Categorization of COVID-19 real-time tweets by sentiment prediction and clustering

Venkata Sai Kusuma Sindhoora Vankayala Siva

Department of Data Analytics, San Jose State University

Data 294: Data Analytics Seminar

Dr. Simon Shim

December 08, 2020

Abstract

The outbreak of novel coronavirus has greatly influenced our daily lives and the entire world is facing an unprecedented health emergency. With the immediate lockdown, social-distancing measures put in place by the government to fight the pandemic, the urge of using social media platforms have exponentially increased. People are constantly looking to connect with others for real-time information, sharing their personal feelings and opinions on platforms like Twitter, Facebook etc. Thus, exploring these evolving tweets can give us insights about the human emotions, sentiments, the mental health of people which can assist the health workers and public information campaigns in framing decisions. The objective of the study is to perform both supervised and unsupervised learning on the COVID-19 twitter data. Along with EDA, the sentiment prediction model is performed to predict people's real tone of voice by labelling the data through a sentiment classifier. Logistic Regression, SVM, Naïve Bayes will be used to evaluate the performance and accuracy of data. And for the unsupervised model, I intend to use both k-means clustering, topic modelling to identify similar clusters or topics in data. The dataset is a collection of global COVID-19 tweets taken from the Kaggle website.

Keywords: Twitter, sentiment prediction, topic modelling, clustering, text mining

Contents

List of Figures

List of Tables

Introduction

Project Goals and Background

The novel coronavirus has already infected sixty-seven million people globally across 215 countries and the death rate has also reached 1.5 million as of December 07, 2020 (Coronavirus Update,2020). Many countries took immediate safety measures such as lockdown, travel ban, implemented stay at home orders, social distance, closure of public places has made employee plans to work remotely etc. to control the widespread of the disease. During this lockdown period, people tend to use social media platforms more such as Twitter, Facebook etc. to know the latest news, communicate with others and share their feelings or thoughts on the internet about the disease. The recent results from the US trend calendar show that Covid-19 has become one of the top trending topics of discussion on twitter since January 2020 and the study shows the discussions have continued to till date. The list of popular hashtags extracted for the period of March 14 to March 30 include #coronavirusupdates, #quarantinelife, #stayathomeorders, #covid-19, #coronaapocolypse (Calendar,2020).These massive real-time tweets on COVID-19 can reveal us valuable insights about human emotions, sentiment dynamics, and topic modelling is important for monitoring the mental health of people especially in this health-crisis situation. Such dynamical analysis is imperative for the government, public health emergency workers for their plan to make the relevant decisions.

Analysis of Requirements

The study is to find answers for (1) Monthly tweet trends since the outbreak. (2) check for outliers and missing values in data (3) identify most common words in the tweets such as top unigrams, bigrams etc. in data (4) identify the popular topics discussed on the twitter (5) How is the sentiment dynamics associated with each topic. (6) insights about fear sentiment prediction of

disease. (7) candidate model that best fits the data. The type of model, data cleaning, feature extraction chosen for training can also be important in non-functional ways. They play a crucial role in the performance evaluation of any machine learning model. Factors such as precision, recall, accuracy, F1 score will be considered to test and evaluate the performance of the algorithms.

Project Deliverables

The project deliverables are a final report written in APA format. It covers the end-to-end analysis of the research starting from data crawling, data preprocessing, model selection, sentiment analysis, topic discovery and experiment. The results of competitive ML models chosen in research will help in predicting the fear sentiment and helps to mitigate the wide spread of disease. This experiment demonstrates the current pandemic and explore the data scientist opportunities for research in this domain.

Technology and Solution Survey

There are different text mining tasks available, one is sentiment classification - used to classify labels as positive, negative, or neutral based on the semantic orientation of the evaluative text. The supervised approach uses the labelled corpora as input and train the different models to predict the underlying sentiment of each text. However, in a practical context, it is hard to find such labelled data in real-time applications. Moreover, the classification trained on one data might not work well on the other data. To tackle this problem, we need a two-step process where the first step is topic modelling i.e. used to detect topics/clusters in the textual data. And the next step is to assign a sentiment label associated for each topic/cluster detected. This approach will make the sentiment polarities intuitively dependent on the topics/clusters rather than performing static sentiment analysis alone.

There are many sentiment analyses tools available to classify the polarity of text data such as Stanford's Core NLP, Textblob, Vader, Senti Strength etc. In this paper, I intend to use VADER to classify the sentiment score of each tweet as it does not require much preprocessing and training (Hutto C.J et al., 2014).

