Automated Content Moderation Using Transparent Solutions and Linguistic Expertise

Veronika Solopova
Freie Universität Berlin, Germany
veronika.solopova@fu-berlin.de

Abstract

Since the dawn of Transformer-based models, the trade-off between transparency and accuracy has been a topical issue in the NLP community. Working towards ethical and transparent automated content moderation (ACM), my goal is to find where it is still relevant to implement linguistic expertise. I show that transparent statistical models based on linguistic knowledge can still be competitive, while linguistic features have many other useful applications.

1 Introduction

My main directions of work include error analysis, explainability techniques, and training transparent models, all using linguistic features. The main challenges I face that motivate my work are:

- The Trade-off between transparency of automated decisions and accuracy in ACM.
- Ethical AI solutions to content moderation in social media and education.
- The trade-off between personal freedoms on the Internet, protection of repressed minority groups and international security.

I investigate such social media concepts as hate speech, fake news, and herd behaviour. In education, we consider automated essay analysis.

2 Methods

Fake news detection. In terms of fake news, I apply my techniques to propaganda detection. I compare two methods of multilingual automated pro-Kremlin state-sponsored propaganda identification, based on Transformers and SVM with linguistic features and manipulative terms' glossary, trained using news articles and Telegram news channels in Ukrainian, Russian, Romanian, French and English. My multilingual BERT model [Devlin *et al.*, 2018] achieved a maximum of 93% F1, while SVM model achieved 89% F1. The addition of a glossary to the morpho-syntactic stylistic features of the SVM model did not influence the results. The SVM performs slightly better on a new genre when I trained the models on news articles only and applied them to telegram posts. Due to its inherent explainability, I looked at the coefficients attributed to each feature in prediction, and I identified the most important features towards different stances. The BERT-based model has a clear tendency towards false positives and performs slightly worse on unseen genres. The overall propaganda style captured by linguistic features is reliable, as I observed that the model's performance does not drastically change for any of the languages in focus. Scalability is, however, a major drawback of feature-based models, while BERT models have important token length limitations.

As the content and form evolve and change with time, this constitutes a challenge to content moderation tools. In my follow-up study, I evaluate my previous models trained on the data from the start of 2022, on the new 2023 subset. Both my models performed very well on the new data, outperforming the results received on the 2022 data. My approach for the SVM error analysis includes testing if distributions of the features have correlations among the output groups: false positives, false negatives, true positives, and true negatives. I witnessed that some distributions of significantly important features from my previous study are now similarly distributed between FPs and the TPs, which causes some errors. In the case of the BERT model, I used a simplified attribution method and I saw that this model is prone to attributing the class according to the news source name mentioned, which can lead to the model predicting as propaganda something debunking it. I also observed that morphological information may be used more than syntactical one for predictions in BERT. Finally, I aligned the extracted predictive words for BERT with my linguistic features set and identified that many of my categories are used actively by BERT.

Hate Speech detection. In our research [Scheffler *et al.*, 2021] I focused on the evaluation of transformer-based and list-based moderation resources. I used a Telegram channel populated by followers of former US President Donald Trump during his presidency. I did not only focus on direct insults and hateful speech, which was the focus of existing works but rather less explicit dividing and offensive language. First, with my co-authors, I created a taxonomy of harmful language and annotated a large portion of our data with it. Second, I compare our manual fine-grained annotations of harmful speech to several automatic methods. I found that while both methods had very low performance on non-explicit language, not once did they make errors in the same sentence.

Herd behaviour on social media. I also investigated shitstorms on social media, which is 'an unforeseen, short-lived wave of outrage in social media [Gaderer, 2018]. With the help of adapted linguistic features set, I examine the propagation of the outrage wave across two media where the shitstorm in focus occurred and test the applicability of machine learning methods to analyze its temporal progression [Scheffler *et al.*, 2022]. Our hypothesis is that supporters remain equally active over time, while the dynamic "ripple" effect of the shitstorm is based on cross-platform participant recruitment. First, using the multilingual BERT model, I classified messages according to the time period in which they appear in the shitstorm. I distinguished three classes: intense beginning, middle to last peak, and end. Second, I classified the contributor's position into three classes: Supporter, Opponent, and Neutral to the target of the shitstorm. I obtained a lower result on this task, largely due to the fact that I grouped a lot of heterogeneous subclasses in the "for" and "against" classes. Reaching 65-67% performance, I reckon that with more data these tasks have a lot of promise.

Student essay automated moderation. I also consider automated feedback on student essays as a case of ACM. I built a controlled system based on natural language understanding (sentiment analysis, emotions detection, reflective component and reflective depth identification, and essay style analysis based on linguistic features). It produces personalised prompts and essay evaluation.

3 Conclusion

3.1 Limitations

My work on social media content moderation seeks to contribute to the efforts to protect human moderators from the constant psychological trauma [Steiger *et al.*, 2021]. I see my work as a first step towards a browser extension flagging harmful content to raise individual awareness. However, the propaganda classifier can be used to block pro-Western news as well, ensuring the impenetrability of echo chambers, and amplifying the effects of propaganda. It can also help Kremlin modify its propaganda so that it does not contain features I identified, making it more difficult to detect. In the context of social media ACM, false positives would mean flagging/filtering a post or banning a person, limiting the freedom of speech. False negatives might lead to posts with propaganda reaching more targets. Hence, the high performance of these tools is extremely important, and automated responses should not be used to ban a user from the platform or restrict the monetisation of content.

In the same way, automated essay analysis at this stage may only be a supporting tool with a teacher-in-the-loop.

3.2 Future work

I would like to look into correlations between hate speech, propaganda and fake news occurrences in posts on social media and see if it is possible to build a model which would predict one phenomenon based on the presence of others. The ultimate goal would be building a formal ethical and legal governance system which could serve as a unified approach for content moderation on social media. The approach would include formal reasoning on the ethical and legal levels over the outputs of several models extracting toxic and harmful behaviour indicators (fake news, hate speech, offensive language and its level, propaganda, its origin and goal) and propose an appropriate moderation method. In the case of student essay moderation, my main goal is to create a controlled chatbot, which would not only be able to produce structural and motivational feedback, as of now but also a professional one, using generative models with restrictions.

References

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[Gaderer, 2018] Rupert Gaderer. Shitstorm. das eigentliche übel der vernetzten gesellschaft. *Felix Meiner*, 2018.

[Scheffler *et al.*, 2021] Tatjana Scheffler, Veronika Solopova, and Mihaela Popa-Wyatt. The telegram chronicles of online harm. *Journal of Open Humanities Data*, Jul 2021.

[Scheffler *et al.*, 2022] Tatjana Scheffler, Veronika Solopova, and Mihaela Popa-Wyatt. Verbreitungsmechanismen schädigender sprache im netz: Anatomie zweier shitstorms. *Hassrede, Shitstorm und Darstellungspolitiken virtueller Affekt Workshop*, 2022.

[Steiger *et al.*, 2021] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.