Using Deep Learning to Correlate Vaccine-lies & COVID-19 Tweets with COVID-19 Time Series*

João Carlos Machado Rocha Pires (UP201806079)
João Francisco de Pinho Brandão (UP201705573)
Pedro Miguel Afonso Teixeira (UP201505916)

Master in Informatics and Computing Engineering
Faculty of Engineering of University of Porto, Portugal

Abstract. Considering the amount of tweets that have been posted regarding COVID-19 and its vaccination process, this paper addresses the preoccupation of identifying which tweets represent correct information and which are not correct. For that, we've created eight models to classify tweets into one of four categories (Accepted, Rejected, Not Relevant, No Stance), achieving the best F1 score of 24.681 with the K-Nearest Neighbor model. We've also compared our predictions with a time series data about World COVID-19 Vaccination progress, analysing semantically the results obtained.

Keywords: COVID-19 · Vaccine · Twitter · Deep Learning · Convolutional Neural Network · Python.

1 Introduction

The first case of COVID-19 to be reported was on 31st of December of 2019. One year later, vaccines from Pfizer, Moderna, Astrazeneca, among others, were distributed and administrated in the population in general. However, since the start of this pandemic and since the start of the production of the first vaccines, a lot has been said about all this, with Twitter being one of the most active social media in terms of COVID-19 and Vaccine related posts. Knowing the potential of the current Natural Language Processing and Machine Learning models and tools, we seek to create a model that is capable of identifying which tweets transmit correct information and which ones are fake news. Our work includes also the comparison between our predictions regarding a subset of tweets and a time series data about number of daily vaccinations in the world for the same tweets time period.

Our contributions are the following:

- Creation of a properly documented Jupyter Notebook with all the code used during this work.
- Building of a K-Nearest Neighbors model to classify a tweet into one of four categories, achieving a binary classification F1 score of 24.681.
- Detailed analysis of the correlation between the predictions made and the time series data about vaccination adherence.

This paper begins with the presentation of related work in which we've based our work. We then describe the data collection, analysis and preparation stages, as well as the presentation of the time series data used, moving then to the models configurations. Before the summary, conclusion and future work, we analyse the models performance and results, comparing the predictions with the time series data about the amount of vaccines administered per day.

2 Related Work

To start this work, we had two baselines. On one side, we had the COVID-19 & Vaccine-lies

* Based on *Using Deep Learning to Correlate Reddit Posts with Economic Time Series during the COVID-19 Pandemic* paper [1]

dataset [2], based on the *CoVaxLies: Identifying the Adoption or Rejection of Misinformation Targeting COVID-19 Vaccines in Twitter Discourse* research paper [3]. This dataset was the primary source of data we had.

On the other side, we had the *Using Deep Learning to Correlate Reddit Posts with Economic Time Series during the COVID-19 Pandemic* paper [1], which described in the detail the procedure of collecting Reddit posts regarding unemployment in the USA during the COVID-19 pandemic and the creation of a model to categorize the posts into one of the defined categories (Unemployment, Cut/Furlough, Employment, Other), as well as to correlate the categorizations with a time series data related with the unemployment rates in the same time period. This paper acted as the baseline for our project in the following aspects:

- Use of the six models, including the corresponding configuration (configuration details presented on the Methods & Models section), described in the paper (Random Forest, Logistic Regression, Linear SVM, CNN 1, CNN 2, CNN 3).
- Data preparation steps, including GloVe 300-dimensions embedding, Stop-Word and Regex-based Removals and VADER sentiment analysis.
- Usage of the binary classification score F1.
- Use of similar approach regarding the comparison between the predictions and the time series data.

3 Dataset

Like we already mentioned, we had the COVID-19 & Vaccine-lies dataset [2] as our primary source of data. This dataset contained a list of development, train and test tweets, identified by their IDs, as well as with at least one of the following categories associated:

- Accept
- Reject
- Not Relevant
- No Stance

We also collected a time series data about the number of new people being vaccinated in order to proceed with a correlation of our model predictions and this data.

3.1 Data Collection

The data collection process for the base dataset was achieved with the usage of the Twitter API, individually configured by each member of the working group according to the individual Twitter Developer Accounts.

For each tweet, we collected the content, the date and the verified flag, which tells us if the author of the tweet has the account verified by Twitter. We consider this a minor personal information, but it adds a lot of value to a tweet, since the majority of the verified accounts are expected to produce correct content.

After the removal of the duplicates, we got ourselves with 4683 tweets in the training dataset and 1299 in the testing one.

Considering that some tweets had more than one labeled category from the four possible ones, we've created a custom function that selects a single category for each tweet. The decision is as follows:

- If one of the categories is 'Reject', then the single category is 'Reject'.
- If the above does not verifies and if one of the categories is 'Accept', then the single category is 'Accept'.
- If the above does not verifies and if the number of occurrences of 'Not Relevant' is equal or higher than for 'No Stance', then the single category is 'Not Relevant'.
- If the above does not verifies, then the single category is 'No Stance'.

