (see Table 2.). While RoBERTa achieved higher scores than the SProp GNN, the degree of difference varied between predicted metrics. While in the case of valence the difference amounted to 0.13 points, for arousal it was as low as 0.02.

Table 2.

Performance results on the emobank dataset

Metric	roberta	lexicon	vader	sprop
Valence	0.75	0.45	0.46	**0.62**
Arousal	0.48	0.25	-	**0.45**

Note. The results written in bold pinpoint the best performance, measured using the Pearson's correlation, in a given metric out of the three alternatives to transformers. Model codes: roberta – finetuned transformer; vader – the Vader approach; lexicon – the lexicon approach; sprop – Semantic Propagation GNN model. The lexicon score has been calculated after pruning stopwords.

The Polish Political Dataset

Finally, the Polish political dataset was used to test the model performance on a mixed set of both multiple and single sentence texts. Here, the SProp outperformed its lexicon counterpart yet again (See Table 3.). Unsurprisingly it was at the same time worse at predicting emotion scores than the RoBERTa model by 0.16 points in the case of valence, and 0.13 points in the case of arousal.

Table 3.

Performance results on the polish political dataset (Pearson's Correlation)

	roberta	lexicon	sprop
Valence	0.88	0.57	**0.72**
Arousal	0.75	0.33	**0.62**

Note. The results written in bold pinpoint the best performance in a given metric, measured using the Pearson's correlation, out of the two alternatives to transformers. Model codes: roberta – finetuned transformer; lexicon – the lexicon approach; sprop – Semantic Propagation GNN model. The lexicon score has been calculated after pruning stopwords.

Political Bias Results

Approach 1: Replicating the Original Regression Procedure

The replication of regressions performed in the original bias study (Plisiecki et al., 2024) yielded no significant results for the SProp GNN model (see Table 4). Neither political affiliation nor gender explained any meaningful variance in the model's valence predictions across any of the stimuli categories: names, neutral sentences, and political sentences. This outcome contrasts sharply with the results obtained for the transformer model in the original study, where political affiliation and gender were significant predictors, explaining 66% of the variance in valence for political sentences, 52% for neutral sentences, and 67% for names alone. For the SProp model, these same predictors explained a negligible proportion of the variance, as shown by the R^2 values of 0.077, 0.135, and 0.103, which are accompanied by non-significant permutation test p-values.

In Table 4, the results are presented for both models. The regression intercepts represent the predicted valence for the reference group, which is Zjednoczona Prawica (the ruling party) and male politicians, while the coefficients for political affiliation indicate how much valence changes for other groups (e.g., Konfederacja, Koalicja Obywatelska, etc.). For the transformer model, the political affiliation coefficients are consistently significant across all stimuli types, confirming substantial bias in its predictions. For example, the valence associated with Koalicja Obywatelska is consistently higher than that for Zjednoczona Prawica, with coefficients such as 5.83 (neutral sentences) and 2.30 (names only), both significant at $p < 0.05$. Gender also has a notable influence in the transformer model, with a coefficient of 9.77 for neutral sentences, indicating that valence predictions for women are substantially more positive than for men in this category.

In contrast, the SProp GNN model shows no significant coefficients for political affiliation or gender in any stimulus category. For instance, the coefficient for Koalicja Obywatelska is close to zero (e.g., 0.18 for neutral sentences, 0.93 for names, and 0.60 for political sentences) and accompanied by p-values well above 0.05. Similarly, the gender coefficient is 3.19 for neutral sentences, 2.11 for political sentences, and while as high as 8.10 for names only, it is not statistically significant. These results suggest that the SProp GNN model's predictions are less systematically influenced by political affiliation and gender, highlighting a potential reduction in bias compared to the transformer model.

Despite these findings, it is important to interpret the results with caution. While the lack of statistical significance in the SProp GNN model suggests an absence of systematic bias, this alone does not confirm that the model is entirely unbiased. The low explanatory power of the regressions and the non-significant results may also reflect limitations in the sensitivity of the statistical tests or the sample size, rather than the true absence of bias. Moreover, differences in residual variance and standard errors between the models indicate that additional factors may be influencing the outcomes. Therefore, complementary analyses, such as those presented in later sections, are essential to provide a more comprehensive understanding of bias reduction in the SProp GNN model.

Table 4.
Regression models – differences in valence

	Transformer			Semantic Propagation GNN		
	Dependent variable: Valence of:					
	Names only (1)	Neutral Sentences (2)	Political Sentences (3)	Names only (1)	Neutral Sentences (2)	Political Sentences (3)
intercept	45.40***	48.61***	43.89***	53.59***	55.02***	44.96***
	(0.67)	(0.92)	(0.26)	(1.87)	(0.71)	(0.49)
3D	5.73**	9.09**	2.72**	2.33	1.44	0.73
	(2.37)	(3.26)	(0.93)	(6.61)	(2.52)	(1.71)
K	6.15*	8.37*	2.56*	3.85	0.83	1.02
	(3.10)	(4.28)	(1.22)	(8.67)	(3.30)	(2.24)
KO	5.83***	3.71*	2.30**	0.93	0.18	0.60
	(1.31)	(1.80)	(0.51)	(3.66)	(1.39)	(0.94)
Left	3.03	5.46	2.51**	-3.75	-3.02	-1.14
	(2.56)	(3.53)	(1.01)	(7.17)	(2.73)	(1.85)
gender	9.77**	10.10*	1.88	8.10	3.19	2.11
	(3.48)	(4.80)	(1.37)	(9.74)	(3.71)	(2.52)
Observations	22	22	22	22	22	22
R^2	0.665	0.520	0.662	0.077	0.135	0.103
Adjusted R^2	0.547	0.370	0.556	-0.211	-0.135	-0.178
Residual Std. Error	3.38 (df=16)	5.21 (df=16)	1.29 (df=16)	6.69 (df=16)	2.75 (df=16)	1.70 (df=16)
P-value (permutation)	0.008***	0.049**	0.018**	0.885	0.686	0.793

Note. *p<0.1; **p<0.05; ***p<0.01 (*t*-test)
Zjednoczona Prawica (ruling party) as intercept, gender: woman=1, man=0
Abbreviations: K – Konfederacja, 3D – Trzecia Droga, KO – Koalicja Obywatelska, Left – Nowa Lewica

Approach 2: Assessing Bias Reduction in the SProp Model

The beta coefficient for bias was β = -0.78, with a permutation p-value significant at p = 0.012. This result suggests that the SProp model's valence predictions exhibit a substantially lower level of bias compared to those of the transformer model. The negative value of the coefficient indicates an inverse relationship, suggesting that the bias introduced by political affiliation and gender in the original model has been largely mitigated in the SProp GNN model. Given the potential for real-world measurement variability and the effects of regression dilution, a coefficient of -0.78 strongly implies that the SProp GNN has no significant residual bias from the original model, or that any remaining bias is minor and unlikely to have practical significance.