Literature Survey of Existing Research

The objective of the study is to detect the topics, sentiments, concerns from the tweets posted by the twitter users on the corona virus disease. Since the COVID-19 outbreak, a lot of research work was conducted on the social media data to understand the public responses and discussions about the pandemic. For instance, P.D. Turney et al. (2020) did most of the work on topic detection where the complete study focus on discovering the topics from a hundred-billion-word corpus, but however they couldn't perform sentiment analysis on the text thus limiting the usefulness of mining results. Jim Samuel et al. (2020) exploited the twitter tweets concerning COVID-19 and found that the most influential tweets are still written by regular users, such as news media, individual reporters, and government officials. Zhou et al. (2020) examined the "tweets concerning covid-19 on Twitter and detected the sentiment dynamics of people living in the state of New South Wales (NSW) in Australia during the pandemic period". The summary of the research found that the overall sentiment polarity of the state was positive.The current related work either performed supervised sentiment analysis or unsupervised topic-modelling alone or failed to perform both. Hence the proposed model will be a composed approach to detect the topic-level sentiment to understand the public emotions and mental health of people for framing relevant decisions.

Data Exploration

Data Exploration Strategy and Planning

Twitter data can be extracted from multiple sources. Some published papers collected the data from Twitter API by using Python and Tweepy scripts. Some used proprietary data that may require special license keys to access the content of the application. Others have used data from public sources, data scraped from automated web-scraping tools that are freely available on the internet. For this research, I will exploit global COVID-19 tweets dataset from the Kaggle website since it is well-vetted and publicly available (Kaggle, 2020).

Data Sources and Dataset Parameters

From July 24 to August 30, a random sample of 1.7L tweets in the English language are collected from Twitter API using python and Tweepy code. The high-frequency hashtags used to query relevant tweets are #covid19, #coronavirus, #coronaoutbreak, #pandemic, #coronavirusupdates and the retweet argument has been set to 'False' to exclude all the retweets in the dataset. The metadata from tweets include 13 variables - full text of the tweet, timestamp, user location, user demographic information, the number of favorites, friends and followers, list of hashtags, source, is retweet.

Collection of Training datasets

The initial exploratory data analysis was conducted on the dataset by using descriptive statistics. The columns such as user location, user description, hashtags, source have missing values in the data. Since the research focus on the sentiment analysis of the tweets, I have replaced the missing values as unknown for now and continued for further analysis. There are no duplicate records identified in the data. Figure 1 shows the plot of uniqueness check performed on the dataset; it was found that the text column has the highest unique values i.e. 178683 observed from the overall

179108 counts and is_retweet has lowest i.e. False. The Table 1 shows the summary of tweets size used per day to conduct the experiment and Figure 2 shows the trends chart for the tweets collected per day.

