A Study of Public Sentiment and Influence of Politics in COVID-19 Related Tweets

Anjishnu Mukherjee[1]*, Souparno Bhattacharyya[2], Kinjal Ray[3], Balraj Gupta[4], and Asit Kumar Das[5]

Department of Computer Science and Technology
Indian Institute of Engineering Science and Technology, Shibpur, Howrah-711103, West Bengal, India
mukherjee.anjishnu@gmail.com*,souparno.1407@gmail.com,kinjalray10@gmail.com,balrajgupta11@gmail.com,akdas@cs.iiests.ac.in

Abstract. An increased usage of social media was observed as the Covid-19 pandemic progressed. We extract tweets relevant to the pandemic from publicly available Twitter data and analyse them to understand the change in public emotions over the year along with a detailed analysis of the topics of discussion. We find public health and politics to be the two most dominant topics. Hence, we perform a study where we compare the performance of existing unsupervised classification methods for the task of detecting whether a tweet is medically relevant or politically motivated.

Keywords: Coronavirus, Covid-19, Covid-19 pandemic, pandemic, outbreak, social media, Twitter

1 Introduction

Towards the end of 2019, some patients in Wuhan, China experienced certain influenza-like symptoms which was later identified as 'novel coronavirus' (2019-nCov). As of 30th January 2021, this virus has affected more than 10 crore people and resulted in more than 22 lakh deaths worldwide[1]. During the early stages, to reduce the infection rate and to ramp up the healthcare facilities, the world went into a lockdown and the entire human civilization transitioned into a digital era. This resulted in people spending more time on social media websites like Twitter and Facebook than usual to interact more with other people. These websites even became a primary source of information, as news broke out faster on social media than through conventional media channels.

The objective of this paper is to use this publicly available data to generate insights on factors influencing the public conversation. We first perform sentiment analysis to understand how the outlook of people towards the pandemic changed over the year. Next, we perform word-frequency analysis, TF-IDF analysis, and topic modeling using Latent Dirichlet Allocation (LDA) to find out the major topics discussed on Twitter during the course of this pandemic. The topics identified were mostly either related to politics or public health. Considering these two as dominant factors in public conversation, we further perform a study using existing zero-shot classification methods to see if an unsupervised classifier can label a tweet as either politics or public health with acceptable performance across different metrics. The goal of this classification study is to see whether unsupervised methods can perform reliably in detecting whether a conversation is politically motivated, and thereby helping to reduce the spread of misinformation.

Following this introductory section, we discuss the data collection and preparation methodology in Section 2, the insights generated from the data in Section 3 and the final conclusion in Section 4.

2 Collecting and Preparing Twitter Data

The dataset we use for our task is not directly available in the required format, as Twitter only offers an API to get tweet IDs directly. In this section, we describe how we have collected and prepared the data for our purpose. The major steps of that involved collecting tweet metadata from tweet IDs, preprocessing the metadata to get clean tweet texts, and preparing a collection of tweets as the datasets for analysis and classification tasks. Each of these steps are described in detail next.

2.1 Tweet collection

A publicly available repository containing the tweet IDs of all tweets related to the Covid-19 pandemic is maintained and updated on Github [2]. This repository contains tweets starting from 21st January 2020 till the current date.

For our analysis, we collect all tweet IDs from the year 2020 and hydrate them using the Twitter Developer API [3]. The output of this step is a structured JSON file on which several pre-processing techniques are applied.

2.2 Pre-processing the tweets

One major challenge of Twitter data is that it is multi-modal and can include text, hashtags, links, images, videos, GIFs, emoticons etc. In this paper, we restrict ourselves to a subset of this problem, by focusing on solely text data.

The following pre-processing steps are applied on the tweets:

- As Twitter supports re-tweets, the same tweet can appear multiple times in the JSON file. The file is first de-duplicated using a Python script to remove all re-tweets.
- Next, the *tweets-preprocessor* [4] library is used to remove URLS, hashtags, mentions, reserved words(RT,FAV), emojis, smileys from the tweets.

2.3 Preparing the datasets

Two separate datasets are considered:

- For **the different analyses conducted**, a total of 1,20,000 tweets (10,000 per month) from the year 2020 are considered.
- For **unsupervised classification**, a set of 1,000 tweets are selected randomly from the above dataset. These tweets are then manually labelled to compare the performance of four zero-shot learning based models.

3 Insights

We perform many different kinds of analysis on our collected data, including word frequency analysis, tfidf analysis, sentiment analysis, topic modelling using LDA and unsupervised classification using zero-shot learning methods. Each of these methods and the results obtained from them are described in detail in this section.