Table 5 presents the regression results, including the adjusted R^2 of 0.535, indicating a moderate fit for the model. While the dummy variables for sentence types were included to account for systematic differences between neutral and political sentences, their specific coefficients are not central to the interpretation of bias reduction. The key finding remains that the SProp GNN model shows a marked reduction in bias, as evidenced by the negative and significant beta coefficient for the bias factor.

Table 5.
Regression model - testing for reduction in bias

	Dependent variable: Bias Reduced SProp GNN Valence Predictions
intercept	53.58***
	(0.96)
bias	-0.78***
	(0.20)
neutral Sentences	-1.07
	(1.21)
political Sentences	-8.62***
	(1.24)
Observations	66
R^2	0.557
Adjusted R^2	0.535
Residual Std. Error	4.04
	(df=62)

Note. *p<0.1; **p<0.05; ***p<0.01 (*t*-test)

Approach 3: Comparing Differences Between Models

The beta coefficient in this analysis explained a significant portion of the variance in valence prediction differences between the two models, with β = 0.78 and a permutation p-value of p = 0.028. This finding strongly supports the conclusion that the SProp model propagates substantially less bias related to political affiliation and gender compared to the transformer model. The positive and significant coefficient indicates that the difference in predictions between the two models is systematically related to the bias factors identified in the transformer model, further highlighting that the SProp GNN effectively reduces the bias originally observed.

Table 6 presents the results of this regression. The intercept (-8.18, $p < 0.001$) represents the baseline difference between the models' predictions, while the bias coefficient ($\beta = 0.78$, $p < 0.01$) accounts for a significant portion of the variance. The inclusion of dummy variables for neutral and political sentences adjusts for systematic differences across stimulus types. While these coefficients (neutral: $\beta = 2.14$, $p < 0.1$; political: $\beta = 7.11$, $p < 0.001$) suggest some variation in the magnitude of prediction differences based on sentence type, the primary finding lies in the bias coefficient itself, which demonstrates the central role of bias reduction in distinguishing the SProp GNN's predictions from those of the transformer model.

Table 6.

Regression model – explaining the difference in predictions

	Dependent variable: Difference between Transformer and SProp GNN Valence Predictions
intercept	-8.18***
	(0.94)
bias	0.78***
	(0.19)
neutral Sentences	2.14*
	(1.19)
political Sentences	7.11***
	(1.21)
Observations	66
R^2	0.415
Adjusted R^2	0.386
Residual Std. Error	4.368
	(df=62)

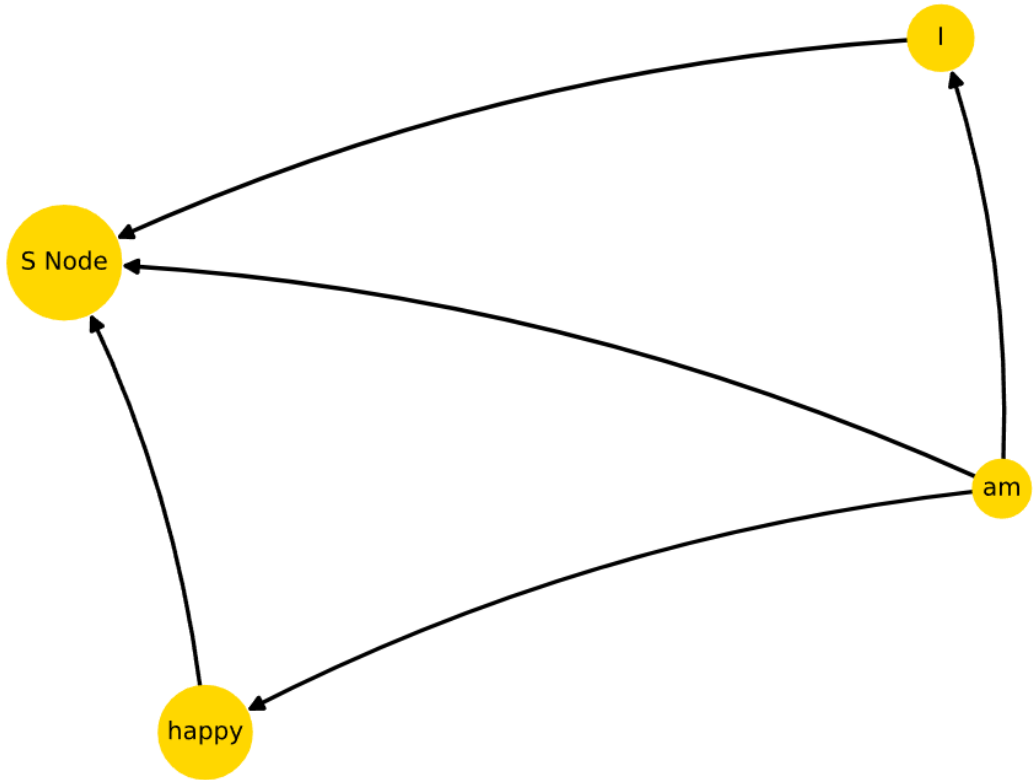
Note. *p<0.1; **p<0.05; ***p<0.01 (*t*-test)

The analysis tested three hypotheses to evaluate the bias robustness of the SProp GNN model compared to the transformer model. The null hypothesis about the SProp GNN being robust to bias was retained, but the results were inconclusive. The second hypothesis, which posited that the SProp model's bias is equivalent to the transformer model's, was rejected, showing a significant reduction in bias. The third hypothesis, testing whether prediction differences between the models are related to bias factors, was also rejected, confirming that the SProp model mitigates the biases observed in the transformer model. These results strongly support the conclusion that the SProp GNN substantially reduces bias and is a reliable alternative for unbiased sentiment analysis.