Figure 1

Plot of unique features in the dataset

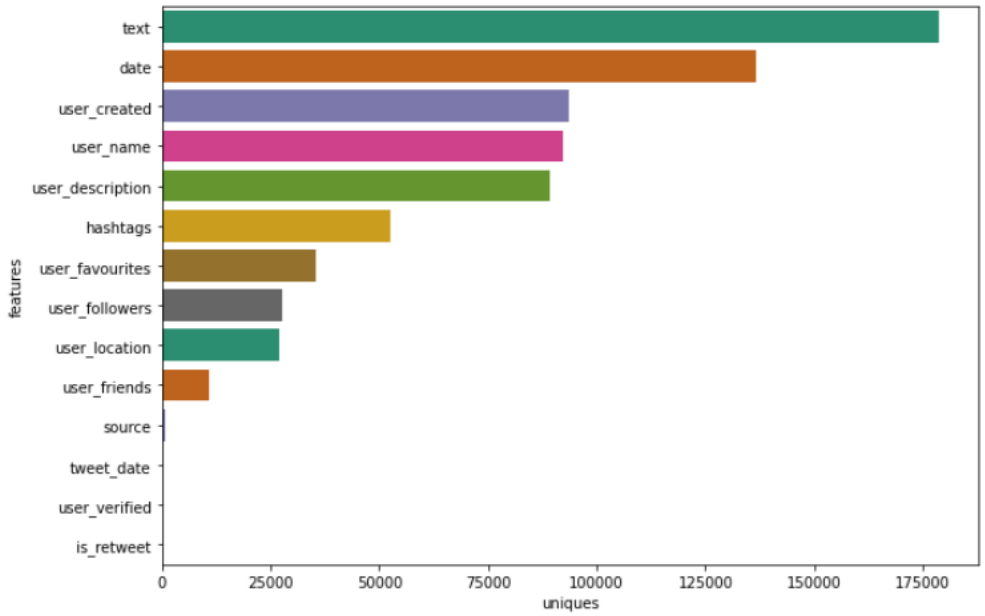

Figure 1. Plot of unique features 1

Table 1

Summary of tweets used for the period of July 24 – August 20

Date	7/24/2020	7/25/2020	7/26/2020	7/27/2020	7/28/2020	7/29/2020	7/30/2020	7/31/2020
Total	295	16881	7500	7500	7500	2780	1980	7500
Date	**8/1/2020**	**8/2/2020**	**8/4/2020**	**8/6/2020**	**8/7/2020**	**8/8/2020**	**8/9/2020**	**8/10/2020**
Total	7500	7500	7500	7214	1060	7500	7500	4891
Date	**8/11/2020**	**8/12/2020**	**8/13/2020**	**8/14/2020**	**8/16/2020**	**8/17/2020**	**8/18/2020**	**8/22/2020**
Total	7500	7500	7500	7500	7500	7500	7500	11555
Date	**8/29/2020**	**8/30/2020**						
Total	4077	8375						

Table 1. Summary of tweets 1

Figure 2

Trends chart of daily tweets per day

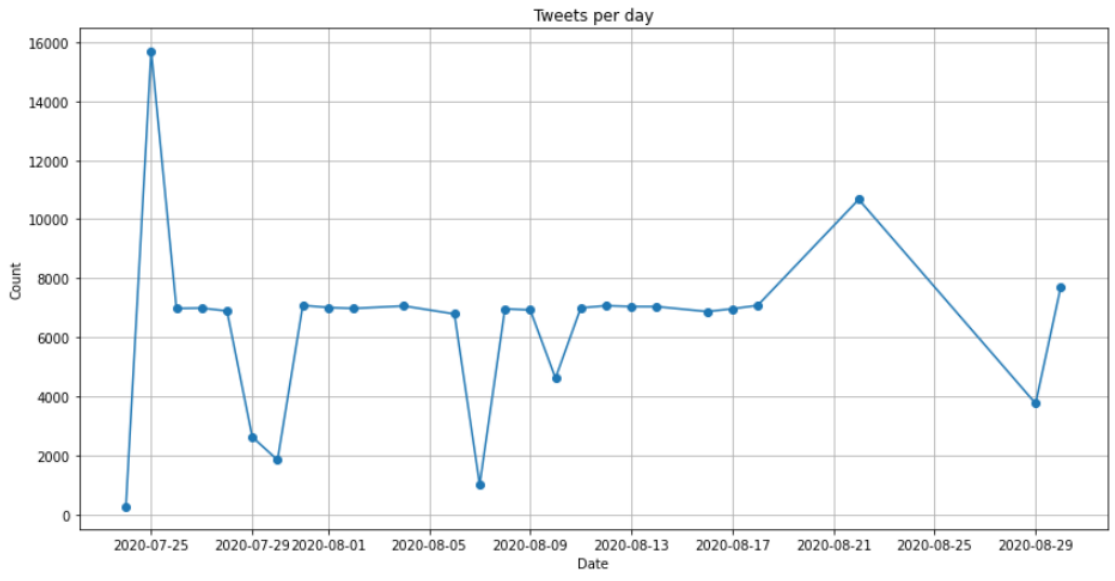

Figure 2. Trends chart of tweets 1

Data Cleansing and Validation

The tweets were cleaned with the help of regular expressions by removing tags, mentions, hashtags, URLs, special characters, punctuations, newline, numbers, and other gibberish content. All the characters in the tweets are turned lowercase and removed the stop words from the data using NLP library. Then, performed snowball stemmer over porter stemming, is that snowball is regarded as an improvement over porter, also it is slightly fast with contributions from the larger community. The length of each text is calculated from cleaned tweets and added to the data frame. An outliers check was performed based on text length and removed all the extremely shorter or longer tweets in the data. Finally, a total of 143396 tweets are retained for further analysis.

There are two sets of training and testing data are prepared to perform both supervised and unsupervised learning on the data. As it is hard to find labelled data in real-time, I have manually created sentiment labels to train the supervised learning model. I used Vader sentiment analyzer, Textblob libraries to compare the polarity scores of the data. The result shows Vader sentiment on cleaned tweets classified 105940 as positive and 37456 as negative tweets. Similarly, Textblob library classified 123061 tweets as positive and 20335 as negative. Since the Vader sentiment classified more negative labels than Textblob, I have considered Vader score and added it to the data frame. K-means, LDA algorithm will be used on unlabeled data for second training, test set to perform the unsupervised approach.

Data Transformation and Tools

The advantage of textual analytics is not only used to identify the popular key words in the tweets but also, we can identify the popular word groups, word chains in the data. I have used N-gram representation to generate feature vectors in the data. The words in the tweets are transformed to generate a matrix of individual words, bi-grams, trigrams etc. Term Frequency and Inter

Document Frequency (TF-IDF) technique were used to normalize the text data to numeric values. The categorical data of sentiment label was converted into sets of 1/0 values through Label Encoder library of sci-kit learn.

