Deep Learning Analysis of Tweets Regarding Covid19 Vaccination in the Serbian Language

Nikola Prodanović*, Adela Ljajić*, Darija Medvecki*, Jelena Mitrović**, Dubravko Ćulibrk*
* The Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, Serbia
** Faculty of Computer Science and Mathematics, University of Passau, Germany
nikola.prodanovic@ivi.ac.rs
adela.ljajic@ivi.ac.rs
darija.medvecki@ivi.ac.rs
jelena.mitrovc@ivi.ac.rs
dubravko.culibrk@ivi.ac.rs

***Abstract*— In this paper, we present an efficient classifier that is able to perform automatic filtering and detection of tweets with clear negative sentiment towards COVID-19 vaccination process. We used a transformer-based architecture in order to build the classifier. A pre-trained transformer encoder that is trained in ELECTRA fashion, BERTic, was selected and fine-tuned on a dataset we collected and manually annotated. Such an automatic filtering and detection algorithm is of utmost importance in order to explore the reasons behind the negative sentiment of Twitter users towards a particular topic and develop a communication strategy to educate them and provide them with accurate information regarding their specific beliefs that have been identified.**

I. INTRODUCTION

The Covid19 pandemic has significantly disrupted the daily lives of individuals and the way in which organizations operate around the world. One of the most effective strategies to tackle the Covid19 pandemic is to achieve collective immunity through mass vaccination. However, people have shown significant resistance and hesitation to the global immunization process. Therefore, the study of the public's attitude toward the vaccination process is of utmost importance. In particular, it is useful to identify the prevailing beliefs and attitudes that may lead to negative sentiment toward vaccination. These are usually related to the efficiency, safety, necessity, etc. of vaccination.

Our strategy can be summarized as follows: In order to study the attitudes and beliefs behind negative sentiment, our goal is to collect as large a number of tweets with negative sentiment as possible. Deciding which tweet is negative is a very difficult process. First, we obtain a large number of tweets with matching keywords via the Tweeter API. However, there is only a very small number of tweets that actually have an attitude towards vaccination, and only a subset of these tweets has a negative sentiment. Therefore, we have decided to develop a deep learning classifier that is able to detect relevant tweets with a sufficiently clear negative attitude towards the vaccination process. The classifier consists of two sequential parts. The first part filters tweets based on their relevance to the topic and the second part filters tweets based on their sentiment.

This kind of public opinion analysis should be done for each geographic region using the tools developed for the languages prevalent in the region. In this paper, we utilize the recently developed BERT language model for Bosnian, Croatian, Montenegrin, and Serbian languages in order to build both of the required classifiers.

II. DATA

For the purpose of building the classifiers, we collected about 11,000 tweets based on the keywords important for the topic of vaccination. The timestamps of the data range from January 1, 2021 to November 23, 2021. This collection includes tweets in Serbian that have not yet been normalized, as the transformer language models require raw text with all morphosyntactic information.

The dataset was manually labeled by human annotators with three classes for sentiment toward vaccination, one class for irrelevant tweets, and one class for irrelevant tweets containing the link. The annotation subjects were text content, hashtags, and emoticons. After the fully individual and separate annotation by the evaluators, a discussion took place on some complex and problematic tweets in order to improve the quality and consistency of the assigned labels. For example, the discussion revealed that tweets closely related to political beliefs were the most challenging. After certain conclusions were drawn and doubts were resolved, the annotators revised their annotations. At the end of the process, we acquired 5791 tweets with a clear message about the three-class sentiment regarding vaccination. Since the first classifier will be trained to detect these relevant tweets, we can safely conclude that the class split for this classifier is around 50%. Among the 5791 relevant tweets, there is also a balanced representation of the three sentiment classes, ranging from 20% to 40%.

III. METHODOLOGY

For the purpose of building both classifiers, a pre-trained language model is transfer-learned and further fine-tuned using annotated data. The state-of-the-art language models are usually based on a self-attention mechanism that efficiently captures the morphosyntax of the language through the pre-training process [1]. Historically, the sequence-to-sequence transduction model was the original model with the attention mechanism [1], but very soon after, the first encoder-only architecture was published under the acronym BERT. [2]

The pretraining strategy for such an encoder is usually defined as Masked Language Modeling, which resembles the autoencoders, and as next sentence prediction task [2]. The most recent proposal for a pre-training strategy was the so-called ELECTRA approach, where the BERT model is trained as a discriminator rather than a generator. This method was utilized to train the first BERT based model for South Slavic languages, known as BERTic [3].