Explainability

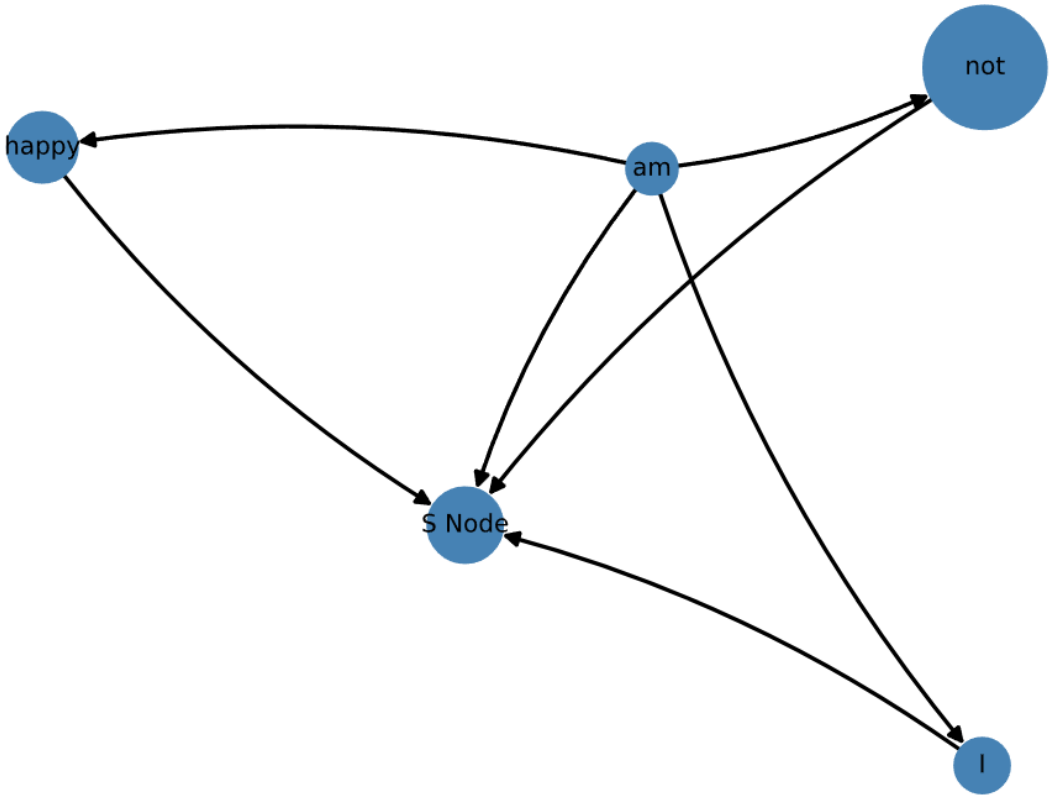
The SProp model trained on the EmoBank dataset was used to assess the valence of two sentences: “I am happy”, and “I am not happy”. The former sentence received a valence prediction of 0.68, and the latter, a valence prediction of 0.43 indicating that the model is able to take negation into account and appropriately modify its prediction. Figures 2, and 3 depict an abstracted representation of what happened inside the model during the prediction. The size of the nodes symbolizes the extent to which the model paid attention to them during prediction, and the arrows symbolize edges through which the emotional information was propagated.

Figure 2. The explanatory graph for sentence “I am happy”



Note. The size of the nodes represents the degree to which the model attended to a given node's feature when extracting information from the graph. The S Node refers to the sentence node.

Figure 3. The explanatory graph for sentence "I am not happy"



It can be seen that in the case of the first sentence (Figure 2), the model relied on the emotional information from the sentence node, and the node that contained the emotional features of the word "happy". This is expected as the sentence node contains information passed from all of the word nodes in the sentence, while the "happy" node is an adjective, and so often conveys emotional information. In the case of the second sentence, however (Figure 3) the model paid significantly more attention to the "not" node features, indicating that it learned that negation could reverse the emotional load of a sentence.

However, as can be seen in Figure 3, the pathway of emotional propagation from the word "happy" does not include the negation node. This means that the mechanism through which SProp GNN operates is partially based on heuristics, rather than just on the propagation of emotional information through the syntactic graph. To test this, a longer sentence including a negation that should not modulate the emotional information was processed by the model. The sentence "I am happy, and not tall" was given a valence prediction of 0.431, while the same sentence "I am happy, and tall" received a rating of 0.69.

The scaling factors used to propagate the emotional information through the syntactic graph were not visualized, as they do not convey a lot of information. This is due to their high dimensionality. In the current models, the dimension of the scaling factors is equal to the hidden dimension of the SProp Layer – 512. This means that their average in all likelihood obscures a lot of information about their underlying mechanisms.

Discussion

The present study introduces the Semantic Propagation Graph Neural Network (SProp GNN) as a novel approach to emotion prediction, addressing the critical issue of bias propagation inherent in machine learning based models. The SProp GNN significantly outperformed other lexicon-based alternatives across two different languages—English and Polish—and two distinct emotion prediction tasks, namely discrete and dimensional emotion prediction. It's ability to utilize the syntactic structure of sentences embedded with emotional information on the single word level allowed it to bridge the gap between simple lexicon-based methods and complex black-box models.

