Stance Classification of Internet Comments on COVID-19 Vaccination

Axel Blom
M.Sc. Data Science & AI
DAT341, Group 36
axelblo@chalmers.se

Andreas Helgesson
M.Sc. Data Science & AI
DAT341, Group 36
andhelge@chalmers.se

Johanna Norell
M.Sc. Data Science & AI
DAT341, Group 36
jnorell@chalmers.se

Abstract

By training a custom natural language processing model based on BERT, we achieve 94.3% accuracy in predicting whether or not comments scraped from internet forums convey a positive or negative stance on COVID-19 vaccination.

1 Introduction

Social media platforms serve as expansive forums for public discussions, enabling individuals to share their viewpoints on a variety of topics related to today's society. In recent years, the question of whether to receive the COVID-19 vaccination has emerged as a particular issue on these platforms, where users engage by commenting on relevant posts. To efficiently assess whether the sentiment of comments on a given post leans towards pro-vaccination or anti-vaccination stances, deploying a machine learning model for comment classification can be highly beneficial. Should the model demonstrate sufficient accuracy, it could significantly expedite the process for researchers or analysts seeking to understand public opinion dynamics on social media platforms. This approach not only streamlines the analysis of social media content but also offers valuable insights into prevailing behavioral trends and attitudes towards health-related interventions.

2 Preparing the Data

The data used was annotated in two or more rounds by different annotators, giving a decision to make in regards to how to handle discrepancy between the annotators. The main decision was between letting the most common value represent the label, or remove all entries where the annotators do not fully agree. The later path was chosen since it was wishful to minimize ambiguous entries, as training the model on these entries would potentially make its performance less reliable.

A performance baseline was established using the Perceptron model, and the data was transformed to be represented as features by using both the *Count Vectorizer* and using the *TF-IDF Vectorizer* for this purpose. This was done to compare the performance of the chosen baseline algorithm with the different vectorizers, since the *TF-IDF Vectorizer* is slightly more sophisticated and therefore is expected to give better performance to the algorithm.

BERT may be seen a a black box model for the purposes of this report, so we will not describe how it processes the features here.

3 Establishing a Baseline

In order to establish a baseline with a simple linear algorithm (not black-box), Perceptron was used. This model is considered as one of the most straightforward artificial neural network. In its single layer structure it is easy to interpret and performs well. However, due to its simplicity, the performance is limited to perform well only on linearly separable input data, which makes it unrefined in comparison to a more advance model like BERT (Gupta, 2023). With the ambition to compare a simple traditional linear machine learning model to one specilized in natural language processing, the Perceptron model was found to be suitable as a baseline classifier in this task.

F1 score was used to evaluate the different models because a balance between precision and recall was wishful. By calculating the F1 score it is possible to achieve a comparison of the performance of the different models in a comprehensive way taking both precision and recall into account.

Since the task is a multiclass problem, we also want to be able to combine the F1 score for the different classes into a single metric to more easily compare the performance between the different models. By averaging the F1 scores for the different classes they are combined into a single metric. Since the dataset is balanced and the averaging should give all classes equal importance, micro averaging is used to compute the averaged F1 score (Sefidian, 2023).

Lastly, when running the Perceptron model on the data, it achieves an F1-score of 80.0% when using *Count Vectorizer*, and 78.5% when using *TF-IDF Vectorizer*. Since it is expected to receive a higher score from the *TF-IDF Vectorizer* when comparing to an ordinary count vectorization, the result is in somewhat surprising. However, the difference is relatively small and no further attention was payed to the difference.

4 Using Natural Language Processing (NLP)

Bidirectional Encoder Representations from Transformers (BERT) is a framework for natural language processing that interprets the meaning of ambiguous words in text by considering the context of each word (Singh, 2023). It is considered to be conceptually simple and empirically powerful (Devlin et al., 2019).

In the context of classifying text as pro- or anti-vaccination, BERT is deemed highly suitable. In the task of text classification, the model's sensitivity to the context, such as how a negating word combined with a positive word can alter the classification of the text, makes it particularly effective. A refined model that accounts for surrounding words is, therefore, considered appropriate for this task.

We use the wrapped version of BERT provided by the simpletransformers package[1].

4.1 Establishing a NLP Baseline

Firslty, a baseline accuracy for BERT by using the pre-trained model *bert-base-uncased*, configured for binary classification, was established.

The F1-Score, in this instance, was recorded at 49.4%. It is observed that the model consistently predicts a classification of 0 (anti-vaccination), indicating the tendency to categorize comments or texts as anti-vaccination. This outcome is not unexpected given the model's lack of specific criteria for classification. The model's strategy of repetitively guessing the same category results in an F1 score that approaches 50%.

Figure 1 shows a confusion matrix representing the performance of the model.