Twitter messages are very specific in their length, informal writing, figurative language, and often contain slang. This presents a significant challenge for natural language understanding models. We aim to test BERTic on this challenging textual content. We expect BERT to be robust to informal language since informal language is usually robustly syntactically correct and lacks morphological correctness. In Ref. [4], the authors stated that BERT outperforms similar models on this type of analysis.

We used annotated data to fine-tune and test BERTic on a downstream task of short text classification. This approach differs from pre-training BERT from scratch on a much larger corpus of tweets [5]. We use this approach for both of our classifiers.

In addition, we intend to use fine-tuned models to statistically analyze significantly larger corpora of Twitter messages that are implausible to analyze by human annotators. However, this is not the main topic of this paper.

In Fig.1, we present the pipeline that we intend to develop for broader and complete analysis of the attitudes towards vaccination. The most difficult part of such pipeline is the process of automatic filtering and detection of tweets with negative sentiment toward COVID -19 immunization. That part of the pipeline is the subject of this paper. The last part of the pipeline, the topic modeling, is outside the scope of this paper, but it is one of the important possible steps and techniques for further processing of this type of identified tweets, which can be used to extract more specific information that can be of great use for a deeper understanding of the negative attitudes towards this topic. In the Conclusion and Future Work section of this paper, we have explained the techniques that we propose and plan to use on this matter in the future.

IV. Results and Discussion

We trained our algorithm on only one iteration of the annotation process because we also wanted to analyze possible human annotation errors. We found that human annotation errors are very common, which was to be expected given the complexity of the semantics of the tweets. However, the algorithms proved to be very resilient to this syndrome and statistically learned very well from the majority of correctly labeled examples.

A. *Relevant Tweets*

The first classifier detects whether the tweet is relevant enough to be considered as an opinion about vaccines. Usually, non-relevant tweets are strongly related to epidemics and politics, but there is no bare attitude towards vaccination. The algorithm performs very well in this part. It was validated on 10 percent of the total number of tweets, which is approximately 1000 tweets in this case. The accuracy is 83 percent.

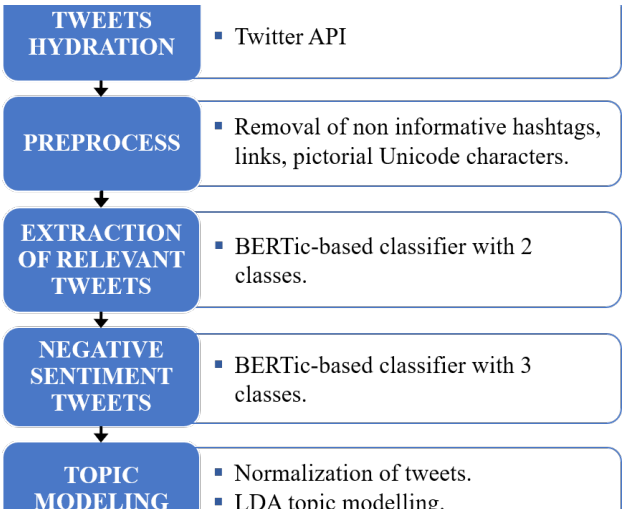

Figure 1. The pipeline that we intend to develop for broader and complete analysis of the attitudes towards vaccination. This work includes first four steps.

The largest disagreement between annotator and the algorithm is in the cases where the algorithm incorrectly predicts that the tweet is relevant, while the annotator marks the tweet as irrelevant. In this regard, the recall for the negative class is 0.62 and that is the weakest point in this part of the pipeline. However, we are not too concerned about these falsely relevant tweets, as we expect that majority of them to be classified as neutral class. Therefore, the overall effect on filtering negative tweets is very small. Other precision and recall values are significantly higher. As mentioned earlier, many of the mismatched tweets are actually an error by the annotator. In other words, annotator incorrectly labeled many of the tweets as irrelevant, likely guided by his or her own emotions. A similar effect occurs with the sentiment classifier. We need to conduct active learning process in the future in order to tackle this effect.

B. *Sentiment*

The second classifier takes as input only relevant tweets and outputs their sentiment towards vaccination. We obtained 5791 relevant tweets and split them in 90/10 percent train-test ratio. This classifier has a more difficult task than the previous one, so we provide some additional training details below.

The number of epochs and batch size were chosen to be optimal for a fixed test set, which may result in a slight but acceptable bias. Furthermore, this is justified by the recommended values of these hyperparameters given in the original paper describing the model BERT [2], namely the number of epochs of 4 and the training batch size of 16 tweets.