The main contribution of this work is the demonstrable reduction of bias in emotion prediction. Such bias can lead to unfair or discriminatory outcomes, both in real world applications such as mental health assessments (Parikh et al., 2019), hiring processes (Kassir et al., 2023), or criminal justice systems (Joseph, 2024), as well as in the academia, where scientific conclusions are required to be fair and objective. The statistical evidence presented shows that the SProp GNN propagates at least significantly less bias than its transformer-based counterpart. This evidence, coupled with the sole fact that the SProp GNN simply does not have access to any bias information it could overfit, since it only processes the syntactic structure of the sentence coupled with external emotional ratings and not words directly, warrants a claim that it is robust to training data bias. By effectively mitigating bias, the SProp GNN not only enhances the fairness and ethical standing of emotion prediction tools but also increases their reliability across diverse populations. This is particularly crucial in a global context where texts may reflect a wide array of cultural, social, and individual differences. The ability of the SProp GNN to provide more objective emotional assessments can contribute to more equitable decision-making processes in applications that rely on sentiment analysis.

While the SProp GNN performs slightly worse than pretrained transformers, it constitutes a viable alternative in applications where objective emotional assessment is of key importance. The development of the SProp GNN highlights the inherent trade-offs between model interpretability, performance, and bias mitigation. Transformer-based models often achieve higher accuracy due to their ability to capture complex semantic nuances; however, they also tend to act as black boxes, making it difficult to understand or control the sources of their predictions, including biases. In contrast, the SProp GNN offers a more interpretable architecture allowing for greater transparency in how predictions are made. Although there is a slight decrease in performance compared to transformers—a mean accuracy difference of 5.70 percentage points on

the GoEmotions dataset, for instance—this trade-off may be acceptable or even preferable in contexts where interpretability and bias reduction are prioritized over marginal gains in accuracy. This balance underscores the importance of aligning model selection with the specific requirements and ethical considerations of the intended application.

The bias robustness of the SProp GNN makes it particularly suitable for applications where fairness and objectivity are paramount. For instance, in the analysis of social media data for public health monitoring (Babu & Kanaga, 2022), using a model that minimizes bias ensures that interventions are based on accurate representations of population sentiments without skewing toward or against specific groups. In mental health contexts, such as suicide risk prediction from text messages (Glenn et al., 2020), unbiased sentiment analysis can lead to more accurate assessments and timely interventions. Increasingly popular tools that analyze student feedback (Dalipi et al., 2021) can also benefit from unbiased emotion assessments to foster an inclusive learning environment. In all these cases and more, the SProp GNN's ability to deliver high-performance emotion predictions while mitigating bias is of significant practical value.

Limitations and Future Research

The SProp GNN seems susceptible to simplistic heuristics, as shown in the explainability section, where the model did not fully capture the nuanced role of negations in complex sentences. This potential discrepancy between the good performance results of the model and its inability to pick up on syntactic cues that are easily understandable to a human can be explained by the low frequency of such sentence structures in the tested datasets and their overall rarity. The datasets used are diverse in language and task types, but they may not encompass the full spectrum of linguistic structures and expressions found in real-world texts. This could limit the generalizability of the SProp GNN to other languages or dialects not represented in the training data.

Future studies can further improve the architecture of the SProp GNN, shrinking the gap between its performance and that of other potentially biased models. One potential avenue for further exploration could be a modification of the syntactic graph creation algorithm. While the syntactic pathways extracted using the spaCy package (Ines Montani et al., 2023) provide useful information about the structure of sentences, a more tailored model that would map the pathway of emotional information propagation directly could achieve even better results, potentially reducing the model’s reliance on heuristics. Expanding the training datasets to include more syntactically diverse sentences could help the model learn to handle complex linguistic structures more effectively.

Furthermore, the model's reliance on lexicons, or norm-extrapolation models could introduce bias present on the word level. Despite the lack of context when annotating emotions at word level, there is still a small possibility that the annotators for the lexicons impacted some sorts of biases on the emotion dictionary. This type of bias, however, can usually be easily explored using the lexicon in question, and its mitigation is as simple as equalizing the emotional load of words that convey it. In circumstances where a specific type of bias could directly impact the

conclusions of a study, checking the lexicon for its presence before the use of the SProp GNN is advisable.

Expanding the Concept of Semantic Blinding

The proposed approach of selectively withholding specific semantic information from the model, termed semantic blinding, is a technique that deliberately limits the model's access to particular semantic details. By preventing the model from associating emotional predictions with specific words or concepts that could introduce unwanted biases, semantic blinding ensures that the model's emotional assessments are free from training data biases related to specific groups or subjects. This technique presents exciting opportunities for future research. It could be extended to other natural language processing tasks where bias could be a concern, such as text-classification. Exploring how semantic blinding can be integrated with transformer-based architectures might also yield models that combine the high performance of transformers with the bias mitigation benefits of the SProp GNN. Additionally, further investigation into the types of semantic information that can be withheld without significantly impacting performance could lead to the development of more robust and fair NLP models across various domains.

Practical Considerations for Deployment

From a practical standpoint, deploying the SProp GNN in real-world applications offers significant advantages in terms of computational efficiency and scalability due to its substantially smaller model size when compared to its transformer-based counterparts. Specifically, the SProp GNN trained on the EmoBank dataset consists of approximately 1.5 million parameters, while the transformer model trained for the same task comprises about 125 million parameters. This significant reduction in model size—over 20 times smaller—translates to lower computational overhead and faster processing times, making the SProp GNN more suitable for deployment on devices with limited resources or for applications requiring real-time analysis. For such applications, the word level prediction stage of the model could be done prior to inference time by generating a very large emotional dictionary a priori. From an academic standpoint, this translates to accessibility for researchers without access to high-performance computing resources. The reduced memory and processing requirements mean that the SProp GNN can be trained and deployed on standard hardware, broadening the scope of researchers who can experiment with and apply this model. This adaptability and efficiency make the SProp GNN a practical and accessible alternative to transformer-based models in sentiment analysis tasks.

In order to allow other researchers to replicate the analyses presented in the current paper and use the SProp GNN architecture for their research, the code, along with detailed comments for this paper has been made available at a GitHub repository (https://github.com/hplisiecki/Semantic-Propagation-GNN). Additionally, the Technical Appendix should serve as additional guide for those willing to apply and further develop the methods here presented.