Raw Data Visualization

Word Cloud in python is the data visualization technique used to represent the important or frequent words in the dataset. Figure 3 shows the word cloud of most frequent words identified in the text corpus. We can see that words such as “case”, “people”, “new”, “pandemic” etc. with larger font size has a higher frequency in the tweets. Figure 4, Figure 5, Figure 6 are the visualization charts of most frequent unigrams, bigrams and trigrams identified in the dataset.

Figure 3

Word Cloud visualization of frequent key words used in tweets

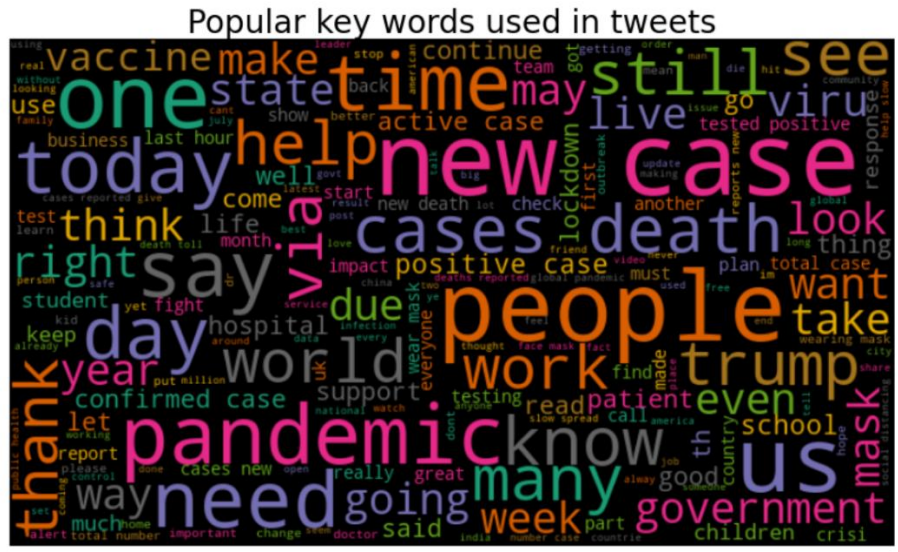

Figure 3. Wordcloud of popular tweets 1

Note. “case”, “people”, “new”, “pandemic” have appeared frequently

Figure 4

Popular Unigrams identified in tweets

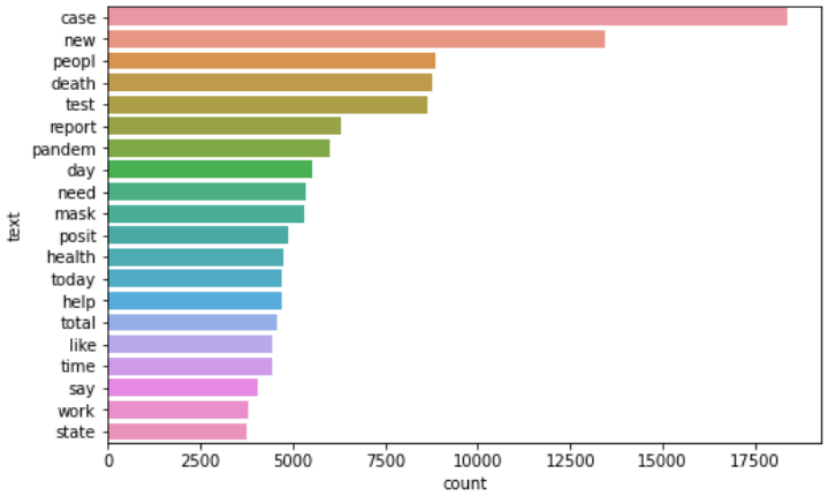

Figure 4. Popular Unigrams 1

Note. "case", "new", "peopl", "death" are popular unigrams

Figure 5

Popular Bigrams identified in tweets

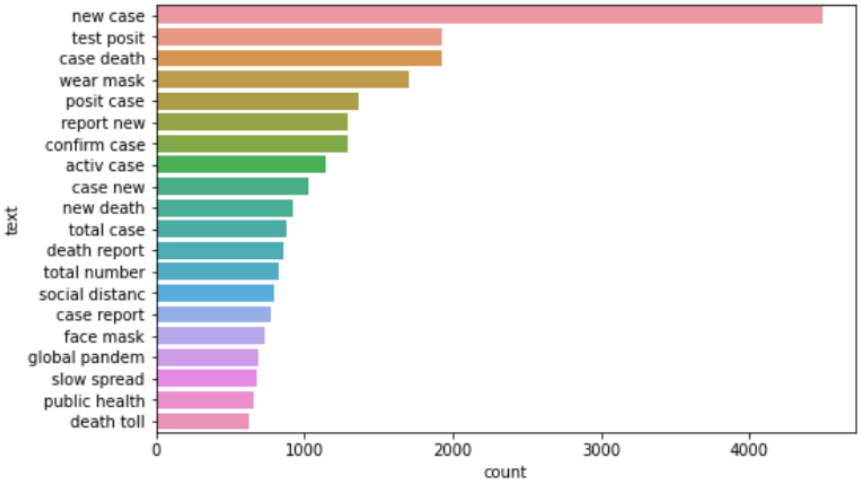

Figure 5. Popular Bigrams 1

Note. “new case”, “test posit”, “case death”, “wear mask” are popular bigrams

Figure 6

Popular Trigrams identified in tweets

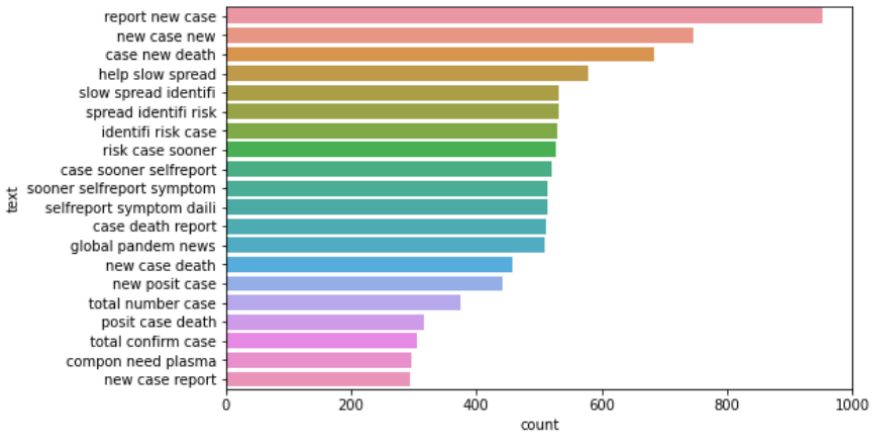

Figure 6. Popular Trigrams 1

Note. “report new case”, “new case new”, “case new death”, “help slow spread” are popular trigrams

Data and Project Management

Project Organization

The best practice to manage the data in a structured way is to use standard frameworks. “CRISP-DM is a Cross-Industry Standard Process for Data Mining- an open standard process model that describes common approaches used by data mining experts (Manasson, A, 2020).” The lifecycle of data management in machine learning involves six phases.

1. Business Understanding
2. Data Understanding
3. Data Preparation

4. Modeling
5. Evaluation
6. Deployment

Once the data collection is done, the next important step is understanding the data. This is called data exploration. Different statistical methods can be performed on the data to check the quality, data types, data redundancy in this process. The patterns in the data can be explored through visualization of data distributions, identifying relations