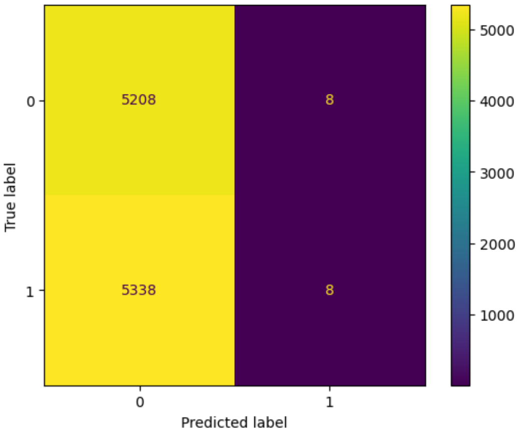

Figure 1: Confusion Matrix for pre-trained BERT model where 0 is the label for anti-vaccination and 1 is the label for pro-vaccination.

4.2 Training a Custom NLP Model

Secondly, a custom BERT model was established by tuning its hyperparameters and running training. The only hyperparameter chosen to be tuned was the number of epochs. Utilizing the fact that the model training function of the simpletransformers package saves a checkpoint copy of the model after each epoch, the model was trained with an arbitrary large number of epochs, and then the copies made after each epoch where loaded to validate their accuracy. These accuracy scores are shown in figure 2.

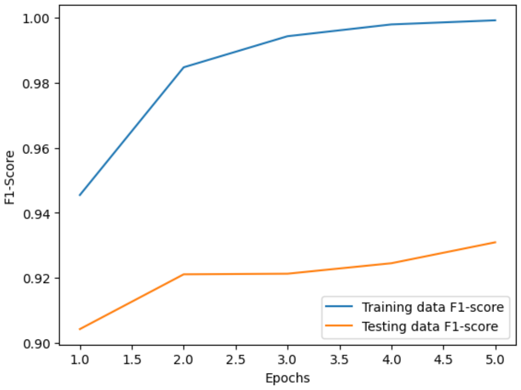

Figure 2: Training data and testing data accuracy for custom trained BERT model over the different number of epochs.

Looking at the training data, the accuracy increases until epoch 2, after which F1-score indi-

[1] https://simpletransformers.ai/

cates that the model tend to overfit the data. From this graph, it is concluded that the best all-round model was achieved after 2 epochs. Evaluating the model at 2 epochs on the testing data gives the result shown in figure 3.

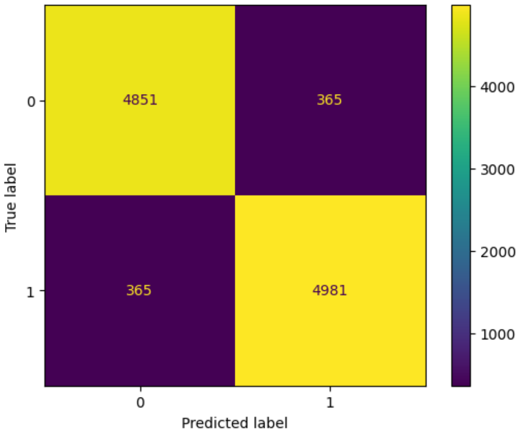

Figure 3: Confusion matrix for custom trained BERT mode where 0 is the label for anti-vaccination and 1 is the label for pro-vaccination.

The performance of the model is significantly improved and reaches an F1-Score: 93.1% on the test data.

5 Evalulating on Evaluation Data

In testing so far, a self trained BERT model trained for 2 epochs achieved the best accuracy score without overfitting.

Training such a model on the entire dataset, to later evaluate it on a separate evaluation dataset, will emulate how the model might perform in a real production environment.

A separate evaluation test dataset is used to simulate a production environment with never-before-seen data. The confusion matrix representing the performance of the model is shown in figure 4. The final F1-score for the evaluation data is 94.3%.

6 Discussion

The performance of the different models is captured and summarized in table 1. The tuned BERT model performance significantly better than the two more traditional models, Perceptron, with Count Vectorizer or with TF-IDF Vectorizer. This result is expected as the BERT model is a lot more advance and built for language tasks. The inner workings of BERT are not considered in the scope

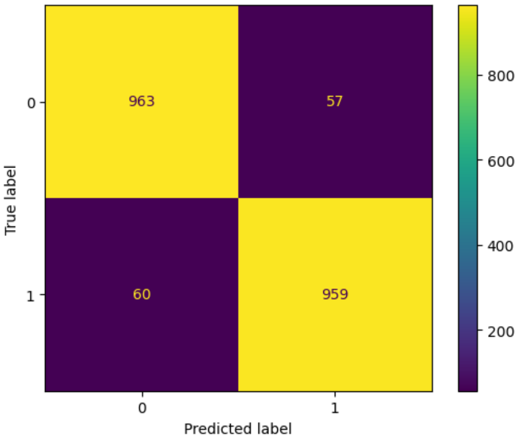

Figure 4: Confusion matrix for custom trained BERT model when run on the evaluation data where 0 is the label for anti-vaccination and 1 is the label for pro-vaccination.

of this report, however the increased training time for the BERT model was significantly higher then the others, hints to a lot more extensive computations in the BERT model. Too better gauge if an accuracy of 94% is good, the results will be compared to how a human would perform, and the incorrect classification of the BERT model will be further looked into.