Conclusion

The SProp GNN represents a significant step forward in developing sentiment analysis models that prioritize fairness and interpretability without substantial sacrifices in performance. The evidence demonstrates that the SProp GNN not only approaches the accuracy of transformer-based models but also at least significantly reduces the propagation of biases. This coupled with the fact that the model does not possess the ability to overfit specific words points towards near total bias eradication. By addressing the critical issue of bias propagation, the SProp GNN offers a viable and ethically sound alternative for a wide range of applications. While there is room for improvement, particularly in handling complex syntactic structures and expanding language coverage, the SProp GNN lays the groundwork for future advancements in unbiased and interpretable sentiment analysis. Future work focused on enhancing the model's architecture, expanding its applicability, and refining the semantic blinding technique holds the promise of further bridging the gap between high performance and bias mitigation in natural language processing

Funding

This research is funded by a grant from the National Science Centre (NCN) 'Research Laboratory for Digital Social Sciences' (SONATA BIS-10, No. UMO-020/38/E/HS6/00302).

References

Babu, N. V., & Kanaga, E. G. M. (2022). Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review. *SN Computer Science*, *3*(1), 74. https://doi.org/10.1007/s42979-021-00958-1

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical report C-1, the center for research in psychophysiology ….

Buechel, S., & Hahn, U. (2022a). *EmoBank* [Dataset]. https://github.com/JULIELab/EmoBank

Buechel, S., & Hahn, U. (2022b). *EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2205.01996

Dalipi, F., Zdravkova, K., & Ahlgren, F. (2021). Sentiment Analysis of Students' Feedback in MOOCs: A Systematic Literature Review. *Frontiers in Artificial Intelligence*, *4*, 728708. https://doi.org/10.3389/frai.2021.728708

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). *GoEmotions: A Dataset of Fine-Grained Emotions* (arXiv:2005.00547). arXiv. https://doi.org/10.48550/arXiv.2005.00547

Diaz, M., Johnson, I., Lazar, A., Piper, A. M., & Gergle, D. (2018). Addressing Age-Related Bias in Sentiment Analysis. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173986

Glenn, J. J., Nobles, A. L., Barnes, L. E., & Teachman, B. A. (2020). Can Text Messages Identify Suicide Risk in Real Time? A Within-Subjects Pilot Examination of Temporally Sensitive Markers of Suicide Risk. *Clinical Psychological Science*, *8*(4), 704–722. https://doi.org/10.1177/2167702620906146

Hills, T. T., Proto, E., Sgroi, D., & Seresinhe, C. I. (2019). Historical analysis of national subjective wellbeing using millions of digitized books. *Nature Human Behaviour*, *3*(12), 1271–1275. https://doi.org/10.1038/s41562-019-0750-z

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), Article 1.

Imbir, K. K. (2016). Affective norms for 4900 Polish words reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Frontiers in Psychology*, *7*, 1081.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, & Henning Peters. (2023). *explosion/spaCy: V3.7.2: Fixes for APIs and requirements* (Version v3.7.2) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.1212303

Joseph, J. (2024). Predicting crime or perpetuating bias? The AI dilemma. *AI & SOCIETY*, s00146-024-02032–02039. https://doi.org/10.1007/s00146-024-02032-9

JULIELab/EmoBank. (2024). [Jupyter Notebook]. JULIE Lab. https://github.com/JULIELab/EmoBank (Original work published 2017)

Kassir, S., Baker, L., Dolphin, J., & Polli, F. (2023). AI for hiring in context: A perspective on overcoming the unique challenges of employment research to mitigate disparate impact. *AI and Ethics*, *3*(3), 845–868. https://doi.org/10.1007/s43681-022-00208-x

Kiritchenko, S., & Mohammad, S. M. (2018). *Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems* (arXiv:1805.04508). arXiv. https://doi.org/10.48550/arXiv.1805.04508

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

Manly, B. F. J. (2018). *Randomization, Bootstrap and Monte Carlo Methods in Biology* (0 ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315273075

Mohammad, S. M. (2017). *Word Affect Intensities* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1704.08798

Motie, S., & Raahemi, B. (2024). Financial fraud detection using graph neural networks: A systematic review. *Expert Systems with Applications*, *240*, 122156. https://doi.org/10.1016/j.eswa.2023.122156

Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing Bias in Artificial Intelligence in Health Care. *JAMA*, *322*(24), 2377. https://doi.org/10.1001/jama.2019.18058

Plisiecki, H., Koc, P., Flakus, M., & Pokropek, A. (2024). *Predicting Emotion Intensity in Polish Political Texts: Comparing Supervised Models and Large Language Models in a Resource-Poor Language* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2407.12141

Plisiecki, H., & Sobieszek, A. (2023). Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection. *Behavior Research Methods*, *56*(5), 4716–4731. https://doi.org/10.3758/s13428-023-02212-3

Rad, R. A., Yamaghani, M. R., & Nourbakhsh, A. (2023). *A survey of sentiment analysis methods based on graph neural network*. https://doi.org/10.21203/rs.3.rs-3173515/v1

Ramteke, J., Shah, S., Godhia, D., & Shaikh, A. (2016). Election result prediction using Twitter sentiment analysis. *2016 International Conference on Inventive Computation Technologies (ICICT)*, *1*, 1–5. https://doi.org/10.1109/INVENTIVE.2016.7823280

Semeraro, A., Vilella, S., Mohammad, S., Ruffo, G., & Stella, M. (2023). *EmoAtlas: An emotional profiling tool merging psychological lexicons, artificial intelligence and network science*. https://doi.org/10.21203/rs.3.rs-2428155/v1

Wang, J., Fan, Y., Palacios, J., Chai, Y., Guetta-Jeanrenaud, N., Obradovich, N., Zhou, C., & Zheng, S. (2022). Global evidence of expressed sentiment alterations during the COVID-19 pandemic. *Nature Human Behaviour*, *6*(3), 349–358. https://doi.org/10.1038/s41562-022-01312-y

Widmann, T., & Wich, M. (2022). Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text. *Political Analysis*, 1–16. https://doi.org/10.1017/pan.2022.15

Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., & Langer, T. (2020). A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, *37*, 1–12. https://doi.org/10.1016/j.ddtec.2020.11.009

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001

Technical Appendix

This technical appendix provides detailed information on the methodologies, models, datasets, and experimental setups used in the paper. It is intended to offer in-depth insights that supplement the main text, as well as to serve as a guide for training similar models in the future.