among attributes. The next step is data preparation, an important step that needs to be performed before feeding to any machine learning algorithm. Firstly, all the categorical data must be converted to machine-understandable numerical data. The data cleaning is performed to impute or remove the missing values, duplicates, and uniqueness check is done. Feature engineering, in general, more crucial than building models is performed to select the important columns in data. Techniques such as PCA, SVD can be used for data reduction and normalization techniques can be applied for data scaling. Modelling responsibilities include to try various models and ensemble them, problem and implementation efficiency, parameters tuning etc. In the model evaluation phase, different metrics are used to measure the performance and if the test results are not satisfied, need to reconsider the previous steps. Finally, deployment is not a part of this research project.

Resource Requirements

The resource requirements for the project include data requirements, software platforms, software licenses, hardware systems, etc.

Data requirements:

Extract covid-19 tweets from Twitter API

Software requirements:

- Python 3.8
- NLTK 3.5
- Pandas 1.1.3
- NumPy 1.19.1
- SciPy
- Seaborn
- Matplotlib
- Scikit-learn
- Genism
- Spacy

Software Licenses: Twitter Developer Labs, python Tweepy

Hardware systems: Intel i-7, 16 GB, GPU

Project Development Methodology

According to the author Jordan "Machine learning projects are highly iterative; as you progress through the ML lifecycle, you'll find yourself iterating on a section until reaching a satisfactory level of performance, then proceeding forward to the next task (which may be circling back to an even earlier step).These steps are sufficiently benign and does provide some structure to data (Jordan, J, 2020)." However, this is best when we use iterative Agile development that focuses on interactivity, engagement, and constant feedback solutions at every phase. This helps the model to improve its learning and thus making the process more adaptive and dynamic. Figure 7 shows the end to end flow of the proposed framework.

