Tweet Sentiment Analysis and Topic Modeling

Azat Dovgeldiyev

Department of College Computing and Digital Media

DePaul University, Chicago, IL 60604

Email: adovgeld@depaul.edu

Abstract

In today's world, social media is everywhere, and everyone is in touch with it every day. More people are using the internet and social media to share their thoughts and views. As a result, the number of user-generated containing sentiment data increased, and with the advent of the COVID-19 pandemic, social media has quickly evolved into a critical communication tool for the production, distribution and consumption of information. The objective of this paper is to compare classification models in predicting whether the text contains positive, negative, or neutral sentiment and applying clustering models to identify topics. Two different datasets used in this work, with 1.6 million and 136 thousand tweets related to coronavirus. The system for classification models was trained using 77% of the dataset and was tested using the remaining 33%. The results show a maximum accuracy rate of 82%, which shows the efficiency of the Support vector classifiers. 3 topics generated with Latent Dirichlet Allocation.

Keywords – sentiment analysis; machine learning; twitter; natural language processing; LDA

I. INTRODUCTION

Social media has become a huge part in our daily lives. It creates a connection between people and the outside world. Social media helps us to display our lives in a private, easy, and self-directed manner. People are more reliant on posts and tweets shared on social media sites such as Twitter, Instagram, and Facebook. Sentiment analysis is one of the most common research topics in the field of natural language processing nowadays, it is text mining that recognizes and extracts subjective information from source content, allowing a company to better understand the social sentiment of its brand, product, or service by analyzing online conversations (1). COVID-19, a form of coronavirus that first appeared in Wuhan, China at the end of 2019, has become one of the most widely discussed and distributed diseases in the world.

This paper deals with an extensive analysis of the sentiments emoted through the tweets since the beginning of 2020 pertaining to the novel COVID-19 using python programming language. While one dataset contains only coronavirus related tweets that have been published during the time span between March 2020 to September 2020, the other dataset contains tweets between April 2020 and May 2009. Sentiments of the tweets have been labeled and WordCloud is presented for every sentiment. Later, different topics were generated from COVID-19 tweets and visualized using pyLDAvis.

The motivation of the paper lies in identifying whether a user-generated text expresses positive, negative, or neutral opinion and in classifying text data based on their topics. Supervised learning applied to classify the polarity level of tweets. To be precise Linear support vector classifier achieved 81% of accuracy on COVID-19 related tweets. Other than the above, this paper proposes topic modeling with Latent Dirichlet Allocation for multi-gram words and visualization tool for selected topics.

The paper is organized as follows: section II provides the literature review. The proposed system is introduced in section III, and finally conclusion and future work are presented in section IV.

II. LITERATURE REVIEW

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Anuja P Jain and Asst. Prof Padma Dandannavar (2016) proposed machine learning approach using Naïve Bayes and Decision trees on three different datasets. This approach employed a ‘bag-of-words’ method using words in the independent features in a feature vector to represent the documents. The performance of the models compared on accuracy, precision, recall and F1-Score. Based on their findings, decision trees performed extremely well showing 100% accuracy, however the size of data was not big enough.

Another approach for sentiment analysis is lexicon based, which is based on finding the opinion lexicon for calculating the sentiment for a given text. Christopher SG Khoo and Sathik Basha Johnkhan (2017) (2) introduced a new sentiment lexicon and compare other existing lexicons using Amazon product review data. Proposed new sentiment lexicon with other three models were equally good for product review polarity level categorization, obtaining accuracy rates of 75%-77%. A limitation of this study was using only single-word terms.

Topic modeling can be useful to find out the different topics that product reviews, survey responses, emails cover. Latent Dirichlet Allocation (LDA) is a popular approach for uncovering hidden topics: multinomial probability distribution over terms. While LDA produces some sensible topics, a prominent issue is the presence of unwanted topics. Thien Hai Nguyen and Kiyoaki Shirai (2015) (3) introduced a new topic model to predict stock price movement using sentiments on social media. New Topic Sentiment Latent Dirichlet Allocation (TSLDA) outperformed LDA model with 6.43%. The advantage of the new model is that it can capture the topic and sentiment simultaneously.