Detailed Emotion Prediction Pipeline

The GNN model proposed in the manuscript consists of three stages:

1. Word level emotion prediction
2. Syntactic graph creation
3. The Semantic Propagation GNN (SProp GNN)

Below each of these stages are described in detail

Word Level Emotion Prediction

The model relies on knowing the emotions of every, or most words in a text, to then propagate this information through the syntactic graph and predict emotion on the text level. The paper, in order to predict the emotions of words, draws on the literature in norm extrapolation (Plisiecki & Sobieszek, 2023) which recommends the use of transformer models for word level emotion prediction. The use of these models might not be necessary given a large enough lexicon of words and their respective emotional values, and a corpus with restrained vocabulary. However, to ensure proper word coverage already trained transformer norm extrapolation models are used, or, in the case of discrete emotion prediction, new ones are trained

Transformer based norm extrapolation models are trained by adding a regression or a classification head to a transformer encoder and training it on an existing norm lexicon. Previous research has also added an additional hidden layer between the encoder, and the regression head, with dropout. To create one for the task of discrete emotion prediction the NRC Emotion Intensity Lexicon was used (Mohammad, 2017) which provides ratings on a scale of 0 to 1 for 5891 unique words for eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust). Not all of the words were rated with regards to each emotion, which required training eight separate prediction models for each of the emotions. The "nghuyong/ernie-2.0-en" model was used for each of them, as this was the model that was also used for valence and arousal prediction in a previous paper and thus can be trusted to model emotional information well (Plisiecki & Sobieszek, 2023). The lexicon was split into seven emotion specific lexicons, and each of these subcorpora was then further divided into train, evaluation, and test sets in the ratio of 8:1:1. Each model was trained for 100 epochs, with a batch size of 500, learning rate of 5e-5, AdamW optimizer with a weight decay of 0.3 and a linear learning rate schedule with warmup steps amounting to 600, a hidden 768 dimensional hidden layer and a dropout of 0.1. Early stopping based on the correlation of predicted scores with the ground truth on the validation set was implemented to prevent overfitting. The test set correlations for each of the emotions are presented in Table 1.

Table 1.

Discrete Emotion Norm Extrapolation Model Performance (Pearson's Correlations)

Emotion	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
Correlation	0.77***	0.68***	0.73***	0.74***	0.71***	0.71***	0.81***	0.72***

* p < 0.6, ** p < 0.1, *** p < 0.001

The performance metrics of already existing valence and arousal models for Polish and English taken from Plisiecki and Sobieszek (2023) are reported in Table 2.

Table 2.

Continuous Norm Extrapolation Model Performance for Polish and English (Pearson's Correlatrions)

Language	English	Polish
Valence	0.95***	0.93***
Arousal	0.76***	0.86***

* p < 0.6, ** p < 0.1, *** p < 0.001

In the pipeline these models were used to assess the emotional value of all words that weren't stop words, punctuations, or negations, as assessed by the *spaCy* package (Ines Montani et al., 2023). To improve the compute time of the emotion prediction pipeline, using similar models to create a very big lexicon prior to inference can be an option.

Syntactic Graph Creation

The current pipeline uses the spaCy package (Ines Montani et al., 2023) to split the text into sentences, and words, followed by an analysis of syntactic dependencies. Each word is connected to the other words it relates to syntactically. For example, in the sentence "I do not feel well," spaCy identifies "feel" as the main verb, with "I" as its subject and "well" as its modifier. Additionally, the negation "not" is linked to "feel," indicating a negative sentiment in the phrase. This information can be represented in a graph form where nodes are words, and edges are syntactic dependencies. Each word is furthermore assigned to a specific part-of-speech category (e.g., PRON (pronoun) for "I" and VERB (verb) for "feel") and each dependency labeled accordingly (e.g., nsubj (nominal subject) for "I" as the subject of "feel" and neg (negation) for "not" modifying "feel").

All of punctuation marks are removed from the text, prior to the construction of the graph, apart for the ellipsis, exclamation, and question marks ('…', '! ', '? ') which were retained as they play a big role in the modulation of text meaning. While spaCy recognizes only around 20 part of speech tags, its taxonomy for dependency types is much larger. For this reason, they have been recategorized to a more manageable taxonomy of 15 separate categories with entries like "Descriptive Modifiers of Verbs", or "Negations". The full mapping is available on the paper's github repository (https://github.com/hplisiecki/Semantic-Propagation-GNN).

The resulting structure is a graph where the nodes (words) are assigned feature vectors with the emotion ratings predicted at the word level emotion prediction stage, along with a number signifying their position in the sentence (word index divided by the number of words in the sentence). Words are also

assigned parts of speech indexes signifying their parts of speech categorization. Finally, each of the edges (connections) within the graph gets assigned their dependency indexes, relating to the dependency type taxonomy.

In order to model not only single sentences but also multiple sentences texts, all words from a sentence are additionally to an additional sentence node. When a text has more than one sentence, the sentence nodes relating to each sentence get connected to each other sequentially in the order they appear in text. These sentence nodes are "empty" in the sense that they are not assigned any emotional information. Instead, their emotion node features are initialized at zero, allowing the graph to propagate the emotion from words into them at inference. Their node features also contain their sentence number indicator (sentence index divided by the number of sentences in the text). Finally, they are also assigned a unique parts of speech category (the same for every sentence), with their edges having a unique dependency category (the same for every sentence).

Semantic Propagation Graph Neural Network

The SPROP GNN rests on the idea of allowing the model to propagate semantic information, in the form of word sentiment scores throughout the syntactic graph as part of the inference. It is able to do it thanks to the custom SPROPConv layer which considers information about the parts of speech each of the two words (nodes) connected in the graph belong to, the emotional information of the receiving node as well as the type of syntactic dependency (edge) between them.

Below is a general overview of the steps that the model performs, followed by a more formal explanation of how the SPROPConv layer works, and a short description of the rest of the model's architecture. Because this paper introduces the SPROPConv layer, much attention is paid to its description. Afterwards, the training setup is described.