Figure 7

Proposed framework

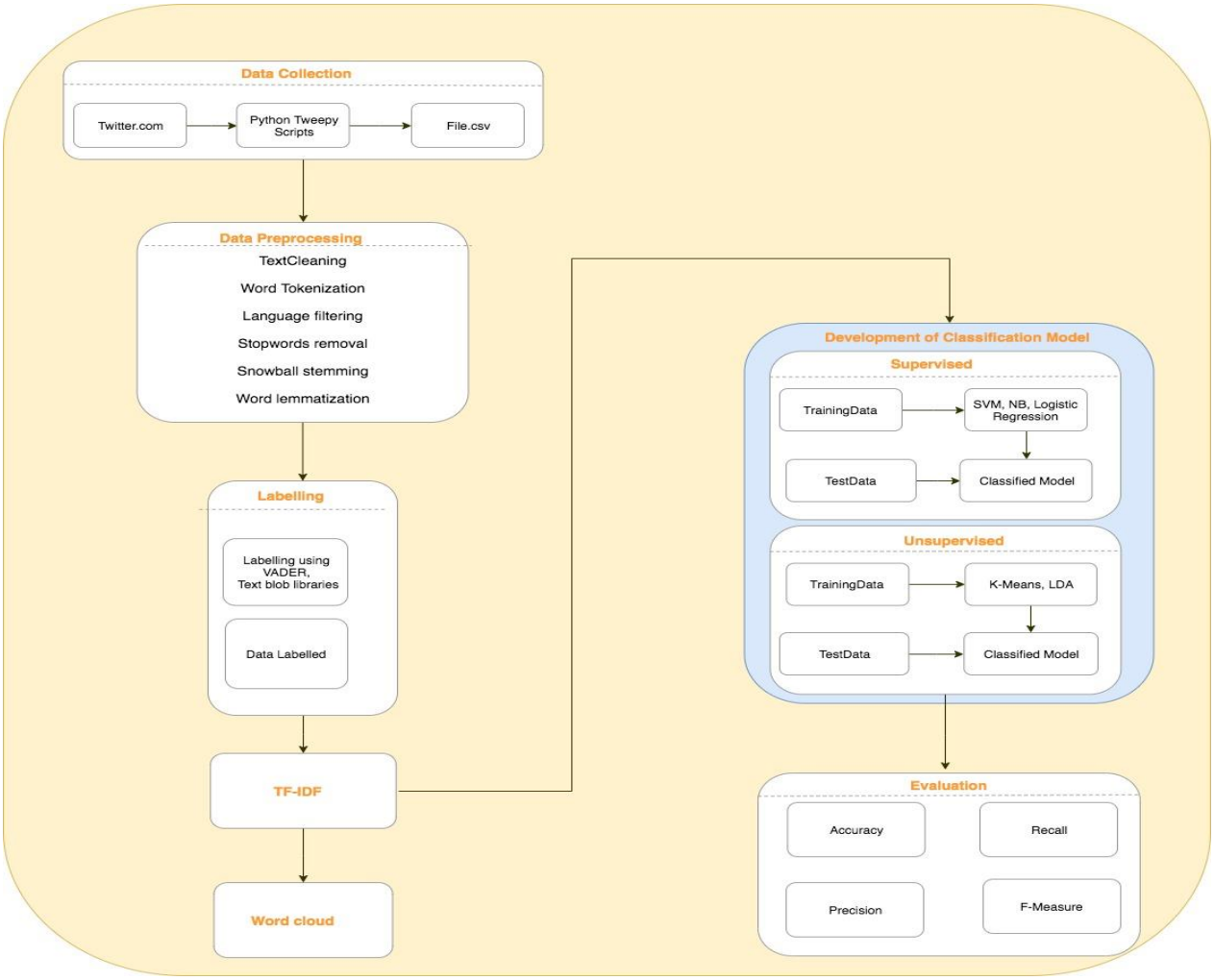

Figure 7. Proposed Framework 1

Project Schedule

Project scheduling is the mechanism to track all the tasks that need to be done, which resources must be utilized, estimated time to complete the sub-tasks, when is the deadline for the final project submission. These sub-tasks are time-bound, and each task has a start and end date, so it can be completed on time. The project planning is important for this research project as there

are many phases in the project life cycle. To organize the project, I have used an online Gantt charting tool to create the tasks, milestones, dependencies etc. The Figure below shows the Gantt chart of project organization built on adapting the CRISP-DM framework blended with iterative Agile development process.