In order to visualize topics based on scores Jason Chuang, Christopher D. Manning, and Jeffrey Heer (2012) (4) presented Termite, a visual analysis tool built in R language for assessing topic model quality. They contributed a novel saliency measure for ranking and filtering terms and introduced a seriation method for sorting terms to reveal clustering patterns.

The purpose of this work is to compare the performance of traditional machine learning algorithms, in addition to find important topics using LDA and K-Means.

II. TWITTER SENTIMENT ANALYSIS PROCEDURE

a) Data

The COVID-19 related dataset was scraped through Twitter using snscrape package, passing "coronavirus pandemic" keyword. The data contains three attributes: id, date and 137 thousand of English tweets from March 2020 to October 2020 (Fig. 1). Due to high volume of tweets, parameters were not included during scraping process such as retweets, replies.

	id	date	tweet
0	1311091266243850240	2020-09-29 23:51:31+00:00	#MannKiBaat\nDuring the time of the coronaviru...
1	1311089691609849856	2020-09-29 23:45:16+00:00	Airlines brace for lower Thanksgiving travel i...
2	1311087858367823873	2020-09-29 23:37:58+00:00	WE hope the Coronavirus Pandemic is vanquished...
3	1311085483800363008	2020-09-29 23:28:32+00:00	A reminder, this first debate will cover: Trum...
4	1311084140180189189	2020-09-29 23:23:12+00:00	Apart from the ongoing coronavirus pandemic, w...

Fig. 1: Initial look of data

We first convert date column into appropriate format, so we can visualize the number of tweets for each month (Fig. 2). Interestingly there are more tweets about coronavirus occur in March and decreasing over the months, probably as the outbreak happened more people talked about pandemic, and it became less relevant later. Our text data 93% unique, we keep duplicate observations for further analysis, since small amount of duplicates do not add any bias.

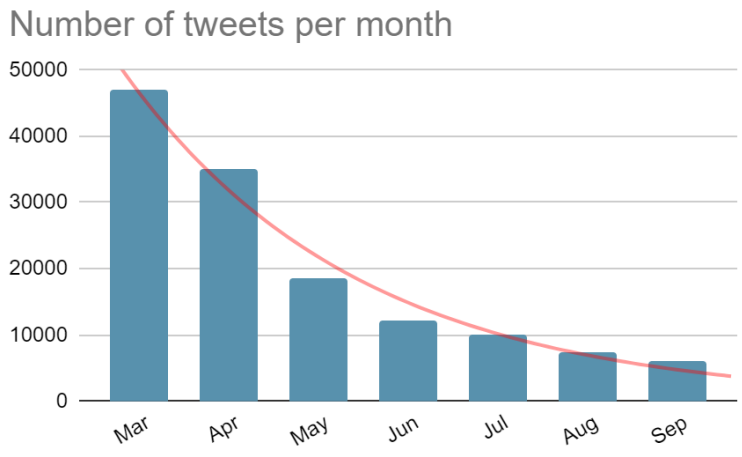

Fig. 2: Number of tweets per month

b) Methodology

i) Calculate polarity level

The data preprocessing can often have a significant impact on the performance of Machine Learning algorithm (5). All words are converted into lowercase to remove difference between words for further processing. The commonly used words such as a, an, the, etc. which do not carry meaning in determining the sentiment of text are removed with "stopwords" extension from nltk library. Furthermore, hashtags URL's and punctuations are removed. Length of text in each observation ranges between 150 and 300 with 30-40 words and most common words in our data is people, trump and world (Fig. 3). After handling abbreviation, we can define polarity level of texts with TextBlob extension. There are more tweets with positive sentiment and less with negative sentiments (Fig. 4).

