Sentiment analysis of Covid19 vaccine tweets
Machine Learning for Natural Language Processing 2020

Alexandre Blain
ENSAE Paris
alexandre.blain@ensae.fr

Oscar Villemaud
ENSAE Paris
oscar.villemaud@ensae.fr

Abstract

Social networks reflect public opinion, and can influence it. The covid-19 pandemic has underlined the divergence of opinions about the utility and dangers of vaccines. Some conspiracy theories even assert that covid-19 vaccines are designed for a malicious purpose. The goal of this project is to classify tweets according to the author's opinion on the vaccine. To achieve this, we will use supervised and unsupervised techniques of increasing complexity, ranging from a simple Vader model to a fine-tuned BERT model.

1 Problem Framing

We are using several techniques of sentiment analysis on tweets talking about the Pfizer vaccine to determine writers' opinion on the vaccine. The data we study is the Pfizer Vaccine Tweets Dataset from Kaggle [1]. The Dataset contains about 8000 tweets, along information about users. Our code is available on Github [2].

2 Experiments Protocol

We restrict ourselves to the tweet texts and perform several unsupervised machine learning techniques on them. But first, we need some preprocessing steps. We remove all # and @ from the tweets as well as all other special symbols (like ” n”). We also remove URLs. We can then have a first look at our data, and for instance see what are the most frequent words in our corpus (Figure 1).

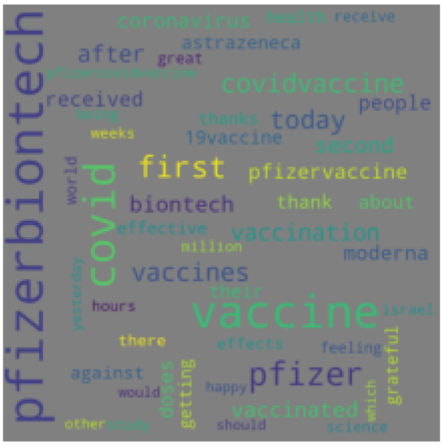

Figure 1: Most frequent words

Then, we apply our first and most naive unsupervised sentiment classification : Vader (Hutto and Gilbert, 2014). We use Vader to compute a probability distribution for each tweet between 3 categories : Positive, Neutral or Negative. From this, by deciding of arbitrary thresholds, we can classify our tweets in 3 categories : Positive, Negative or NaN. This approach gives us an interesting baseline. We can check qualitatively the result by reading manually some tweets from each category.

Our second unsupervised technique is based on Word2Vec (FastText (Bojanowski et al., 2017)). We generate word embeddings using a pre-trained FastText and perform k-Means on these embeddings (one tweet being a data point). The idea is to hope that by dividing the data in 2 or 3 clusters, those will match Positive/Negative tweets or Positive/Neutral/Negative tweets. We can then measure our performance quantitatively by using labels of the Vader analysis and keep the best ones among several K-Means restarts to explore different random initializations.

Finally, we aim at reaching a better performance, so we use a pre-trained BERT (Devlin et al., 2018) classification model through the HuggingFace library. This model needs fine tuning,

[1] https://www.kaggle.com/gpreda/pfizer-vaccine-tweets
[2] https://github.com/alexblnn/NLP-ENSAE

so we have to feed it some labeled data. We try two approaches : training it on the labels we got from the Vader analysis (which represent only a small fraction of the data) and training it on another labeled text database (UCI Sentiment Labelled Sentences (Kotzias et al., 2015)). We analyse our results both qualitatively by printing and reading tweets and their scores and quantitatively by comparing two different approaches we tried with BERT.

3 Results

3.1 Vader

Vader tends to put a very big probability on Neutral (more than 0.7 for about 90% of the tweets). Consequently, we decide to use an arbitrary criterion and to classify tweets as Negative or Positive if the difference between the two probabilities is more than 0.2. This leaves us with about 300 Negative tweets and 1000 Positive ones. The labels we obtain this way seem very accurate according to what we sampled and read. The rest (6500 tweets) is considered unlabeled.

3.2 FastText

When using K-Means on FastText embeddings we can't necessarily notice (by qualitative examination) that each of the two clusters corresponds to a clear positive or negative sentiment. We deepen this analysis by comparing the clusters with Vader labels, try both label permutations to identify which cluster is which. We sometimes get a good accuracy for one of the two sentiments (i.e. 0.76 accuracy on negative sentences according to Vader) but generally the clusters don't represent sentiments, and overall accuracy is low as a result.

3.3 BERT

For our first approach using a pre-trained BERT model, we trained on 1300 positive/negative labels obtained with Vader. Qualitatively, the result is satisfying, as our model successfully detects positive or negative sentiment in most cases. Tricky sentences, such as neutral ones or sentences that include irony or subtext can be misclassified. For instance, the sentence ”the vaccine is a 5G chip” - inspired by a conspiracy theorist's tweet - is classified as positive, albeit with less certainty (0.87 probability) than for easier examples ($\geq$ 0.99 probability).

For our second approach, we used a UCI dataset containing labeled Amazon comments. Qualitatively, the results are satisfying and resemble the results of our first approach. To obtain a more precise evaluation of the two models, we used 1000 tweets from the initial tweet database that were labeled as ”neutral” by Vader. Since our first model is only trained on positive/negative tweets, these neutral tweets can be validly used as a testing set. Here are some quantitative results :

	Precision	Recall	F1-score	Support
Negative	0.49	0.88	0.63	356
Positive	0.88	0.49	0.63	644

Table 1: Vader labels vs UCI labels

We notice that the models disagree often (F1-score = 0.63), which is relatively surprising given our qualitative evaluation. By looking at the sentences on which the two models disagree, we notice that the UCI-trained model makes surprising mistakes on seemingly easy examples such as ”Canada to begin PfizerBioNTech Covid19 Vaccine roll out” (classified negative with probability $\geq$ 0.99) or ”The agency also released new information for health care providers and for patients as the US shipped millions of doses” which was also classified as negative with a high probability.

Perhaps this discrepancy is due to the fact that the UCI based model was trained on data that is not relative to the covid vaccine. Our problem is an aspect-based sentiment analysis problem rather than a straightforward sentiment analysis problem; we can obtain good results with a direct classification approach by training only on tweets relative to the vaccine, but this is not the case if we train on data about an entirely different topic.

4 Discussion/Conclusion

Surprisingly, the simple Vader model yielded coherent labels for about 15% of our tweets. To extend this result we tried two different approaches, FastText and BERT. While the results of our approach using FastText were disappointing, mainly due to the fact that K-Means clusters don't necessarily represent sentiments, fine-tuning a pretrained BERT model with our Vader labels proved to be efficient. However, this model could be improved to better detect irony and subtext in tweets.

References

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 597–606.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.