Figure 8

Gantt chart showing Agile development process

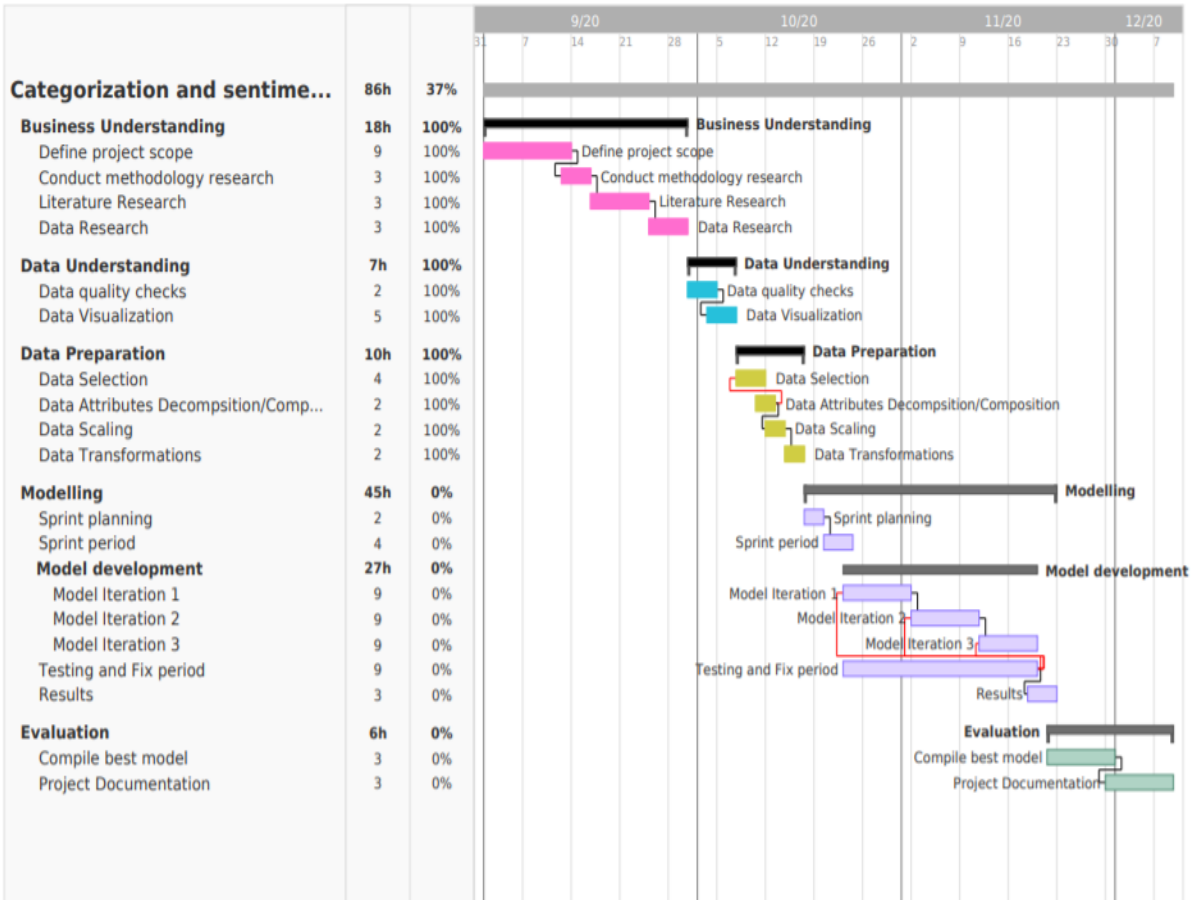

Figure 8. Gantt chart for project manage 1

Note. Gantt chart of project organization built on adapting the CRISP-DM framework blended with iterative Agile development process

Problem Formulation and Model Selection

Problem formulation

The objective of the study is to perform sentiment analysis on COVID-19 tweets. This problem can be solved by composed approach of both supervised and unsupervised learning, that requires a classifier which takes input as text data, and output is to distinguish between positive, negative sentiments. The goal of the study is to compare both approaches and identify the best fit algorithm for the model.

Foundation of Proposed Solutions

Naive Bayes (NB)

Naïve Bayes classifier is simple and effective method that can be used in text analytics of supervised model (Liu, B et al.,2013). It is a probabilistic classifier based on Bayes theorem with strong independence assumptions between the predictor variables. According to theorem, the conditional probability can be calculated based on the below formula.

$$\mathbb{M}(\mathbf{q}) = \underset{l \in levels(t)}{\arg\max} \left(\left(\prod_{i=1}^{m} P(\mathbf{q}[i] \mid t = l) \right) \times P(t = l) \right)$$

There are three types of classifiers in Naive Bayes: Multinomial NB, Gaussian NB, Bernoulli NB. A comparative study among all classifiers shows that MNB works better for document classification than other classifiers (Pulung et al.,2020). The advantage of Naïve Bayes is it works well with real-time data and it can effectively handle dimensionality reduction (Jim Samuel et al., 2020).

