Polarity Based Sentiment Analysis about COVID-19 Using Reddit Data

1st Fathima Mirza
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
fathima.mirza@g.bracu.ac.bd

2nd Md Yusuf Khan
ICT & Computer Science
International Hope School Bangladesh
Dhaka, Bangladesh
yusuf.khan@ithsbd.net

Abstract—2020 has presented us with a unique problem, COVID-19- a pandemic reaching almost all over the globe. People all over the world have been reaching out to social media to express their views and opinions. Reddit is discussion-based website and therefore a good source of data for opinions of people regarding different subjects. This paper aims to scrape the data from the reddit regarding the latest information on COVID-19 and run various traditional machine learning algorithms on them to perform polarity based sentiment analysis to see what reditters feel about COVID-19.

I. Credits

This document has been adapted by Yulan He from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2020 by Steven Bethard, Ryan Cotterrell and Rui Yan, ACL 2019 by Douwe Kiela, Ivan Vulic,´ NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Michell and Stephanie Lukin, 2017/2018 (NA)ACL bibtex suggestions from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL-2016 by Margaret Mitchell, ACL-2012 by Maggie Li and Michael White, those from ACL-2010 by Jing-Shing Chang and Philipp Koehn, those for ACL-2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, those for ACL-2005 by Hwee Tou Ng and Kemal Oflazer, those for ACL-2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats. Those versions were written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the International Joint Conference on Artificial Intelligence and the Conference on Computer Vision and Pattern Recognition.

II. Introduction

Coronavirus disease 2019 (COVID-19) is an infectious dis-ease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It is said to have originated in Wuhan, the capital of China's Hubei province in 2019 and has since spread globally, leading to a pandemic. The outbreak in 2019–2020 has caused at least 713,171 infections and 33,597 deaths
[1]. Understandably, people are feeling scared, overwhelmed, panicked and various other emotions over this. In today's generation, the people turn to social media for expressing their views. Reddit is one such platform where there is various discussions amongst people under various subreddits. According to redditmetrics.com on 30 March 2020, the subreddit about Corona Virus has been the second most trending for the month of March 2020. This creates an interesting scope to perform a sentiment analysis on the opinions gathered from reddit. Sentiment analysis (also called opinion mining) refers to the application of natural language processing, computational linguistics, and text analytics to identify and classify subjective opinions in source materials (e.g., a document or a sentence) [11]. Sentiment Analysis is performed basically in two ways, using statistics only, and on the basis of both statistics and linguistics. For the former, it includes algorithms such as Bag of Words (BOW) where the context is ignored [12]. The latter tries to incorporate linguistic features along with the statistical ones [13]. Sentiment analysis can also be polarity based [14] or have degrees of sentiment [15].

This paper first scrapes the reddit for titles within subreddits that are related to corona virus. Once the data is collected from reddit, the polarity (whether the sentence is positive or negative) of the sentence is gathered by using the built-in sentiment analyzer. The data is then cleaned by converting all the text to smaller alphabets, removing the stop works, punctuations, and tokenizing them. This is done for both positive sentences and negative sentences. Once the polarity is set, and traditional machine learning methods are applied to see how well they perform.

III. Related Works

[2] published the first twitter dataset for COVID-19 for enabling informed solutions and prescribing targeted policy interventions to fight this global crisis. [3] attempted to find a solution to quantifying the misinformation that is spreading like rapid fire on twitter. They identified that healthcare accounts were least likely to contain misinformation. They also discovered which keywords and hashtags were most unreliable. [4] also attempted to do quantify misinformation along with the panic it causes people. [5] tried to perform predictive analysis on the outbreak of COVID-19 on the basis of information from social media. Twitter data has been and is been exploited by researchers to perform a number of tasks, such as identifying misinformation, aiding in relieving

panic, performing predictive analysis on outbreaks, just to mention a few. However, another equally, if not richer source of information is available on reddit. [6] is one of the few research works targeted at COVID-19 that targeted reddit. This paper identifies the fact that there is no good dataset for reddit data as well as there is not sufficient work done into exploring the polarity of sentiments of reddit users (who are some of the most expressive people on the internet) for COVID-19. This paper aims to find a method to get data from reddit for COVID-19 as well as automatically find the polarity of the sentences. Traditional machine learning techniques are then applied to the dataset to see how they perform.

