Dr hab. Małgorzata Marciniak, prof. IPI PAN
Instytut Podstaw Informatyki PAN
ul. Jana Kazimierza 5
01-248 Warszawa

Warszawa, 27.11.2025

Review of the doctoral dissertation by Hubert Plisiecki, MA, entitled "Words, Vectors, and Feelings: Advancing Psychological Emotion Research Through Natural Language Processing"

Hubert Plisiecki's doctoral dissertation consists of a collection of the following four articles preceded by an introduction:

- "Emotion Topology: Extracting Fundamental Components of Emotions from Text Using Word Embeddings", by Hubert Plisiecki and Adam Sobieszek, published in *Frontiers in Psychology*;
- "Extrapolation of Affective Norms Using Transformer-Based Neural Networks and Its Application to Experimental Stimuli Selection" by Hubert Plisiecki and Adam Sobieszek published in *Behavior Research Methods*;
- "High risk of political bias in black box emotion inference models" by Hubert Plisiecki, Paweł Lenartowicz, Maria Flakus, and Artur Pokropek, published in *Scientific Reports*;
- "Eradicating Social Biases in Sentiment Analysis using Semantic Blinding and Semantic Propagation Graph Neural Networks" by Hubert Plisiecki published on arXiv.

Three of the articles are published in peer-reviewed journals, and the fourth is a preprint. This means that they meet the requirements for the number of articles that make up a doctoral thesis. The dissertation, however, does not, in my opinion, meet the requirements for thematically related scientific articles. The usefulness of machine learning methods and language models is not a hypothesis that needs to be proven. For fifteen years, sentiment analysis has a track at big language conferences like COLING and ACL. Conference proceedings and journals include numerous articles on the application of machine learning and language models in psychology. The query containing the terms *psychology*, *language models* and *machine learning* in Google Scholar returns almost 60,000 papers. The PhD candidate formulates research questions for each article separately but does not show how they are related.

Furthermore, I am unable to determine which concepts and works were developed by the PhD candidate. Although the first and third articles contain notes specifying the work carried out by the authors, they are imprecise and indicate the same contribution being declared by at least one author other than the PhD candidate. The second article does not contain any information about the authors' contribution. The declaration attached to the dissertation (only signed by the PhD candidate) doesn't specify participation in any way. It is a list of activities with a YES answer marked on all items in the form except for the last one, which refers to the acquisition of funds. The place for describing the candidate's contribution to each article is in the introduction to the dissertation, but the candidate did not take advantage of this.

Two articles in the dissertation are printed in a font that is too small, requiring them to be enlarged on a screen. This doesn't impact the evaluation of the work, but it makes reviewing it harder than it needs to be.

I will now move on to discuss the problems that I identified in subsequent parts of the PhD thesis. First, I would like to stress that I am reading this dissertation as a specialist on natural language processing (NLP) and not psychology science, so it is those aspects I focus on. Below, I highlight issues that, in my opinion, are controversial or that have not been adequately addressed in the articles.

Introduction

The introduction provides general information about natural language processing using language models, i.e., it is a historical overview supplemented by one-page summaries of articles. Unfortunately, it contains statements concerning NLP that are either imprecise or unfortunate. Below are some examples:

The author writes, "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." Text analysis is one of many subfields of NLP; these are not synonyms.

What does the author mean by "traditional linguistic quantification" in the sentence "it is essential to understand the fundamental tools that have enabled the shift from traditional linguistic quantification to modern computational approaches"?

The subsection of the introduction "An Overview of the Main Tools in Natural Language Processing" doesn't mention a single NLP tool. Instead, it mentions the successive development stages of the vector representation of language.

"an encoder processes input text, converting it into a compact numerical representation (embedding)" I would not call embeddings a compact representation.

The thesis contains very long and difficult-to-understand sentences such as the following: "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." I don't understand what "performance problem" the author might encounter when using a dictionary such as LIWC. What does the author mean by "NLP-based models"? How does the author propose to evaluate the validity and reliability of the algorithms that produced language models?

"APIs" should appear without an apostrophe in the phrase "with the help of available API's".

The following claim, "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." is clearly untrue.

"Emotion Topology: Extracting Fundamental Components of Emotions from Text Using Word Embeddings"

In the paper, the authors use the GoEmotions dataset, which consists of 58,000 Reddit comments annotated with 28 emotions. The authors create texts consisting of comments annotated with a given emotion. A text vector is counted for each text with the doc2vec model. The authors state that such text vectors correspond to emotion and call them "Emotion vectors", analyze their features, and cluster them. The PhD candidate states in the Introduction "The study showcases a new method of extracting vectors that relate to psychological phenomena".

In section 2.4.2, the authors claim that they trained 3x3x9=81 models (a lot) and selected the best one, but we aren't told which model turned out to be the best.

The authors write, "we recoded the original GoEmotions dataset from 28 emotions into positive and negative labels." Originally, the set of 28 emotions was divided into three groups: positive, negative, and neutral. The article does not explain why the authors abandon the neutral emotions group and classify "confusion" and " surprise" as negative emotions, while classifying "curiosity" and "realization" as positive ones.

The dataset is not sufficiently described. We know that it consists of 58,000 Reddit comments and that the final dataset consists of 155,663 pairs (text-label). So, it is almost three times more. It means that on average, each comment has three labels. Each comment is therefore a component of three texts related to three various emotions. If two texts contain the same sentences/phrases, they are necessarily more similar than those that do not overlap. In other words, a direction in embedding space of say, two emotions, may simply be the result of label co-occurrence (the same texts sharing embedding labels), and similar visualizations to those the authors present could only be obtained using label co-occurrence analytics. The experiment is therefore

incorrectly constructed. Correct data preparation in machine learning methods is crucial. If a comment has multiple annotations, we should know whether these emotions refer to different fragments or to the entire comment. Perhaps those with multiple annotations should be excluded from the experiment? The length of texts being compared is also important.