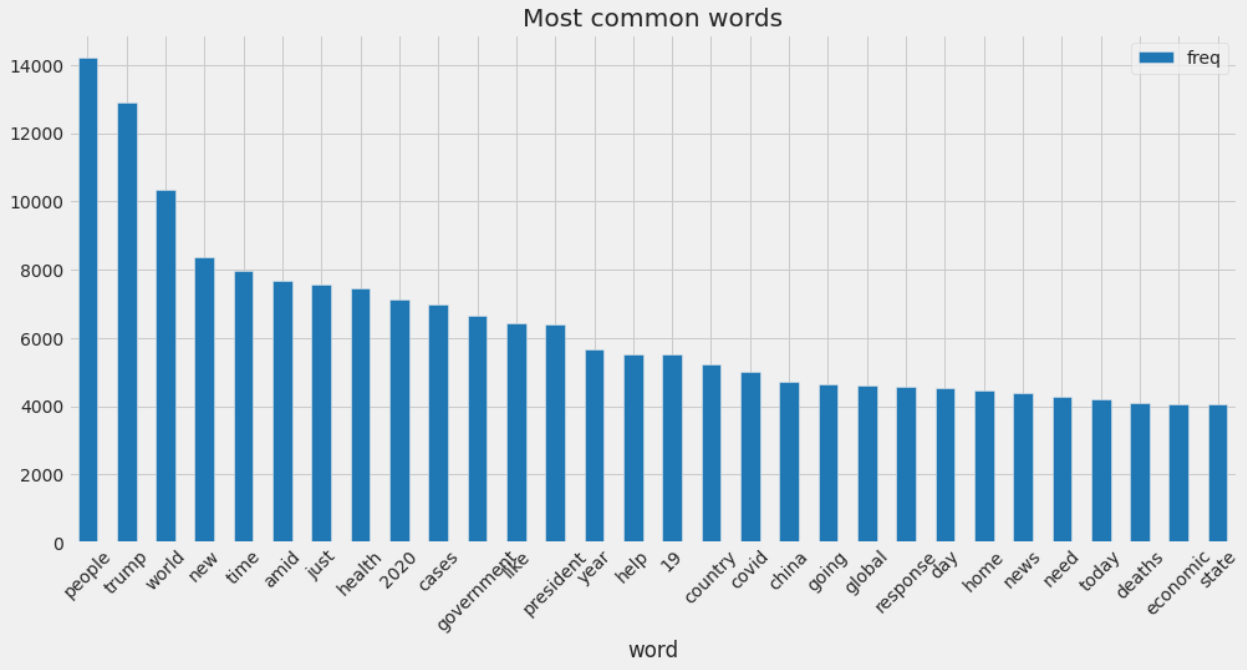

Fig. 3: Most common words using Count vectorizer

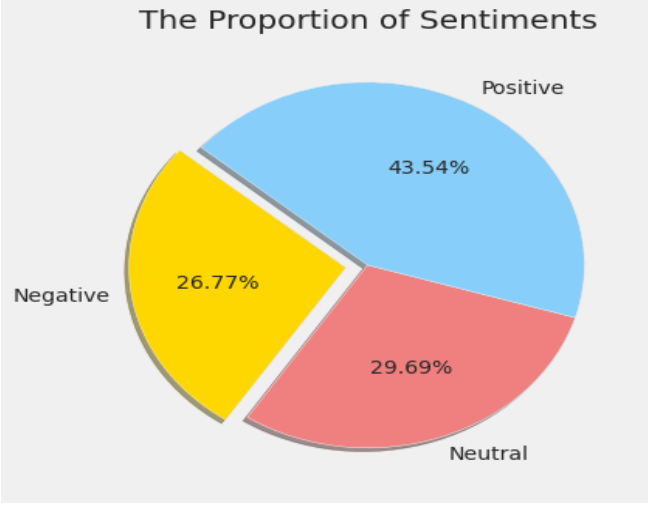

Fig. 4. The proportion of Sentiments

ii) Modeling with supervised learning

To apply any supervised learning models to our data, target variables should be classified based on the text sentiments, and a collection text documents tokenized in building a vocabulary of known words and encoding new documents using existing vocabulary. Also, machines read words in numbers, hence, term frequency-inverse dense frequency applied, which is used to find meaning of sentences consisting of words and cancels out the incapability of Bag of Words technique (6). To put it in more formal mathematical terms, the TF-IDF score for the word *i* in the document *j* from the document set is calculated as follows:

$$tfidf_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right)$$

$$tf_{i,j} = total\ number\ of\ occurences\ of\ i\ in\ j$$

$$df_i = total\ number\ of\ documents\ containing\ i$$

$$N = total\ number\ of\ documents$$

The data then split into training and testing portions with 0.33 test size. 3 different classification models applied to classify sentiment of tweets: Logistic regression, Linear SVC and XGBoost classifier. Precision, recall, accuracy, and f1-scores are calculated to get the results (Tab. 1).

Models	Precision	Recall	Accuracy	F1-Score
Logistic Regression	0.76	0.71	0.74	0.72
Linear SVC	0.82	0.81	0.82	0.81
XGBoost	0.69	0.59	0.63	0.59

Tab.1. – Comparison of final models on test set

iii) Clustering with K-means and LDA

Next, to find topics in our data we performed clustering with K-means using single word terms, and LDA using multi terms. For k-means clustering number of clusters calculated using elbow method, we initialized centroids that would help us select random data points to set them as initial cluster centers (6). To simplify the calculation of the distances pairwise distance from sklearn with Euclidean metric is applied, which returns the distances to each cluster. Two very important functions summarize our model, predict function returns the corresponding predicted cluster label for each data point, and function that fits the algorithm. To visualize our clusters graphically in 2 dimensions, Principal components with 2 components are applied (Fig. 5). Three distinct clusters visualized not far from each other, indicating that there is not much difference in terms of our data, where subjects are coronavirus and pandemic.

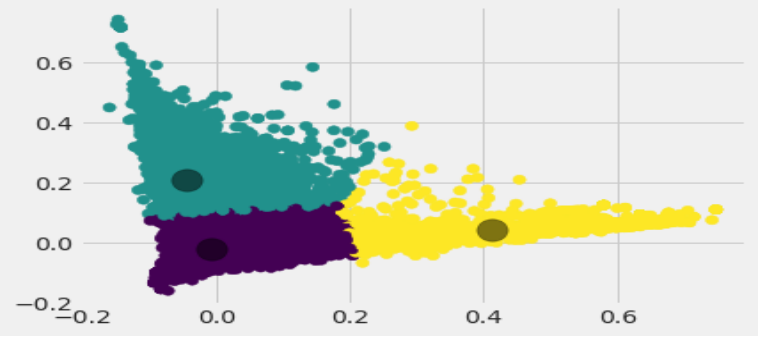

Fig. 5 k-means clustering

Because we are mainly interested in seeing the commonalities between words in each cluster, top words for each cluster are generated and cluster 0 emphasizes news on health and people, cluster 1 is more about political views and cluster 2 is more about news on coronavirus updates (Fig. 6).