General Steps Performed by the SPROP GNN

1. **Process Syntactic Graph with SPropConv Layer:**

   The model processes the syntactic graph of the text using the custom SPropConv layer, enriching each word's representation with information from related words based on their grammatical structure and roles.

2. **Concatenate with POS Embeddings:**

   Each word's updated features are combined with its part-of-speech (POS) embedding, adding grammatical context to each word's representation within the graph.

3. **Apply Attention Pooling:**

   The concatenated embeddings are passed to an attention pooling layer, which identifies and weighs the most relevant words in the graph for predicting the text's emotional tone. These weighted embeddings are then aggregated using a global addition pool to create a cohesive text representation.

4. **Pass Through Fully Connected Layers:**

The pooled text representation is further processed through fully connected layers. These layers refine and adjust the representation to reach the dimensionality needed for the final prediction.

5. **Generate Final Prediction:**

   For continuous emotional metrics, the output layer uses sigmoid activation to predict values between 0 and 1 for each metric. For discrete emotion categories, a softmax activation generates probabilities across categories, identifying the most likely emotion.

Mathematical Formulation of the SPropConv Layer

The SProp layer operates through a series of steps that involve transforming node features, computing messages between nodes, and updating node representations.

1. Node Feature Transformation

Each word in the sentence is initially represented by a feature vector which consists of the emotional score of each word, alongside its index sentence divided by sentence length. These features are transformed to a hidden representation using a linear transformation:

$$h_i = W_x x_i + b_x$$

- h_i: Hidden representation of node iii.
- W_x, b_x: Learnable parameters (weights and biases).

2. Message Passing

For each edge from node j to node i (representing a syntactic dependency), the model computes a message that incorporates:

- The hidden representation of the source node h_j
- The embeddings of the POS tags for both nodes: t_i for node $\boldsymbol{i}$ and t_j for node $\boldsymbol{j}$.
- The embedding of the edge type (syntactic dependency) s_{ij}

These components are concatenated and passed through a linear transformation followed by a hyperbolic tangent activation (tanh) to compute a scaling factor s_{ij}

$$\boldsymbol{s_{ij}} = \mathbf{tanh}\left(\boldsymbol{W_s}\left[\boldsymbol{h_j}; \boldsymbol{t_i}; \boldsymbol{t_j}; \boldsymbol{e_{ij}}\right] + \boldsymbol{b_s}\right)$$

- $[\cdot;\cdot]$: Concatenation operation.
- W_s, b_s: Learnable parameters.

3. Message Computation

The message from node ***j*** to node ***i*** is calculated by scaling the hidden representation of node ***j*** with the scaling factor s_{ij}:

$$m_{ij} = \boldsymbol{s_{ij}} \cdot h_j$$

This step allows the model to modulate the influence of node ***j*** on node ***i*** based on their syntactic and semantic relationship.

4. Aggregation

For each node ***i***, the incoming messages from all its neighboring nodes are aggregated using summation:

$$a_i = \sum_{j \in \mathcal{N}(i)} m_{ij}$$

- $\mathcal{N}(i)$: Set of neighboring nodes of node iii.

5. Update

The node's hidden representation is updated by combining its original hidden state with the aggregated messages, followed by a rectified linear unit (ReLU) activation:

$$h_i' = \mathbf{ReLU}(h_i + a_i)$$

- h_i': Updated hidden representation of node iii.

The Remaining Architecture

After the syntactic graph has been processed using the SPropConv layer, the model concatenates the graph's matrix representation with the parts-of-speech embeddings for each word in the syntactic graph. This concatenated embedding is then passed to an attention pooling layer, which identifies the words in the graph that contain the most relevant information for predicting the text's emotional tone, assigns them weights, and aggregates these embeddings using a global addition pool.

This representation is then passed through fully connected layers that gradually bring them to the dimensionality required by the prediction. In the case of continuous emotional metrics, this means one output dimension per predicted metric, with a sigmoid activation applied to scale the output between 0 and 1. For discrete emotion prediction, the final layer instead uses a softmax activation, outputting probabilities across predefined emotion categories.

Specific Architectures and Training Setup

The three SProp GNN models trained on the three datasets GoEmotions, EmoBank, and the Polish Political Dataset share a similar architecture. Each model contains a single SProp layer with 512 hidden dimensions, alongside embedding layers for both parts of speech (node types) and dependency relationships (edge types), with dimensions matching those of the SProp layer. This is followed by a

global attention mechanism, which applies a gated attention layer configured with two linear transformations (1024 to 256, and 256 to 1) and a ReLU activation in between. The attention weights are computed by applying softmax across nodes within each graph, and graph-level features are subsequently aggregated using a global addition pool.

The differences between the models lie in the final sequence of linear layers. In the case of discrete predictions, these layers have the form of three linear transformations (1024 to 1024, 1024 to 512, and 512, to the number of discrete emotions), separated by dropout and relu activations. Alternatively, in the case of the two continuous metric prediction models there are only two linear layers (1024 to 100, and 100 to 1), also separated by a dropout and a relu activation. These differences stem from free experimentation with different amount of final linear layers. A systematic exploration of alternative architectural setups is beyond the scope of this study.

The hyperparameters for the three SProp GNN models were chosen using a Bayesian hyperparameter sweep on the wandb platform (*Wandb/Wandb*, 2017/2024). The hyperparameter options for the three models were the same: dropout - 0, 0.2, 0.4, 0.6; learning rate - 5e-3, 5e-4, 5e-5, and weight decay - 5e-3, 5e-4, 5e-5. All models were trained using the AdamW optimizer with the epsilon equal to 1e-6 and betas equal to 0.9, and 0.999. The discrete model used the cross-entropy loss, while the continuous metric prediction models used the mean squared error loss. The final models were trained using the best performing parameters from the sweeps.

Comparative Experiments

This section will outline the data wrangling performed on the datasets that were used to compare the SProp GNN with other methods, along with the explanation of how each of the alternative methods were implemented.

