

Identifying Situational Information during Mass Emergency

Sumit Anand¹, Mehuly Chakraborty² and Diptaraj Sen³

¹University of Engineering and Management Kolkata, India

²University of Engineering and Management Kolkata, India

³University of Engineering and Management Kolkata, India

Abstract

In the advent of Natural Language Processing, what finds itself in much use is analysis. This research paper finds itself in reference to the same that enables it in analysing sentiments of a text. The tasks that were covered in working with NLP includes – firstly, differentiating tweets on the basis of claims and facts, and secondly to create an effective classifier that finds out if a tweet is anti-covid vaccine, pro-covid vaccine or neutral. The beauty of our paper resides in the fact, that we have hit high end accuracies without using hefty algorithms, namely 93% for the first task using Random Forest and 45.4% for the second task using BERT's Algorithm. Our accuracies are the best among all the teams working on the same tasks, which deepens the effect that this paper resonates. The details of the IRMiDis 2021 data challenge have been discussed elaborately here, and we hope our paper marks its significance by virtue of its own merit.

Keywords

Random Forest, BERT's Algorithm, Micro-blogging, Natural Language Processing, classifier

1. Introduction


In periods of such dire needs, where mankind is at loss of life, humanity struggles to stay put in every way possible. Computer Science plays a massive role here by trying to make things more accessible to people, in efficient and sophisticated ways. Social media posts are one of the most important bullets that help coders analyse how and what need to be done in case of massive worldwide emergencies. Our tasks have led us to a discovery of what people think about covid vaccines, leading us to understand what has to be further coped up with to increase awareness in society, and also to create a model that separates claims and facts. Taking help from twitter data has solved our purpose meticulously, and we say this with immense confidence that micro-blogging will be used further in plethora of fields that Computer Science has blessed us with. Both our tasks are but an analysis that micro-blogging has provided us with. To explain more, let's sketch out a brief overview of tasks 1 and 2. For the first task, we were required to differentiate facts and non-facts, the data sets being extracted from twitter. The second task consisted of 2 data sets- the train and the test- that were also extracted from micro-blogging sites, where a classifier was to be built to separate opinions-for, against and neutral-regarding

Forum for Information Retrieval Evaluation, December 13-17, 2021, India

✉ sumit.anand@uem.edu.in (S. Anand); mehuly25@gmail.com (M. Chakraborty); diptaraj.work@gmail.com (D. Sen)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

the covid vaccine. With the help of two efficient algorithms, namely Random Forest [1] and BERT's [2], we have conjured accuracies of 0.93 and 0.45 on tasks 1 and 2 respectively. The paper, hence describes intricately the procedures that we've undertaken to get the highest accuracies among the other teams working on the same IRMiDis 2021 data challenge.

2. Tasks

The tasks were pretty simple, yet immensely engaging.

The first task involved the data set of 11000 tweets from twitter related to the Nepal Earthquake in April 2015. Along with the dataset, sample of few claims or fact-checkable tweets and non-fact-checkable tweets were also provided in text format. What needed to be done was to identify claims and fact checkable tweets.

Examples of claims:

1. @mashable some pictures from Norvic Hospital *A Class Hospital of nepal* Patients have been put on parking lot.

2. @ Refugees: UNHCR rushes plastic sheeting and solar-powered lamps to Nepal earthquake survivors [url]

Example of non-fact checkable tweets:

1. Students of Himalayan Komang Hostel are praying for all beings who lost their life after earthquake!!! Please do...[url]

2. We humans need to come up with a strong solution to create earthquake proof zone's.

The second task was in reference to the present scenario of covid wrenched pandemic that has kept everyone in tatters of their own luck. Two data files were provided namely the Train dataset and the Test data set. The train dataset contains stances of tweets towards COVID-19 vaccines, crawled between November-December 2020, whereas the test dataset contains tweets between March-December 2020 with various vaccine related keywords. What needed to be done was to identify how many people are still skeptical about the covid vaccine. Hence a classifier was to be built for 3 class classification as stated below:

1. AntiVax - the tweet is against the use of vaccines.

2. ProVax - the tweet supports / promotes the use of vaccines.

3. Neutral - the tweet does not have any discernible sentiment expressed towards vaccines or is not related to vaccines

3. Dataset

We have used the datasets provided to us by IRMiDis FIRE 2021 [3], which include a wide range of data ranging from Nepal Earthquake in 2015 to tweets deciphering how many people hold negative views regarding the covid vaccine. For the first task, we have a dataset containing around 11,000 microblogs [4] (tweets) from Twitter that were posted during the Nepal earthquake in April 2015. Along with the dataset, sample of few claims or fact-checkable tweets and non-fact-checkable tweets are also provided as text files, in the following format- Tweetid <||> Tweettext

Example:

592568567247212544<||>RT @NewEarthquake: 4.7 earthquake, 25km S of Kodari, Nepal. Apr 26 13:21 at epicenter (21m ago, depth 10km).