Model	Description	F1 score
Perceptron	Count Vectorizer	80.0%
Perceptron	TF-IDF Vectorizer	78.5%
BERT	out-of-the-box	49.4%
BERT	tuned	94.3%

Table 1: F1 score for each model evaluated in this task.

6.1 Humans Versus AI

Determining whether a text exhibits a pro- or anti-vaccination stance is far from straightforward. To establish a benchmark for human consensus on varying labels, the data were annotated multiple times. In this task, the baseline for complete agreement was set at 84.4%, and 85.5% when considering the majority score. This baseline ought to be viewed as the maximal expected accuracy that the model could feasibly achieve.

Upon analyzing the performance of the final model, it attained a score of 94.3%. This outcome suggests that, for this specific dataset, the model outperforms human annotators.

It is important to underscore that our dataset is relatively modest in size, and deploying the model on a more extensive dataset could potentially yield a lower F1 score. Nevertheless, we regard the model's performance in this context as commendably effective.

6.2 Incorrect Classifications

To analyze the incorrect classifications it is useful to look at some of the instances the model classified incorrectly:

> ”Do these vaccines prevent Covid19”
>
> ”People who refuse to take the vaccine are the 21st century Rosa Parks”
>
> ”count me in too. im now an antivaxxer”

It is understandable that the model had difficulties in correctly classifying these phrases. Many are ambiguous, see for example “Do these vaccines prevent Covid19”, which is a question rather than an expression of opinion. Removing the first word of this phrase would also change the classification of the phrase, which makes it typically hard to classify the text.

Also, see “People who refuse to take the vaccine are the 21st century Rosa Parks.”. With context, we understand that this comment expresses an anti-vaccination stance (the commenter implies that refusing the vaccine is comparable to Rosa Parks refusing to give up her bus seat as an act of defiance against discriminatory laws and traditions). However, as the model has no knowledge of what Rosa Parks did or who she is, it is understandable that the model would have issues correctly classifying these comments.

There are also examples that are very clear, and which are surprising the model did not get right. For example, see “count me in too. im now an antivaxxer”. Though the first part of the comment is a bit ambiguous, the latter does provide a clear anti-vaccination stance.

6.3 Possible Future Improvements

The different vectorizers that was first tested could be improved by testing other models than the Perceptron, we will however here discuss how to improve the BERT model. First and foremost, a larger data set could be gathered that, if trained, would enable the model to perceive more nuances in the comments for the model to achieve higher accuracy. Further the gathered data could be checked for symbols and emojis that could be hard for a language model to handle. The data instances where annotators disagreed, are probably harder to categorize for a data model as well. Hence, by trying to fix them by manually checking and then including them in the data set, the accuracy could improve.

Lastly, other advanced models similar to BERT and/or different BERT versions could be tested for higher accuracy. Some of the incorrect labeled comments, for instance the Rosa Park comment, would require a very large and advanced language model, to be categorized correctly. When the advanced chat bot language model ChatGPT was given the mentioned comment, it could categorize it correctly as the model was familiar with the historical context surrounding Rosa Parks.

7 Conclusion

The exploratory effort to classify internet comments regarding COVID-19 vaccination stances through a custom natural language processing model based on BERT yielded promising results. By annotating the dataset and establishing a human agreement baseline, we received an accuracy which we aimed the model to be close to. Remarkably, our custom-tuned BERT model surpassed this benchmark, achieving an F1-score of 94.3% in comparison to the human agreement baseline of 85%. This underscores the model's proficiency and demonstrates its usefulness.

In comparison to the Perceptron model, used as a baseline in the sense of a simpler less specified model, the performance of using BERT the turnout was significantly improved as the evaluation score increased from 80.0% to 94.3%.

Future improvements would include exploring a larger dataset to make the model able to handle even more nuanced text. It is crucial to acknowledge the limitations presented by the relatively small size of our dataset and the potential variability that might arise when applying this model to a larger, more diverse corpus of text.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805.

Binay Gupta. 2023. Perceptron learning algorithm. https://www.scaler.com/topics/machine-learning/perceptron-learning-algorithm/. Accessed: 2024-02-17.

Amir Masoud Sefidian. 2023. Understanding micro, macro, and weighted averages for scikit-learn metrics in multi-class classification with example. https://iamirmasoud.com/2022/06/19/understanding-micro-macro-and-weighted-averages-for-scikit-learn-metrics-in-multi-class-classification-with-example/. Accessed: 2024-02-17.

Sumit Singh. 2023. Bert explained: State-of-the-art language model for nlp. https://www.labellerr.com/blog/bert-explained-state-of-the-art-language-model-for-nlp/. Accessed: 2024-02-17.