If the authors claim that the obtained vectors represent emotions, why didn't they, for example, attempt to check the similarity of the vectors representing words to those obtained from texts annotated with emotion? This would demonstrate the correctness and usefulness of such an approach.

Furthermore, why don't they compare their results to those described in the paper introducing the dataset, "GoEmotions: Dataset of Fine-grained Emotions" Demszky et al. 2020?

In my opinion, the article does not bring anything new and does not have a clear thesis or evaluation method.

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

The authors work on affective norms in six languages. The article does not explain important issues concerning data and the method of processing, so if I were a reviewer, I would ask for corrections.

The authors process words (lists of words but no texts), so what benefits do they expect from using "highly contextual representation"? Transformer (Bert-like) models are useful for processing sentences where context modifies the vector of the word to clarify its meaning.

It is unclear what data the authors are working with (see section "Linguistic materials and data curation", page 48). The ANEW dataset (test set) contains 1034 words, while the authors have 1030 words. Then it drops to 983 as they subtract the test set from the training data (Warriner et al. 2013). Why reduce the size of the test data? For explanations, the authors refer the reader to Tables 1 and 2, but these are incorrect references. On page 48, they write that ANEW is a test set, while on page 50 they write, "high-quality validation set—the ANEW corpus". So, is it the test or validation set?

The comparison of the results of previous and current work (Figure 2) is questionable. The authors claim "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." The ANEW corpus is annotated with valence, arousal, and dominance. The results in Table 2 are also given to the concreteness and age of acquisition, which are not annotated in ANEW. In Vankrunkelsven et al., 2018, the results for the ANEW set are also given, but in the table, they are for cross-validation. The results in

Vankrunkelsven et al., 2015 concern Dutch, but the paper doesn't state this. The results for Dutch reported by the authors in Table 3 and those from Vankrunkelsven et al., 2015, are sometimes a bit worse and sometimes a bit better—a comment explaining why this is so would be useful, as it is inconsistent with the thesis of the article that the transformer-based method is substantially better.

The authors write "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." XML Roberta is a multilingual model trained on 100 languages and released in 2019.

The authors write "Each of the models was trained for 1000 epochs with early stopping" How many epochs actually was enough, because 1000 is definitely too many for such data.

As the authors write " First we gathered a list of all words that appeared at least five times in Polish language corpora" then these Polish corpora should be mentioned.

When describing the second experiment, it is worth referring to the results presented in the article: "ANEW+: Automatic Expansion and Validation of Affective Norms of Words Lexicons in Multiple Languages" Samira Shaikh, Kit Cho, Tomek Strzalkowski, Laurie Feldman, John Lien, Ting Liu, and George Aaron Broadwell, LREC 2016.

In addition, the article contains numerous typos and long sentences in which the authors' intentions are difficult to follow. Since it has already been published and cannot be corrected, I will not list them all. Some examples:

- the last sentence in the first column on page 55 remained after changes;
- LSTMS and CCNS should be LSTMs and CCNs;
- The results of Bestgan and Vinche (2012) should include "concreteness" (Table 2)

"High risk of political bias in black box emotion inference models"

The authors examined political bias in language models tuned using manually annotated data. In conclusion, they "propose using lexicon-based systems as an ideologically neutral alternative". However, it should be mentioned that the use of dictionary-based methods carries the risk of problems with interpretation, e.g., irony and metaphors, and produces poorer results.

This article perfectly illustrates the importance of proper preparation of training data in machine learning. The instructions for annotators encouraged them to subjectively represent feelings. This includes political beliefs. It is therefore not surprising that a model trained on such annotations reproduces these beliefs. Additionally, the annotators were students, 80% of whom were women (representing only a fraction of the political views of society). It seems that careful preparation of training data may substantially reduce this phenomenon.

The article lacks detailed information on the data used to conduct the experiments. On page 63, it mentions 24 politicians, including two women, which seems to be a small number. We also do not know their distribution among the five political groups. Often, the reader is referred to an appendix that is not included in the doctoral thesis.

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

I recommend the PhD candidate consult this paper with NLP specialists. Graph B in Figure 1 does not represent syntactic connections in the sentence "I am not happy". The problem of negation in syntactic analysis (parsing) has a broad literature. It is possible for the correct interpretation of the emotion in the phrase "not happy" to be the opposite to "happy ". Antonyms can be extracted, e.g., from the WordNet.

I understand that the author relies on a dependency parser. It would be helpful to include an example graph for the sentence being analyzed and describe the principles of sentiment propagation. I don't know how the graphs in Figures 2 and 3 were created and how they relate to dependency graphs.

The title of the article doesn't match its content. What kind of social bias is being removed from the GoEmotion collection?

Conclusion

Considering all the comments made in my review, I conclude that the thesis doesn't meet the formal requirements in § 3 item 3.5. This is due to the lack of thematic connection between the series of articles and the lack of precise indication of the PhD candidate's contribution to the co-authored articles. The review also contains critical comments on both the content and form of the articles included in the thesis. Since these are published articles, they cannot be improved. This removes the possibility of improving the PhD thesis. The doctoral dissertation doesn't meet the conditions specified in art. 187 of the Act of 20 July 2018 – Law on Higher Education and Science (Journal of Laws, item 478, 619, 1630)

Małgorzata Marciniak