3.1 Word Frequency Analysis

Word Frequency Analysis is conducted to find out the words which occurred most commonly in the tweets across the year.

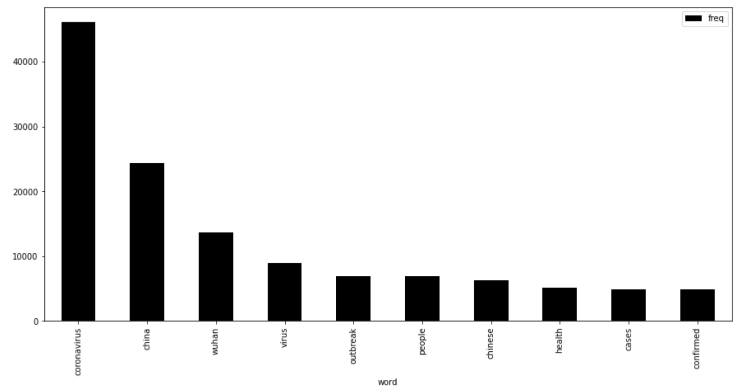

Fig. 1. Most Frequently Occurring Words (2020)

From Fig. 1, “coronavirus” is the most frequent word. This is followed by “china” and “wuhan”, indicating that the people discussed a lot about the origin of the virus as well.

3.2 TF-IDF Analysis

Term Frequency - Inverse Document Frequency (TF-IDF) is an important measure that depicts how important a particular term is to a document. It is very useful in document search and information retrieval. A higher TF-IDF value indicates more importance. Using this score, the keywords of a document can be identified.

We use this analysis to check if the most frequently occurring words are the most important ones as well. Fig. 2 represents the TF-IDF score for the entire year.

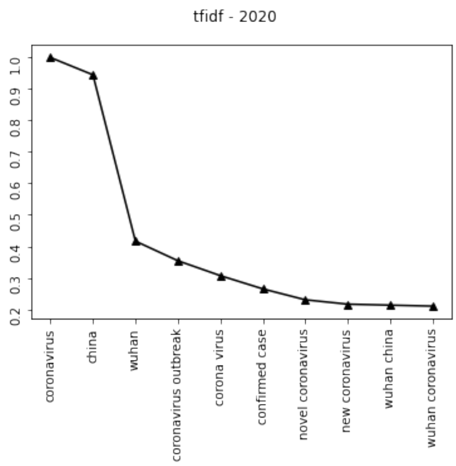

Fig. 2. TF-IDF (2020)

It is seen that the first 3 words are the same in both Word-Frequency analysis, as well as TF-IDF analysis. However, the other words are different in the 2 cases, indicating that frequency alone is not a good indicator of the most important words present in the tweets dataset.

3.3 Sentiment Analysis

Sentiment analysis can be used to find the underlying attitude and emotions from text data. Using sentiment analysis, the overall outlook of people towards the pandemic (whether the majority had a positive or negative feeling) is studied.

This analysis is performed using the TextBlob library [5]. TextBlob returns a polarity score which is a float in the range [-1,1]. For our analysis, tweets having a polarity score greater than 0 are classified as "Positive", while the rest are classified as "Negative". Considering all tweets, we see that 66.98% tweets have a negative outlook, while 33.02% tweets have a positive outlook in the year 2020 as shown in Fig. 3.

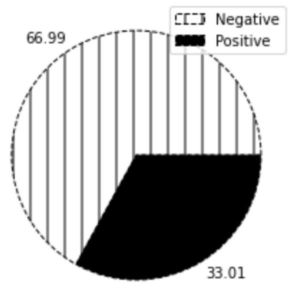

Fig. 3. Overall sentiment distribution across 2020

Hence, it is evident that people, in general, have a negative outlook towards the Covid-19 pandemic. However, to get deeper insights, we take a look at the distribution for every month of the year. Fig 4. shows this distribution.

In January, when Covid-19 was restricted only to China, the positive sentiments are higher. Gradually as it spread to other parts of the world, the positive sentiment decreases (as seen in February and March). Since then, this sentiment remains almost the same till August. The positive sentiment gradually starts increasing from September as lockdown was gradually lifted across the world. A marked increase is seen in the month of December when the news of the arrival of vaccines started coming. Thus, the sentiment analysis of tweets correlates with the developments of the Covid-19 pandemic.