IV. Methodology

The methodology shall first discuss how the data was collected, then move on to how the data was labelled and what Natural Language Processing techniques were applied, and finally, how the traditional machine learning algorithms are applied in this case. PRAW is a python-based Reddit API wrapper which allows researchers to scrape data from reddit legally. This paper has also abided by the guidelines and used PRAW for scraping the data from reddit. At first, this paper tries to create a dataset using subreddits. Only the title field is used in this case because the titles of reddit are usually set in a manner that expresses the entire sentiment of the post. Using the subreddit r/covid19 and r/CoronaVirus, a set of above 1824 titles were collected. The data collected does not have any polarity labels as of yet. The data is labelled using the built in sentiment analysis tool in nltk. Employing SentimentIntensityAnalyzer from the nltk package, the titles are analyzed and tagged as positive, negative or neutral. The information is then saved into a csv file, with the titles and their corresponding labelling. Figure 1 demonstrates a plot that shows the percentage of sentences identified as positive, negative and neutral.

The positive and negative sentences are separated. The neutral sentences are dropped because they are assumed to have no impact on the sentiment. Only adjectives from the positive and negative sentences are extracted to form the feature set on the assumption that adjectives are highly informative of positive and negative sentiments – especially given the nature that the dataset is extremely small [7,8,9,10]. The adjectives are extracted using part of speech tagging method. Using the re module of python, and a regular expression of '[^(a-zA-Z)ns]' punctuations are removed for both positive and negative sentences. Afterwards, using the stopwords module for English from nltk package, the stopwords are also filtered out from the positive and negative sentences. The sentences are then tokenized, the adjectives are extracted and their frequency distribution is calculated. 5000 features are saved for testing and 20000 features are used for training. The features are selected randomly. Na¨ıve Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, Stochastic Gra-dient Descent and Support Vector Classifier was trained using the training set and validated using the testing set to classify the polarity of the reddit titles regarding COVID-19.

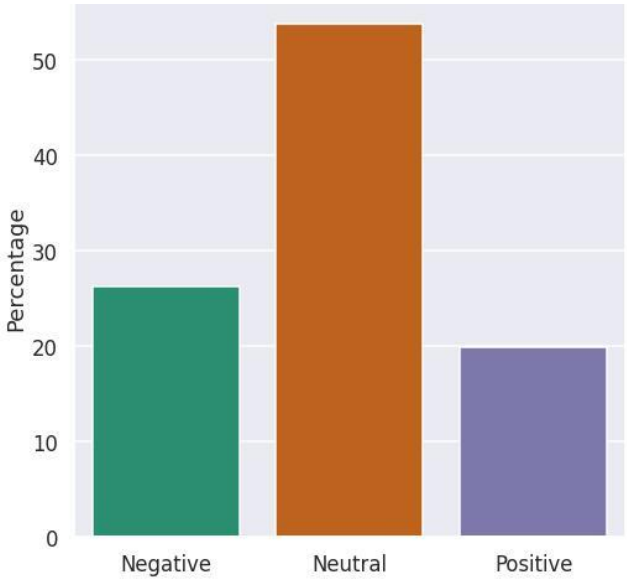

Fig. 1. Percentage of sentence in the dataset that are positive, negative and neutral

V. Results

On analyzing the six machine learning algorithms, it is dis-covered that Stochastic Gradient Descent Algorithm provides the best accuracy for classifying the polarity of sentiment in titles of reddit regarding COVID-19 and Support Vector Machine Algorithm comes a close second. The accuracy of Naive Bayes Algorithm is 82.00%, Multinomial Naive Bayes Algorithm is 81.45%, Bernoulli Naive Bayes Algorithm is 78.54%, Logistic Regression Algorithm is 79.27%, Stochastic Gradient Descent Algorithm is 85.45% and Support Vector Machine Algorithm is 83.45%. This is shown in Table 1. Figure 2 demonstrates this graphically. Figure 3-8 shows the performance metrics, Accuracy, Precision and F1 Score of each of the 6 different algorithms demonstrated graphically. Table II-VII shows the confusion matrix for the 6 different algorithms.

The results below show the Stochastic Gradient Descent is the algorithm that performed most successfully at identifying positive and negative titles from reddit titles. This is probably because gradient descent employs a trick that minimises the cost function and hence this optimization resulted in the best result.