After the first annotation iteration, the accuracy of the model in the test phase was about 68%. Most of the confused examples fall between neutral and the other two

classes. Precision and recall are approximately equally dispersed among the classes. Upon closer inspection, it was confirmed that this type of annotation task is really difficult for people to perform and to decide objectively and with utmost certainty which labels to assign.

As mentioned earlier, the algorithm often outperforms its supervisor, by about 12%. This leads to the conclusion that annotation was an emotionally and mentally difficult process in which the annotator makes typical human mistakes.

BERTic, on the other hand, learns statistically from the majority of correctly labeled examples.

Nevertheless, there is overfitting present in the fine-tuning process, indicated by extremely high training accuracy. This indicates that more data would improve the algorithm.

The supervisor outperforms the algorithm in about 8% of the examples. These are the examples that usually contain complex emotional content and figurative language. For many of these examples, a broader knowledge is required.

Clearly mixed cases account for 12%. These examples are mostly long tweets with multiple contradictory statements. Any disagreement is therefore justified. Further inclusion of intermediate values would likely lead to improvement on this basis.

If we consider mixed examples and examples where the algorithm performs better than the annotator as justified, we arrive at an overall accuracy of about 90%.

All this suggests that the algorithm would improve if we were to apply some revised annotations through the so-called active learning approach [6]. The already explained overfitting in combination with the annotators' mistakes may lead to a slight bias and degradation of the overall performance of the classifier. However, we expect this to be a weak effect since the vast majority of examples are correctly labeled and the algorithm learns robustly and statistically from the majority of correctly labeled examples.

V. Conclusion and Future Work

The contribution of this paper is twofold. First, we have developed a very efficient two-stage classifier for automatic detection of tweets with negative sentiment on the topic of Covid19 immunization. Considering the possible errors that human annotators might make in labeling the tweets - which is an emotionally and mentally difficult process given the nature of tweets, especially on this topic where there are situations in which a single tweet can reflect multiple opposing opinions - we expect that such a two-stage classifier should achieve a deployment accuracy of about 80%. We plan to use these classifiers in selecting a large number of tweets with negative sentiment and perform further analysis using standard topic modeling techniques [7]. Second, we have successfully tested the first BERT language model developed for Bosnian, Croatian, Montenegrin, and Serbian languages on a downstream task of classifying short texts in tweets written in Serbian language.

These classifiers will be further used as part of our larger pipeline in order to analyze the beliefs behind negative sentiment of Twitter users. These two classifiers will be combined to automatically extract a large number of sufficiently relevant tweets with clear negative sentiment of the message related to the immunization process. We will then use this large number of tweets to perform topic modeling [7]. Obtained topics will be then

matched with topics that we have already identified through the annotation process. We expect to obtain an accurate quantification of various beliefs that are present in the part of the Twitter community that tweets in the Serbian language.

Acknowledgment

This work was partially supported by the Government of Republic of Serbia and partially by organizations USAID and the United Nation's Development Programme.

References

[1] A. Vaswani *et al.*, “Attention Is All You Need.” arXiv, Dec. 05, 2017. Accessed: May 26, 2022. [Online]. Available: http://arxiv.org/abs/1706.03762

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *ArXiv181004805 Cs*, May 2019, Accessed: Jan. 20, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805

[3] N. Ljubešić and D. Lauc, “BERTi\’c -- The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian,” *ArXiv210409243 Cs*, Apr. 2021, Accessed: Sep. 16, 2021. [Online]. Available: http://arxiv.org/abs/2104.09243

[4] Q. G. To *et al.*, “Applying Machine Learning to Identify Anti-Vaccination Tweets during the COVID-19 Pandemic,” *Int. J. Environ. Res. Public. Health*, vol. 18, no. 8, Art. no. 8, Jan. 2021, doi: 10.3390/ijerph18084069.

[5] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, “BERTweet: A pre-trained language model for English Tweets,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, Oct. 2020, pp. 9–14. doi: 10.18653/v1/2020.emnlp-demos.2.

[6] V. Ilić and J. Tadić, “Active learning using a self-correcting neural network (ALSCN),” *Appl. Intell.*, vol. 52, no. 2, pp. 1956–1968, Jan. 2022, doi: 10.1007/s10489-021-02515-y.

[7] Topic Modeling Technique on Covid19 Tweets in Serbian, Adela Ljajić, Nikola Prodanović, Darija Medvecki, Bojana Bašaragin, Jelena Mitrović, unpublished