To perform a deeper analysis, the emotional quotient from the tweets are calculated using the NRCLex library [6]. NRCLex categorises emotions into eight categories: fear, anger, anticipation, trust, surprise, sadness, disgust and joy. Among the eight emotions, "trust" and "joy" are positive emotions, while "anger", "sadness", "fear", and "disgust" are considered negative emotions. "Surprise" and "anticipation" can be either positive or negative depending on the context. Fig. 5 shows the obtained results.

The negative emotions like fear and sadness are the most prevalent amongst the people, while positive emotions like joy are ranked pretty low. However, the percentage of joy has increased in December from March, just like the marked increase in positive sentiments during the same phase. This supports our findings in the previous section.

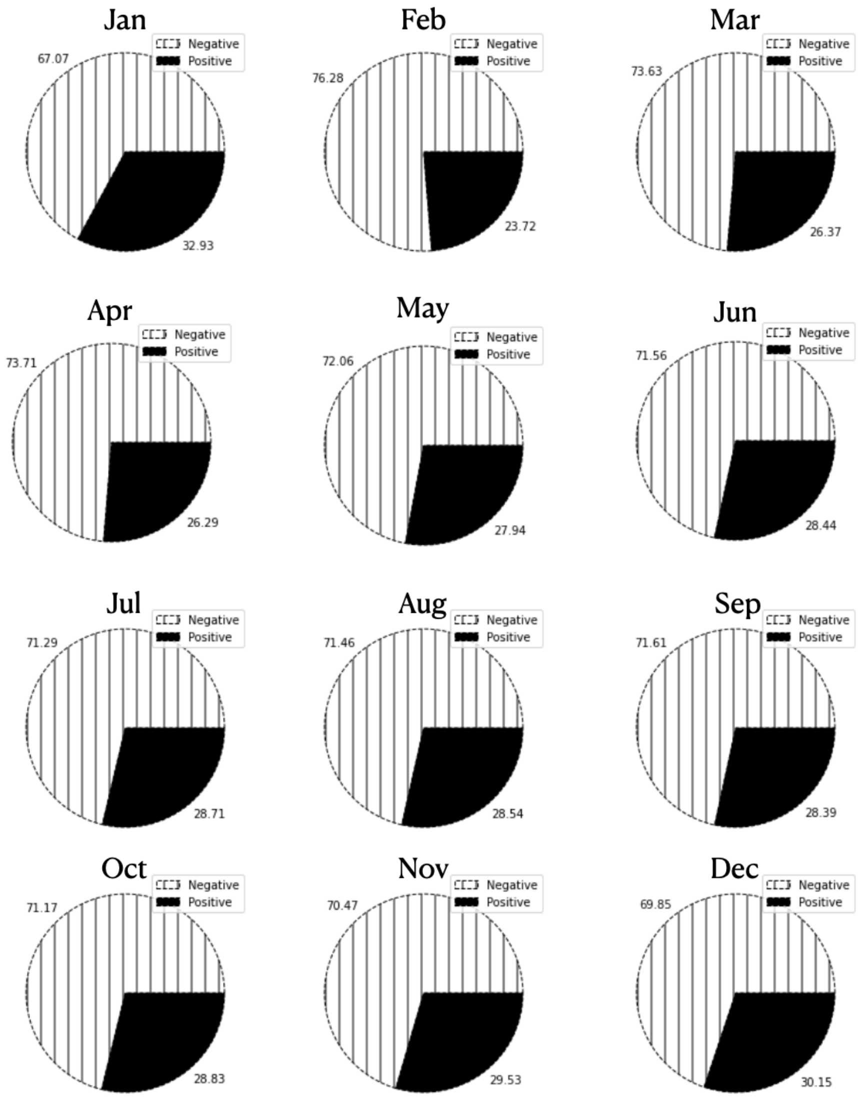

Fig. 4. Monthwise distribution of positive and negative sentiments across 2020

3.4 Topic Modelling using LDA Model

Topic modelling is a statistical modelling technique that is used to find out the abstract topics that appear in a document. Latent Dirichlet Allocation (LDA) [7] is a type of topic model. Using the LDA model, 10 topics are found from all the tweets. The topics are shown in Table 2.