For the second task, IRMiDis FIRE 2021 has contributed 2 datasets – namely the training and the testing. To describe more of it –

The training data set consisted of stances of tweets towards COVID-19 vaccines, crawled between November-December 2020. From this dataset 2,792 crawled tweets texts along with the tweet-IDs and the classes were produced for the same. The testing data set, on the other hand, had in it tweets between March-December 2020 with various vaccine-related keywords. There were tweets annotated by three crowd workers. For 1600 tweets, there was at least majority agreement, i.e., at least 2 out of the 3 annotators provided the same label. The test dataset is formed of these 1600 tweets; each tweet was tagged in with the tweet ID and the tweet text. Referring to these master datasets, the IRMiDis 2021 data challenge was accomplished with great results, that helped us analyse skilfully with no hindrance in the least.

4. Methodology

This section, describes in details the process that we followed to get to the desired results. We have tried to apply Random Forest and BERT's Algorithm to build up the basics of our task, and have been successful in leading out predictions at higher accuracies.

4.1. Task 1: Classifying tweets into Facts and Non-facts

The first task, as mentioned before, asks us to differentiate tweets into facts and non- facts. Hence, to perform this, we have figured out a proper model algorithm that gives us an accuracy of 93%. This task can be divided into 3 non lapping phases, namely – Preprocessing, Feature Selection and Model Selection.

4.1.1. Preprocessing

The first phase helps us clear unnecessary data that hold no relevance to our task of interest. The dataset, hence, was pre-processed before moving on to further phases. We removed links that were present in the data list, along with stop words and tweeter id. Also, user id and punctuations of any kind were removed, hence what was left was a dataset that simply has letters and numbers. Then for the ease of our working, we converted the texts to lower case. This was our pre-processed dataset.

4.1.2. Feature Selection

Proceeding to the next step, we come to Feature Selection, where we try to extract certain features from our newly reduced dataset. This helps us to understand the dataset in a clearer fashion. Here, after pre-processing, we have added an additional column which consists of 1 and 0. The assigning of 1 and 0 is done in the following way – if a tweet has a total number of digits to be 5 or more, we have assigned a 1 in the respective column, while for total digits 4 or lesser, it has a designated 0 in the column.

For example,

1. 'Nepal is the only Hindu country in the world, we need to protect and provide relief in this crisis, hats off to.'

2. ':Earthquake helpline at the Indian embassy in Kathmandu: +977 98511 07021, +977 98511 35141'

Here in the first example, the number of digits is 0, hence the assigned value is 0 and it is a non-fact. In the next example, we see there are 14 digits in total, so likewise the assigned value in column is 1, and also it is a fact. The reason we thought about this criterion is because, while manually inspecting the tweets, we realized that most of the facts are those that involve digits. So now, we experimented with our model by feeding different inputs of the digit count starting from 2,3,4 and 5. What we came to understand is that if the total count of digits is more than or equal to 5, it has a greater tendency of being a fact. Hence, we implemented this, and proceeded to the final phase.

4.1.3. Model Selection

This is the final stage of our task, that involves selecting a proper model to train our data to. After experimenting with many models and algorithm, we found out that Random Forest Classifier gives us the best results. Training our data list after Feature extraction in Random Forest, gives us an accuracy of 0.93. We have incorporated five-fold cross validation for better results. Hence, we could successfully conclude our first task, by efficiently differentiating tweets on the basis of facts and non-facts [5].

4.2. Task 2: 3-class classification on tweets regarding their stance towards COVID-19 vaccines.

As explained previously, task 2 asks us to build an effective classifier for 3-class classification on tweets regarding their stance towards COVID-19 vaccines. The 3 classes are namely – AntiVax (against covid vaccine), ProVax (for covid vaccine) and Neutral. Hence, we have to train our training data through an algorithm, that will provide us with higher accuracies while tested with the testing data. Like Task 1, here too we can divide the process in 3 non-overlapping phases – Preprocessing, Feature Selection and Model Selection.

4.2.1. Preprocessing

Now to get rid of redundant data, we have performed certain functions that helped us lay more focus on our required task. We started by removing all the usernames and hashtags, removing URLs and links and all kinds of special characters, emojis and emoticons. The remaining texts were converted to lowercase to ease our work. Hence our data was pre-processed successfully, and we led onto the next step.

4.2.2. Feature Selection

In this step, we have tried to extract certain features from our pre-processed data so that training the data becomes easy. The classes have been level-coded such that – AntiVax=-1 ProVax=1

Neutral=0. Along with this, we have tokenized the tweets for better working.