TABLE I
Accuracy Metrics for different algorithms

Algorithm	Accuracy
Naive Bayes	82.00%
Multinomial Naive Bayes	81.45%
Bernoulli Naive Bayes	78.54%
Logistic Regression	79.27%
Stochastic Gradient Descent	85.45%
Support Vector Machine	83.45%

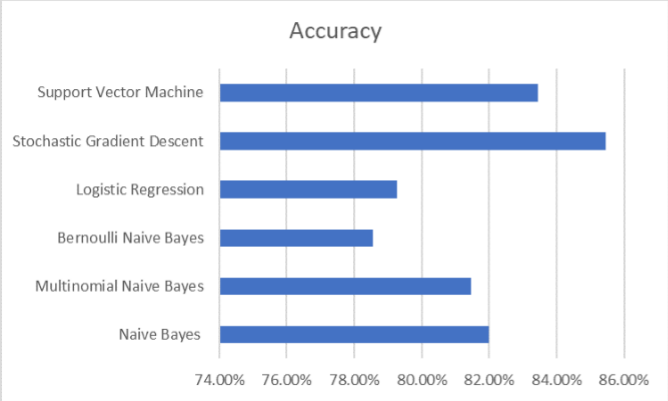

Fig. 2. Accuracy Metrics of different algorithms

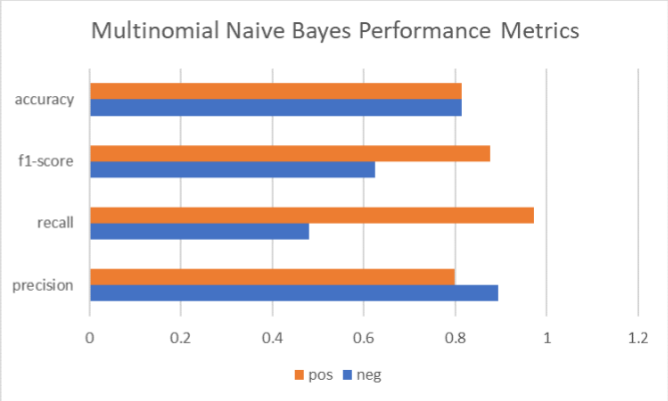

Fig. 5. Multinomial Na¨ıve Bayes Performance Metrics

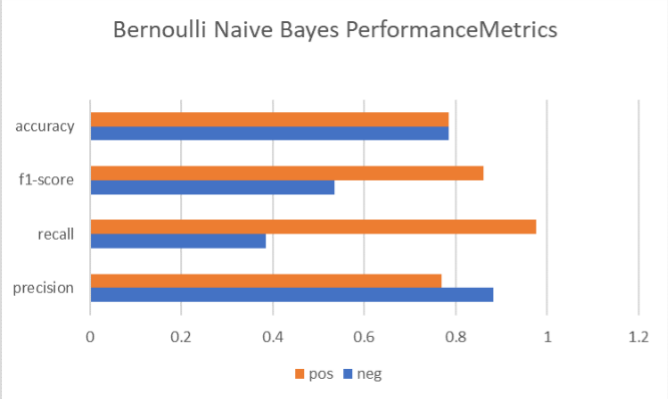

Fig. 3. Bernoulli Naive Bayes Performance Metrics

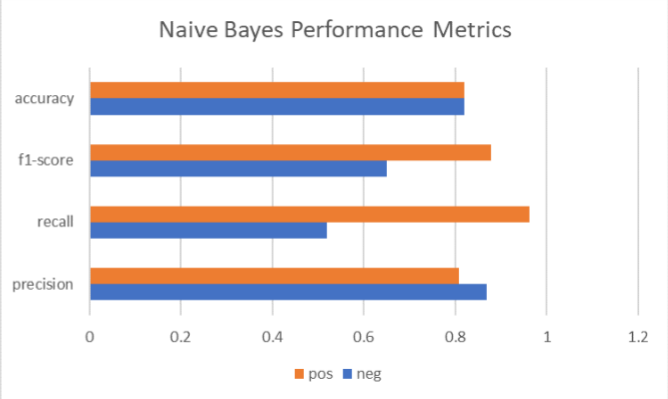

Fig. 6. Na¨ıve Bayes Performance Metrics

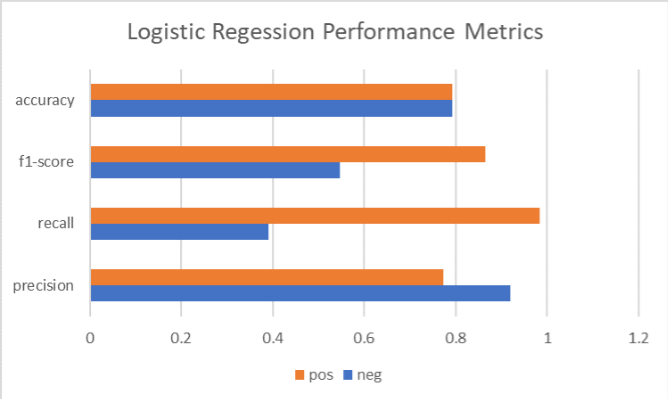

Fig. 4. Logistic Regression Performance Metrics

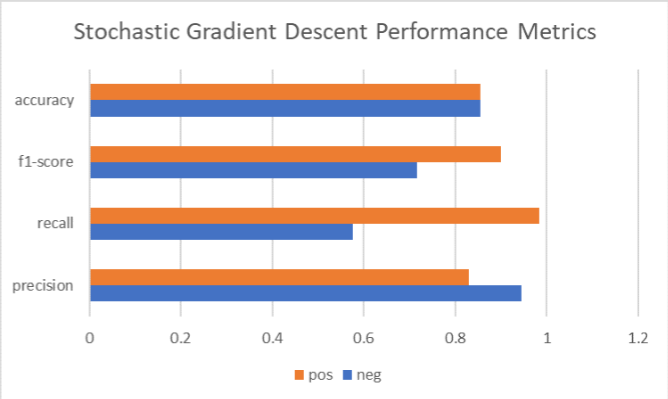

Fig. 7. Stochastic Gradient Descent Performance Metrics

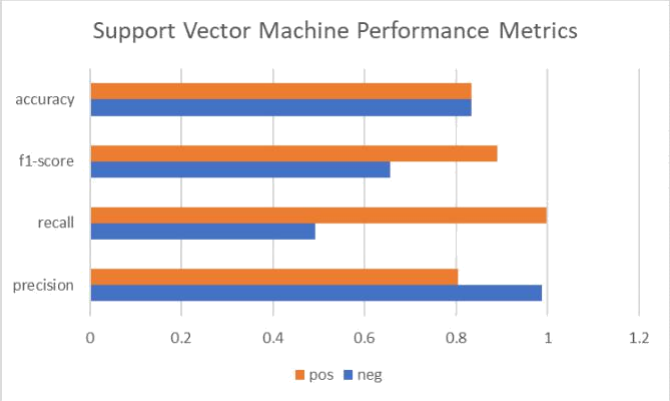

Fig. 8. Support Vector Machine Performance Metrics

TABLE II
CONFUSION MATRIX FOR MULTINOMIAL NA¨IVE BAYES