3.2 Data Analysis

To start the Data Analysis stage, we've analysed, from the training and testing datasets, the amount of tweets whose author has the account verified. For the training dataset, 357 (7,6%) tweets had the corresponding author's account verified, while in the testing dataset the value was similar in relative terms, 96 (7,4%).

Then, we've analysed the most frequent words, after removing the stop-words using the NLTK removal, in both the training and testing datasets. Since they were quite similar, we present only the table of the most frequent words for the training dataset [The most frequent words for the training dataset.], as an example. It's noticeable that the two most frequent words are the ones related to the dataset subject. The word 'people' and 'I' denote also the personal references, which might indicate that tweets are always referring either to a personal opinion or to the vaccine effect on others. Last, but not least, the 'get' and 'immune' words are referring to the vaccine administration process.

For both training and testing datasets, 18 is the average number of words per tweet. The distribution of tweets containing the words 'covid(-19)' and 'vaccine' is also similar between the training and testing datasets. For the training dataset, 85,80% of the tweets contain the term 'covid(-19)', while 93,40% contain the term 'vaccine(s)'. For the testing dataset, 84,31% of the tweets contain the term 'covid(-19)', while 91,69% contain the term 'vaccine(s)'.

3.3 Data Preparation

The data preparation step included, as mentioned, the drop of the duplicates. For all the models, we've also replaced the tweets' content by a 300-dimensional embedding. The used embedding was GloVe (Global Vectors - Stanford University), which was pre-trained on Wikipedia 2014 and Gigaword 5.

For all the models except the CNN 1, we've also removed stop-words, using the NLTK list of stop-words, and performed a regular expression-based removal of special characters, discarding all non-alphanumeric characters. We've converted the four possible categories for each tweet into a numerical representation (1 to 4).

3.4 Time Series Data

In order to compare our best model with a time series data, we started by looking for a time series data regarding COVID-19 disinformation. Since we did not find that time series data, which would be the ideal for our work, we've then proceeded with another time series data that gathers the information about the World Vaccination Progress [4].

After getting the data by importing it from an CSV file to our Jupyter Notebook, we've filtered it by the time interval from the oldest to the newest tweet dates, considering all the tweets from the training and testing datasets.

We've then plotted the time series data for that time interval alongside with the percentage of accepted and rejected tweets, from both the training and testing datasets, considering that the testing dataset includes the prevision from our best scoring model. Results and the corresponding analysis is described in detail on the Results & Discussion section.

4 Methods & Models

The models we've used in our work are a combination of the six models described in the baseline paper and two extra models we've implemented (Naive Bayes and K-Nearest Neighbors) [Models configuration and results.]. All the models, except the CNN 1, included the GloVe 300-dimensional embedding, the Stop-word and Regex removal. The CNN 1 included only the GloVe 300-dimensional embedding. The CNN 3, more than those, also had the VADER sentiment analysis added to the pre-processing stage.

The CNN models had all the following layer structure:

1. Convolution 1D
2. Max Pool
3. Dense F.C.
4. Dropout
5. Dense F.C.
6. Dropout
7. Softmax Output

To train the CNN models, we've used a Stochastic Gradient Descent optimizer with a learning rate of 0.0001 and the Sparse Categorical Cross Entropy loss function, training with 150 epochs.

Table 1: The most frequent words for the training dataset.

Word	Occurrences
vaccine(s)	4459
covid(-19)	3635
the	903
people	702
I	497
get	429
immune	394

4.1 Model Evaluation

The results of our models are presented in Table 2 [Models configuration and results.]. We used the same evaluation metric of the baseline paper (F1 score). The table presents both the macro and weighted F1 scores (harmonic mean of precision and recall), computed using Scikit-Learn. The model that achieved the best score was K-Nearest Neighbors.

4.2 Ablation Studies

As we can see from the Table 2 [Models configuration and results.], the best CNN was the CNN with the configuration 2, which included the configuration of all the other non-CNN models. This denotes that the VADER sentiment analysis didn't add any major value to our model, which also happened in the baseline paper.

From all the models, the K-Nearest Neighbors was the best, achieving a F1 Macro score of 23.257 and a F1 Weighted score of 24.681, which is still far from being a good F1 score, but it's acceptable due to the nature of our dataset and the amount of pre-processing steps taken before the training stage.

5 Results & Discussion

By looking into the plots that compare the acceptance rate and the rejected rate with the World vaccination progress [Acceptance Rate vs. World Vaccination Progress., Rejection Rate vs. World Vaccination Progress.], we can identify a peak for both rates in an early stage of the vaccination process, as well as a lot of variation in the rates. As the World vaccination progress is moving forward, and the number of people getting the vaccines are increasing, despite the fact we keep having a considerable rate of rejected tweets, we can also see that the rejection rate is quite close to 0 much more times than the acceptance rate, which denotes that, the more vaccines are being administrated, the lower the number of rejected tweets we get.