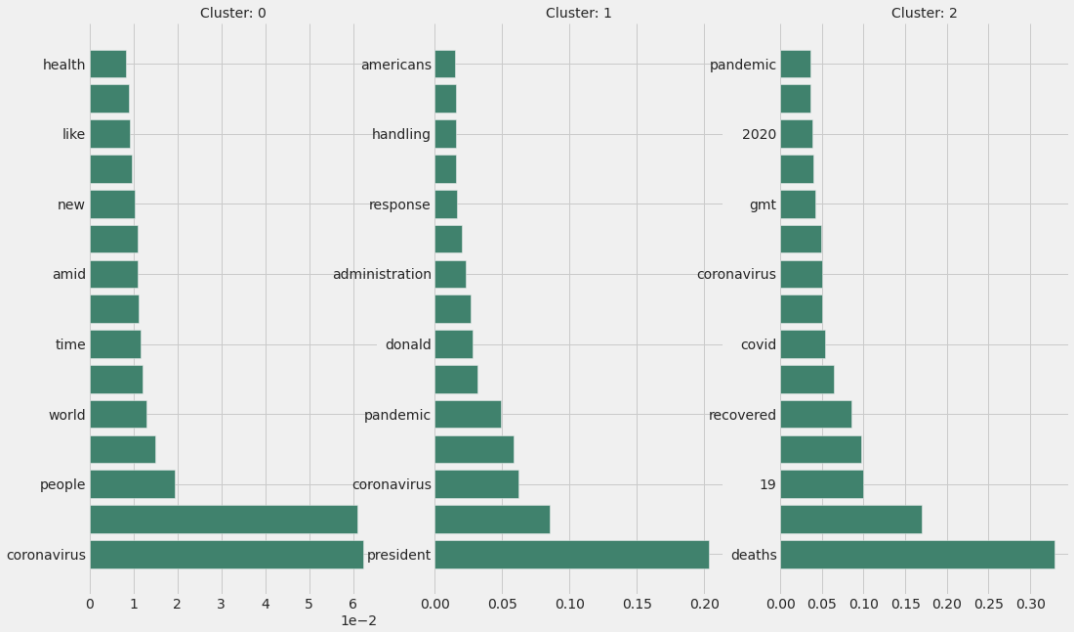

Fig. 6. 3 types of topics

Compared to k-means, LDA in this work was performed to get multi terms with genism library. Several attempts were applied to find right amount of topics, and the best parameters include 3 number of topics, size of document looked at every pass is 500, 10 passes through documents and 400 iterations with phrases that occur 20 times or more, completed in 10 minutes. We are able to define our topics based on the findings, and pyLDAvis is useful to visualize these terms (Fig. 7). The blue bars below show the overall term frequency. Selecting each topic on the right, modifies the bar chart to show the relevant terms for the selected topic. The left panel circles represent different topics and the distance between them. Similar topics appear closer and dissimilar topics farther, since we applied multi terms to our model, we could define that the first topic=0 is about alerts on pandemic, topic1 is about negative feelings and topic=2 is about politics when relevance of lambda is high.