		Actual Class: Positive	Actual Class: Negative
Predicted class	Positive	TP =85	FP =92
	Negative	FN=10	TN = 363

TABLE III
CONFUSION MATRIX FOR BERNOULLI NA¨IVE BAYES

		Actual Class: Positive	Actual Class: Negative
Predicted class	Positive	TP =68	FP = 109
	Negative	FN=9	TN = 364

TABLE IV
CONFUSION MATRIX FOR LOGISTIC REGRESSION

		Actual Class: Positive	Actual Class: Negative
Predicted class	Positive	TP =69	FP = 108
	Negative	FN=6	TN = 367

TABLE V
CONFUSION MATRIX FOR STOCHASTIC GRADIENT DESCENT

		Actual Class: Positive	Actual Class: Negative
Predicted class	Positive	TP = 102	FP =75
	Negative	FN=6	TN = 367

TABLE VI
CONFUSION MATRIX FOR SUPPORT VECTOR MACHINE

		Actual Class: Positive	Actual Class: Negative
Predicted class	Positive	TP =87	FP =90
	Negative	FN=1	TN = 372

TABLE VII
CONFUSION MATRIX FOR NA¨IVE BAYES

		Actual Class: Positive	Actual Class: Negative
Predicted class	Positive	TP =92	FP =85
	Negative	FN=14	TN = 359