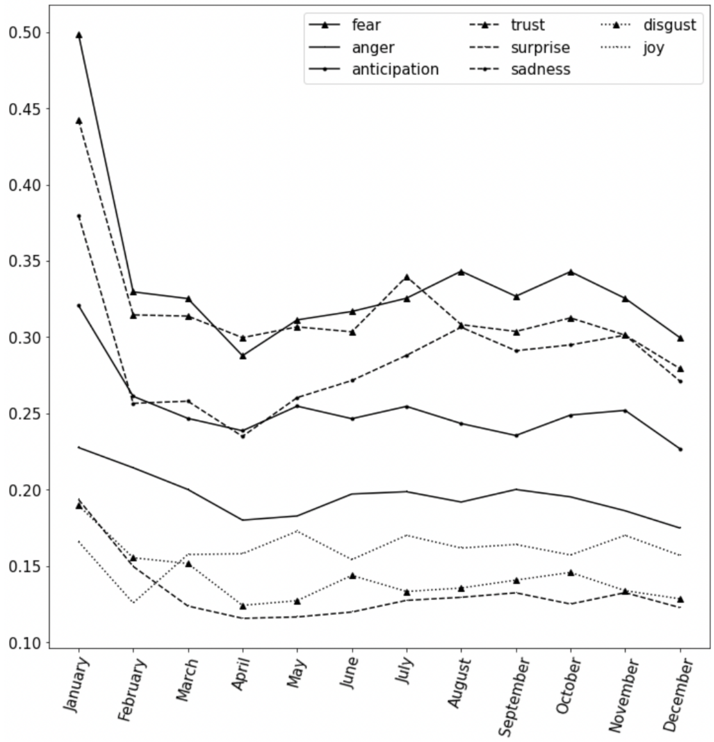

Fig. 5. Change of Sentiment over the Months(NRC Lexicon)

Analysing the above topics, we find most of the tweets belonging to one of 2 categories. They either discuss something related to public health, or they discuss something related to politics. For example, Topic 3 contains words such as "vaccine" and "mask" indicating that it is related to public health, while Topic 9 contains the word "trump" indicating that it is related to politics. On the other hand, Topic 2 contains the words "pandemic" and "country", indicating that it is ambiguous.

3.5 Unsupervised Tweet Classification

Traditionally, zero-shot learning (ZSL) [8] is a technique used to train a classification model for a set of labels and then evaluate it on a different set of labels which the classifier hasn't been exposed to yet. A common approach for ZSL for text data is to use a single model to embed data and class labels into the same space. This can be done using something as simple as word2vec [9] or using more recent techniques like Sentence-BERT(Sentence level Bidirectional Encoder Representations from Transformers)[10]. In this paper, we use a differ-

Table 1. Topics found using LDA

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
need	pandemic	vaccine	people	lockdown	corona	new	coronavirus	trump	like
health	country	get	time	home	virus	day	hospital	make	week
american	still	mask	say	stay	got	case	president	let	since
million	state	know	would	work	bill	one	family	everyone	back
relief	good	wear	go	check	died	year	casos	stop	last

ent approach by treating classification as a natural language inference(NLI) [11] problem. This method embeds sequence-label pairs to the same latent space and also allows us to compare their distances in that space. Using transformer architectures like BERT(Bidirectional Encoder Representations from Transformers) [12], NLI datasets can be modelled using sequence-pair classification. This approach was first proposed by Yin et al. [13] in 2019 and it uses a pre-trained MNLI (Multi-Genre Natural Language Inference)[14] sequence-pair classifier for ZSL with decent accuracy.

We use 4 different zero shot learners based on the approach above to generate labels for 1000 pre-processed tweets. We choose to consider it as a multi-label classification task, with the models predicting probability scores of each tweet for the labels **public health** and **politics**. Based on our data analysis earlier, we have found these 2 topics to be the most dominant ones in the dataset, hence we choose to use them as labels for classification. To get a single label from the predicted class scores, we use Algorithm 1.

Algorithm 1: Choosing a single label from predicted scores of model

Result: Predicted Label
a = output score of model for **public health**;
b = output score of model for **politics**;
c = threshold score for ambiguity;
if $a \leq c$ **or** $b \leq c$ **then**
  predicted label = ambiguous;
else
  if $a \geq b$ **then**
    predicted label = a;
  else
    predicted label = b;
  end
end

Using this procedure, we get a total of 4000 labels from our models. Next, we manually label each of the 1000 tweets using majority voting to resolve any confusions. We treat these labels as the true labels. For the true labels, we are unable to label tweets in language other than English, so we treat them as ambiguous. However, some of our models can understand some languages other than English. To remove this disparity, we remove all the tweets which are not written in English. We are left with 710 tweets after this. For these tweets, we compare the accuracy of the 4 models and also some other metrics as can be seen in Table 3.