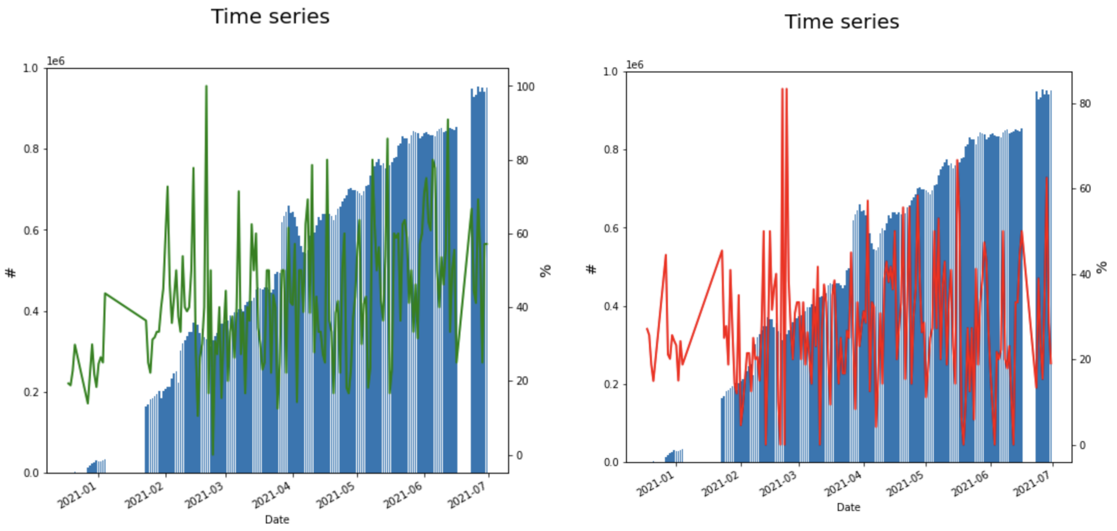
6 Summary, Conclusion & Future Work

In this work, we've created eight models to categorize a set of tweets regarding COVID-19 and the corresponding vaccination. We've included a pre-processing pipeline with GloVe 300-dimensional embedding, Stop-word and Regex removal and even VADER sentiment analysis. Our best scoring model was K-Nearest Neighbors, achieving a F1 score of 24.681.

We've also compared a time series data about the World Vaccination Progress with our best model predictions, analysing semantically the parallel evolution of both. We believe that the time series date we used was not the ideal, as daily vaccinations happening in the world can have a lot of other factors conditioning the total amount (e.g. limited stock, availability, national holidays, prioritization, among others). A much better time series data would be one related to fake news or disinformation regarding COVID-19, but we could not find any resource or similar online.

Table 2: Models configuration and results.

Model	F1 Macro	F1 Weighted	Configuration
Random Forest	19.083	20.349	GloVe 300, Stop-word + Regex removal
Logistic Regression	18.060	19.744	GloVe 300, Stop-word + Regex removal
Linear SVM	18.061	19.703	GloVe 300, Stop-word + Regex removal
K-Nearest Neighbors	23.257	24.681	GloVe 300, Stop-word + Regex removal
Naive Bayes	12.667	12.502	GloVe 300, Stop-word + Regex removal
CNN 1	17.020	20.229	GloVe 300
CNN 2	17.237	20.331	GloVe 300, Stop-word + Regex removal
CNN 3	16.321	18.887	GloVe 300, Stop-word + Regex removal, VADER



(a) Acceptance Rate vs. World Vaccination Progress.

(b) Rejection Rate vs. World Vaccination Progress.

As future work, we identify the need to have a more adequate time series data, regarding fake news or disinformation regarding COVID-19. We also identify the need to improve NLP-related tasks and try different CNN configurations/parameters, possibly using the Optuna Framework.

References

1. Hossu, P., Parde, N.: Using Deep Learning to Correlate Reddit Posts with Economic Time Series during the COVID-19 Pandemic. Location: Department of Computer Science, University of Illinois at Chicago.
2. COVID-19/Vaccine-lies Dataset, `https://github.com/Supermaxman/vaccine-lies/tree/master/covid19`. Last accessed 14 Jul 2022
3. Weinzierl, M., Harabagiu, S.: Identifying the Adoption or Rejection of Misinformation Targeting COVID-19 Vaccines in Twitter Discourse, In: Association for Computing Machinery, Proceedings of the Web Conference 2022, New York, NY, USA https://doi.org/10.1145/3485447.3512039
4. COVID-19 World Vaccination Progress, `https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress`. Last accessed 15 Jul 2022