Data Wrangling

Each of the datasets was processed for the task of using them to compare alternative approaches to emotion prediction. Considerable attention was paid to the description of the Polish political dataset as it is a far less known dataset when compared to the other two.

The Goemotions Dataset

The goemotions dataset was developed by a team at Google (Demszky et al., 2020). It consists of 57565 unique texts and 210622 annotations. Each comment received annotations from three English-speaking raters from India, with additional raters assigned when agreement was low. The most voted for emotion per each text was computed and those texts for which two emotions were assigned the same number of votes were dropped. This resulted in a dataset of 47136 unique texts. From these texts, those that were assigned one of the following emotions: anger, disgust, fear, joy, surprise; were retained leaving 4819 unique texts. The choice of emotions was dictated by the availability of emotion norms in the NRC Emotion Intensity Lexicon (Mohammad, 2017). This dataset was then split into the training, evaluation, and test sets in the proportion of 8:1:1.

The Emobank Dataset

The EmoBank dataset, created by Buechel and Hahn (2022), consists of 10,062 English sentences from sources like news, blogs, fiction, and letters, annotated along three emotional dimensions: Valence,

Arousal, and Dominance (VAD). Each sentence was rated by multiple annotators from the crowdsourcing platform CrowdFlower for both *writer* and *reader* perspectives, giving insights into both expressed and perceived emotions. Each sentence was annotated by 5 annotators. In accordance with the recommendations of the researchers, the dataset with the weighted average of the reader and writer perspective labels provided at their online repository was used for training (*JULIELab/EmoBank*, 2017/2024). The ratings for valence and arousal were normalized to a 0 to 1 range by subtracting the lowest score and dividing by the number of Likert scoring options prior to splitting into the training, evaluation, and test sets in the proportion of 8:1:1.

The Polish Political Dataset

The Polish Political dataset (Plisiecki et al., 2024) was created by sampling text data from social media profiles of Polish journalists, politicians, and non-governmental organizations (NGOs) across YouTube, Twitter, and Facebook. Posts from 2019 onward were collected for 69 profiles. A total of 1,246,337 text snippets were gathered, with breakdowns of 789,490 tweets, 42,252 YouTube comments, and 414,595 Facebook posts. To handle the varying text lengths, Facebook posts were split into sentences, and only texts under 280 characters were retained. Social media artifacts, such as dates and extraneous links, were removed, and non-Polish texts were filtered using language detection software. To prevent overfitting, online links and usernames were standardized as "*link*" and "*user*."

To create a dataset with richer emotional content, neutral texts were filtered out, leaving only those with higher levels of emotional valence, arousal, and dominance. This selection process used a lexicon-based approach, where each text was assessed for emotional intensity across these dimensions, resulting in 8,000 emotionally charged texts. An additional 2,000 neutral texts were included to balance the dataset, preserving original platform proportions. The final 10,000-text dataset, comprising 496 YouTube comments, 6,105 tweets, and 3,399 Facebook texts, was then annotated by 20 psychology students well-versed in Polish political discourse. Each text was rated by five randomly assigned annotators on six emotions (happiness, sadness, disgust, fear, anger, and pride) and two emotional dimensions (valence and arousal), using a 5-point Likert scale. Before formal annotation, annotators received an introduction to valence and arousal, and comprehensive guidelines were provided to ensure consistency. For clarity, the annotators received the following English instruction for evaluating valence and arousal:

"Go back to the text you just read. Now think about the sign of emotion (positive / negative) and the arousal you read in a given text (no arousal / extreme arousal). Rate the text on these emotional dimensions."

This instruction was designed to provide a standardized understanding of emotional dimensions, ensuring alignment in annotators' assessments across the dataset.

For the purposes of the current experiment, all of the emoticons and symbols were filtered out and the dataset was split into the training, evaluation, and test sets in 8:1:1 proportion.

Comparative Approaches

This section outlines the details of how each of the alternative approaches was set up and trained for performance comparison.

The Lexicon Approach

For the lexicon analysis of the EmoBank, and the Polish Political dataset I utilize the norm extrapolation transformer-based models for Polish and English described in the "Word Level Emotion Prediction" section above. Each test set text was first split into words using the *spacy* package (Ines Montani et al., 2023). Each word that wasn't a stop word was then fed into the norm extrapolation model, and the emotional prediction was averaged to get the text level emotion score.

The Vader Approach

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a rule-based model designed for sentiment analysis, particularly effective in capturing sentiment from social media and informal text. I have used the 3.3.2 version of the Vader package to get the valence/positivity scores for the Emobank test set.

The EmoAtlas Approach

The EmoAtlas utilizes an extensive lexicon-based network to profile emotions by mapping syntactic and semantic relationships in text, effectively capturing nuanced emotional cues without extensive model training. The EmoAtlas performance results for the goemotion dataset were taken directly from the original paper (Semeraro et al., 2023).

The Transformer Approach

For the Goemotion and EmoBank datasets the *roberta-base* transformer model developed by Facebook was finetuned on the two English datasets (Liu et al., 2019). A fully connected layer, with the dimensions equal to 768 was added on the top of the base model with dropout and a layer norm, with either a regression head for the sake of predicting valence and arousal, or a classification head for predicting discrete emotions. A Bayesian hyperparameter sweep was performed using the wandb platform (*Wandb/Wandb*, 2017/2024) for both models with 20 runs and the following hyperparameter options: dropout – 0.0, 0.2, 0.4, 0.6; learning rate – 5e-4, 5e-5, 5e-6; weight decay – 0.0, 0.2, 0.4, 0.6; and warmup steps – 300, 600, 900. Both models use the AdamW optimizer for training with the epsilon equal to 1e-6 and betas equal to 0.9, and 0.999, alongside the linear learning rate scheduler with warmup. In the case of discrete prediction, cross-entropy loss was used, while in the case of continuous emotion metric prediction mean squared error loss was chosen. Finally, the final models have been trained using the best performing hyperparameters from the sweep. The performance of each of the models is reported in the results section of the main manuscript. The training code for these models can be found in the following Google Collab:
https://colab.research.google.com/drive/1pA3oBbHg0pza1yF5kyuddK36-RGKtHxo?usp=sharing