Table 2. Comparison of model performances

Model Name	Class	Precision	Recall	F1-Score	Support	Accuracy
bart-large-mnli	ambiguous	0.59	0.90	0.71	243	**0.70**
	public health	0.80	0.64	0.71	290	
	politics	0.87	0.52	0.65	177	
	macro average	0.75	0.69	0.69	710	
	weighted average	0.74	0.70	0.70	710	
bart-large-mnli-yahoo-answers	ambiguous	0.75	0.39	0.51	243	**0.65**
	public health	0.70	0.81	0.75	290	
	politics	0.53	0.74	0.62	177	
	macro average	0.66	0.65	0.63	710	
	weighted average	0.67	0.65	0.63	710	
distilbart-mnli-12-1	ambiguous	0.44	0.98	0.61	243	**0.53**
	public health	0.79	0.25	0.38	290	
	politics	0.86	0.37	0.51	177	
	macro average	0.70	0.53	0.50	710	
	weighted average	0.69	0.53	0.49	710	
distilbart-mnli-12-6	ambiguous	0.65	0.74	0.69	243	**0.72**
	public health	0.79	0.74	0.69	290	
	politics	0.73	0.65	0.69	177	
	macro average	0.72	0.72	0.72	710	
	weighted average	0.73	0.72	0.72	710	

The 4 models we use are BART (Bidirectional and Auto-Regressive Transformers) [15], a version of BART fine-tuned on Yahoo Answers topic classification [16], and 2 distilled versions of BART created using No Teacher Distillation method [17] which vary in the number of layers they use. One of the Distilbart models perform best on this data, followed closely by the original BART itself. The BART model fine-tuned on Yahoo Answers is slightly inclined towards predicting label as politics, to the extent that it misclassifies even some blank tweets as politics.

We also use a maximum majority voting ensemble of the 3 best models and get an overall accuracy of 73 % on the data, which showed not much significant increase compared to the best model, which has an acuracy of 72%.

4 Conclusion

In the scenario of a global pandemic like Covid-19, we analyse public conversation on social media platforms like Twitter to understand the main topics people are interested in and also the change in their sentiments over the months. We also compare unsupervised classification techniques for labelling the data and see that these methods are accurate to some extent when it comes to automating the labelling process for large amounts of such data. Being able to understand whether a conversation is politically motivated in a disaster scenario is very helpful not only for the people affected but also for moderating conversations online, and our study shows that unsupervised classifiers can help in this regard. This work can be extended for different general disaster scenarios by extracting information from social media other than Twitter as well, by performing multi-modal analysis.

References

1. World Health Organisation official dashboard for tracking covid-19, `https://covid19.who.int`
2. Tweet IDs related to COVID-19 pandemic, `https://github.com/echen102/COVID-19-TweetIDs`
3. Twitter API for Developers, `https://developer.twitter.com/en/docs/tools-and-libraries`
4. Preprocessor Library, `https://github.com/s/preprocessor`
5. Sentiment Analysis using TextBlob, `https://github.com/sloria/TextBlob/blob/dev/textblob/en/sentiments.py`
6. NRCLex Library, `https://github.com/metalcorebear/NRCLex`
7. David M. Blei, Andrew Y. Ng and Michael I. Jordan: Latent Dirichlet Allocation, `https://jmlr.org/papers/volume3/blei03a/blei03a.pdf`
8. Ming-Wei Chang, Lev Ratinov, Dan Roth and Vivek Srikumar: Importance of Semantic Representation: Dataless Classification, `https://www.aaai.org/Papers/AAAI/2008/AAAI08-132.pdf`
9. Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space, **arXiv:1301.3781**
10. Nils Reimers and Iryna Gurevych: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, , **arXiv:1908.10084v1**
11. Repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state-of-the-art for the most common NLP tasks, `http://nlpprogress.com/english/natural_language_inference.html`
12. Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, **arXiv:1810.04805**
13. Wenpeng Yin, Jamaal Hay and Dan Roth: Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach, **arXiv:1909.00161**
14. Adina Williams, Nikita Nangia and Sam Bowman: The Multi-Genre NLI Corpus, `https://cims.nyu.edu/~sbowman/multinli/`
15. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov and Luke Zettlemoyer: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, **arXiv:1910.13461**
16. BART finetuned on Yahoo Answers topic classification dataset, `https://huggingface.co/joeddav/bart-large-mnli-yahoo-answers`
17. Distilled BART, `https://github.com/huggingface/transformers/tree/master/examples/seq2seq#distilbart`