VI. CONCLUSION

This paper successfully described a way to collect data from Reddit titles and label those data as positive sentiment, negative sentiment or neutral sentiment. It further showed the performance of six traditional machine learning algorithms for the purpose of classifying titles from Reddit as positive or negative where Stochastic Gradient Descent Classifier gave the best results. However, this paper only intends to set a rudimentary template for sentiment analysis using reddit data. Lots of improvement could have been done. For example, more subreddits could have been used to enrich the dataset. Because the dataset is very small in size, overfitting could easily occur. Title of a post along with its comment should have also been analyzed to get an accurate depiction of the sentiment of the post. Instead of using the default nltk sentiment analyzer, humans should have labelled the data as positive or negative to get an actual representation of the sentiment of the post. The bag of words methods only considered adjectives to denote the features whereas it is better to use adverbs along with adjectives for better results [16]. Furthermore, no contextual or linguistics methods were employed in this paper. The authors intend to solve all these issues soon as well as implement a method where CNN can be used for automatic feature extraction and LSTM or MHSA (depending on which gives better results) can be used to contextual and semantic dependencies [17,18].

REFERENCES

[1] Wikipedia contributors. (2020, March 29). Coronavirus disease 2019. In Wikipedia, The Free Encyclopedia. Retrieved 21:07, March 29, 2020, from https : ==en:wikipedia:org=w=index:php?title = Coronavirus disease 2019&oldid = 948019644

[2] Chen, Emily & Lerman, Kristina & Ferrara, Emilio. (2020). COVID-19: The First Public Coronavirus Twitter Dataset.

[3] Kouzy R, Abi Jaoude J, Kraitem A, et al. (March 13, 2020) Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. Cureus 12(3): e7255. DOI 10.7759/cureus.7255

[4] Cuan-Baltazar, Jose & Munoz˜-Perez, Maria & Robledo-Vega, Carolina & Perez´-Zepeda, Maria & Vega, Elena. (2020). COVID-19 misinformation on the internet: The other epidemy (Preprint). 10.2196/preprints.18444.

[5] Jahanbin, Kia & Rahmanian, Vahid. (2020). Using twitter and web news mining to predict COVID-19 outbreak.

[6] Cinelli, Matteo & Quattrociocchi, Walter & Galeazzi, Alessandro & Valensise, Carlo & Brugnoli, Emanuele & Schmidt, Ana & Zola, Paola & Zollo, Fabiana & Scala, Antonio. (2020). The COVID-19 Social Media Infodemic.

[7] Benamara, Farah & Cesarano, Carmine & Picariello, Antonio & Refor-giato Recupero, Diego & Subrahmanian, Vs. (2005). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. ICWSM.

[8] Sharma, Raksha & Gupta, Mohit & Agarwal, Astha & Bhattacharyya, Pushpak. (2015). Adjective Intensity and Sentiment Analysis. 2520-2526. 10.18653/v1/D15-1300.

[9] Andreevskaia, Alina & Bergler, Sabine. (2006). Sentiment Tagging of Adjectives at the Meaning Level. 4013. 336-346. 10.1007/11766247 29.

[10] Bhadane, Chetashri & Dalal, Hardi & Doshi, Heenal. (2015). Sentiment Analysis: Measuring Opinions. Procedia Computer Science. 45. 808-814. 10.1016/j.procs.2015.03.159.

[11] Luo, Tiejian & Chen, Su & Xu, Guandong & Zhou, Jia. (2013). Sentiment Analysis. 10.1007/978-1-4614-7202-5 4.

[12] Zhang, Yin & Jin, Rong & Zhou, Zhi-Hua. (2010). Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics. 1. 43-52. 10.1007/s13042-010-0001-0.

[13] Agarwal, Alekh & Bhattacharyya, Pushpak. (2005). Sentiment Analysis: A New Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified.

[14] Wang, Min & Shi, Hanxiao. (2010). Research on sentiment analysis technology and polarity computation of sentiment words. 1. 10.1109/PIC.2010.5687438.

[15] Kim, Evgeny & Klinger, Roman. (2018). A Survey on Sentiment and Emotion Analysis for Computational Literary Studies.

[16] Benamara, Farah & Cesarano, Carmine & Picariello, Antonio & Refor-giato Recupero, Diego & Subrahmanian, Vs. (2005). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. ICWSM.

[17] Yang, Peng & Zhao, Guangzhen & Zeng, Peng. (2019). Phishing Website Detection based on Multidimensional Features driven by Deep Learning. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2892066.

[18] Xiao, Xi & Zhang, Dianyan & Hu, Guangwu & Jiang, Yong & Xia, Shutao. (2020). CNN-MHSA: A Convolutional Neural Network and multi-head self-attention combined approach for detecting phishing websites. Neural Networks. 125. 10.1016/j.neunet.2020.02.013.