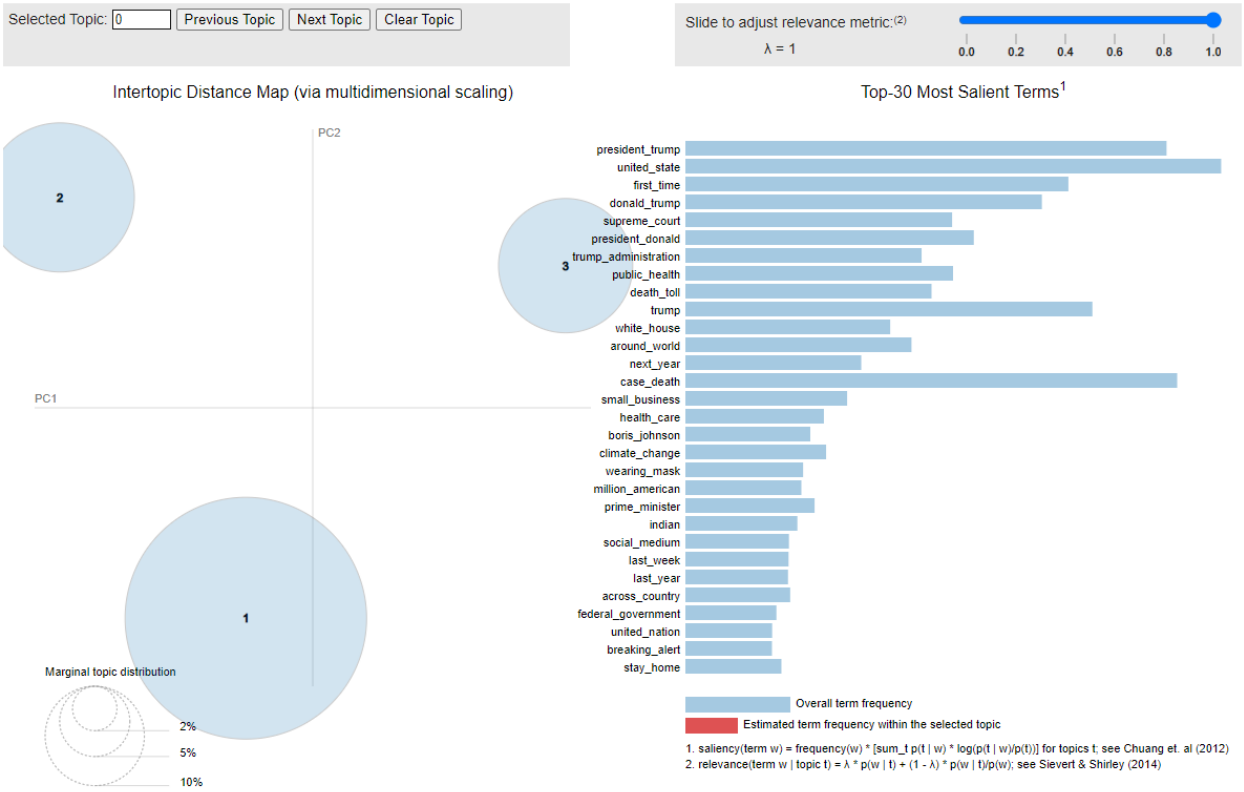

Fig. 7 Visualizing topics from LDA

IV. Results and Conclusion

Sentiment analysis can be performed using lexicon-based approach or machine learning based approach. We identified whether a user-generated text expresses positive, negative, or neutral opinion and in classified text data based on their topics. This paper explains in detail various steps for performing sentiment analysis on twitter data using machine learning algorithms. A machine learning classifier requires a labeled dataset which is divided into train and test set. Once an appropriate dataset is collected, the next step is to perform
preprocessing on data (tweets) by using NLP based techniques, followed by feature extraction method in order to extort sentiment relevant features. Finally, a model is trained using machine learning classifiers like Logistic Regression, Linear SVC and XGBoost and is tested on test data (Tab. 2). The performance of the model can be measured in terms of accuracy, precision, recall and F-score. The results show that Linear SVC performs well showing 82% accuracy on test data. XGBoost achieved the worst result with bad timing on three target variables.

Models	Precision	Recall	Accuracy	F1-Score
Logistic Regression	0.76	0.71	0.74	0.72
Linear SVC	0.82	0.81	0.82	0.81
XGBoost	0.69	0.59	0.63	0.59

Tab.2. – Comparison of final models on test set

Topic modeling was applied with k-means using single word term and LDA, using multi terms. For every model 3 different topics determined. Final model for k-means include 3 topics related to news on health,

political views, and coronavirus updates, while LDA model generated alerts on pandemic, negative feelings, and politics.

Future work aims to combine emoticons and text for sentiment analysis and using hybrid-based approach combining lexicon and machine learning model to identify sentiments with the fact there are huge numbers of tweets generated every second.

REFERENCES

1. Jain AP, Dandannavar P. Application of machine learning techniques to sentiment analysis. In2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) 2016 Jul 21 (pp. 628-632). IEEE.
2. Khoo CS, Johnkhan SB. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. Journal of Information Science. 2018 Aug;44(4):491-511.
3. Nguyen TH, Shirai K. Topic modeling based sentiment analysis on social media for stock market prediction. InProceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) 2015 Jul (pp. 1354-1364).
4. Chuang J, Manning CD, Heer J. Termite: Visualization techniques for assessing textual topic models. In Proceedings of the international working conference on advanced visual interfaces 2012 May 21 (pp. 74-77).
5. Gupta Shashank. Sentiment Analysis: Concept, Analysis and Applications. Medium 2018 Jan 7.
6. Madan Rohit. TF-IDF/Term Frequency Technique: Easiest explanation for Text classification in NLP using Python (Chatbot training on words). Medium 2019 May 30.