4.2.3. Model Selection

After much speculation, we applied a pretrained BERT Model to train our data, that has proved to be immensely effective. Our BERT model is called 'distilBERT-base-uncased' that is a transformers model, smaller and faster than BERT, which was pretrained on the same corpus in a self-supervised fashion, using the BERT base model as a teacher. This means it was pretrained on the raw texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts using the BERT base model. On applying this, the accuracy that we received while testing our training data is around 45.4%, which is a good accuracy compared to all the other algorithms that we have tried. Hence, this led us to the end of the second task [6].

5. Evaluation

This section displays the output of all our algorithms and concepts that we have applied to complete our required tasks. Hence, this consists of our accuracies, precision, recall, MAP, MAP Overall and macro-F1 score that we have retained after training our data through the models which we decided to work on. For both tasks our results were inclined towards the higher ends, that invariably imply that our tasks were a success.

5.1. Task 1: Classifying tweets into Facts and Non-facts

Task 1 was successfully completed by applying Random Forest Classifier to our pre-processed data. The accuracy was 93% which is a very good output based on our dataset. Equivalently our other results have borne amazing outcomes namely a precision (out of 100) of 0.9100, recall (out of 1000) of 0.2165, MAP (out of 100) of 0.0669, and a MAP Overall of 0.1543. The facts and the non-facts were efficiently differentiated, and the results yielded were more than satisfactory.

Team ID	Precision@100	Recall@1000	MAP@100	MAP Overall
ByteCrackers	0.9100	0.2165	0.0669	0.1543

5.2. Task 2: 3-class classification on tweets regarding their stance towards COVID-19 vaccines.

The second task required us to build up an effective model to classify between 3 classes namely - ProVax, AntiVax and Neutral. We've applied a pretrained BERT's model to train our data into, called 'distilBERT-base-uncased' that has yielded us an accuracy of 45.4%. Likewise, our macro-F1 score is 0.440, which is a good outcome based on our data. The model works dexterously and hence, we have successfully completed our second task.

Team ID	Accuracy	macro-F1 Score
ByteCrackers	0.454	0.440

6. Conclusion

This paper holds an amalgamation of two beautiful working algorithms – the Random Forest Classifier and the DistilBERT Algorithm [7]. As mentioned in the paper, we have achieved higher end accuracies in both the tasks. This IRMiDis 2021 data challenge has been much more than a learning experience, for as coders, we have had a working idea about the surrounding society, and their thoughts about issues troubling the nation. Programmers can further utilize this data to procreate something better to treat these societal issues [8]. Our tasks can be further modified to make it better and efficient. Theoretically we have thought about implementing clustering along with BERT's for task 2, and extracting more eminent features other than digits to complete task 1. We have now a much wider grip on Machine Learning and we hope to implement our theories to these tasks to build well-structured models that perform much higher accuracies.

References

- [1] Biau G., Scornet E. A random forest guided tour. TEST 25, 197–227 (2016).
- [2] Miller D., Leveraging BERT for Extractive Text Summarization on Lectures. arXiv:1906.04165 [cs.CL]
- [3] Basu M., Ghosh S., and Ghosh K. 2018. Overview of the FIRE 2018 track: Information Retrieval from Microblogs during Disasters (IRMiDis). In Proceedings of the 10th annual meeting of the Forum for Information Retrieval Evaluation (FIRE'18). Association for Computing Machinery, New York, NY, USA, 1–5. DOI: <https://doi.org/10.1145/3293339.3293340>
- [4] Dutt R., Basu M., Ghosh K., Ghosh S. Utilizing microblogs for assisting post-disaster relief operations via matching resource needs and availabilities, Information Processing and Management, Volume 56, Issue 5, 2019, Pages 1680-1697, ISSN 0306-4573. DOI: <https://doi.org/10.1016/j.ipm.2019.05.010>.
- [5] Chatterjee S., Deng S., Liu J., Shan R., Jiao W. (2018). Classifying facts and opinions in Twitter messages: a deep learning-based approach. Journal of Business Analytics. 1. 29-39. 10.1080/2573234X.2018.1506687.
- [6] Shekhar H., Gangisetty S. (2015). Disaster Analysis Through Tweets. 10.1109/ICACCI.2015.7275861.
- [7] Victor S., Lysandre D., Julien C., Thomas W. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL]
- [8] Elaziz M., Hosny K., Salah A., Darwish M., Lu S., et al. (2020) New machine learning method for image-based diagnosis of COVID-19. PLOS ONE 15(6): e0235187. <https://doi.org/10.1371/journal.pone.0235187>