Support Vector Machines (SVM)

SVM is a fast, dependable supervised machine learning model that can be used to classify both linear and non-linear data. The classifier is used to identify good linear separators that works

better when we have different classes. If the data is not linearly separable, then we can use kernel tricks of mapping our space to a higher dimension. According to Joachims T., " text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories (Joachims T, 1997)."

Logistic Regression (LR)

Logistic regression is a probabilistic classification method that can be used to predict binary outcome. "In LR, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function (Pedregosa, F et al., 2011)". The algorithm uses sigmoid function to calculate the estimated class. As said by Samuel "the main aim of the objective function is to minimize the errors in training data using cross-entropy loss function. The stochastic gradient descent optimization algorithm is widely used to optimize the objective function (Jim Samuel et al., 2020)". Below is the function form of the algorithm.

$$P(t) = \frac{\exp(\alpha + \beta t)}{1 + \exp(\alpha + \beta t)}$$

K-means Clustering

K-means clustering is a popular algorithm for data clustering, and it takes input as number of clusters(k), random data points will be initialized as cluster centroids. It outputs the coordinates of calculated cluster centroids based on Euclidean distance between the data to each centroid point. The formula to calculate Euclidean distance is shown below. The advantage of k-means is scalable to large datasets and are simple to implement (Muhammad Ihsan Zul et al., 2018).

$$D(i,j) = \sqrt{(xi1 - xj1)2 + (xi2 - xj2)2 + \cdots + (xip - xjp)2}$$

Where,

D(i,j)= distance of i^{th} data to cluster center j

X_{ik} = i^{th} on the k^{th} data attribute

X_{ij} = j^{th} center point on the k^{th} data attribute

Topic modelling

It is the process of learning, extracting the hidden patterns across a corpus of unstructured text. A popular method for this is Latent Dirichlet Allocation (LDA) is first proposed by Blei which is used to detects topics in an unstructured text (Blei D.M. et al.,2006). The priori input of number of topics should be provided and output will be the mixture models i.e. each tweet or document will be assigned to more than one cluster.

Feature Engineering

Feature engineering is the process to select the most important features that can be used in our training data before feeding to a machine learning algorithm. In general, the raw data is transformed to numerical i.e. machine understandable format.TF-IDF (Term Frequency-Inverse Document Frequency) was used to convert words into vectors. The feature vectors have terms and its frequencies. TF is used to measure how frequent a word appearing in the document whereas IDF is used to measure how important a term is. This is called TF-IDF weighting and formula is shown below where D= Number of documents in the corpus, dfx: Number of documents where (t) appears in document d.

$$W(x,y) = tf\ (x,y) * \log\ (D/dfx)$$

The most popular words must not be considered in general because they cannot be used as key differentiator for sentiment analysis. Hence, the term inverse document frequency is used in log part of the formula.

Model and Solution Comparison, Justification, and Discussion

Three models SVM, Naïve Bayes, Logistic Regression can be considered as candidates for supervised model of sentiment analysis. The results of the comparative study by Pulung Hendro Prastyo1 shows that "SVM algorithm with the Normalized Poly Kernel is the best algorithm in predicting sentiments and outperformed MNB in all test models. The SVM provided the highest accuracy in sentiments with an average accuracy of 82.00%, the precision of 82.24%, the recall of 82.01%, and the f-measure of 81.84%. Therefore, SVM can be used as an intelligent algorithm to conduct a sentiment analysis for new data (Pulung et al.,2020)".

Another study conducted by Jim Samuel experiments how different machine learning models perform when we have tweets of varying lengths. The results show that "we observe a strong classification accuracy of 91% for short Tweets, with the Naïve Bayes method. We also observe that the logistic regression classification method provides a reasonable accuracy of 74% with shorter Tweets, and both methods showed relatively weaker performance for longer Tweets (Jim Samuel et al., 2020)".

A study conducted by Muhammad Ihsan Zul shows "the accuracy of the system is tested 10 times at k different points for each k value (k=6, 7, 8, 9 and 10). As the result, the combination of K-Means and Naïve Bayes has lower accuracy than the accuracy produced by Naïve Bayes without the combination of K-Means. The accuracy of Naïve Bayes algorithm is from 80.526%-82.500%, while the combination of Naïve Bayes and KMeans has 80.323%-81.523% accuracy. The accuracy that was generated by combination of K-Means and Naïve Bayes was not better than the accuracy

of using Naïve Bayes without combination of K-Means. So, Naïve Bayes without the addition of K-Means is more effective than the combination of K-Means and Naïve Bayes (Muhammad Ihsan Zul et al., 2018)".

References

B. Liu., E. Blasch., Y. Chen., D. Shen & G. Chen. (2013) "Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier," 2013 IEEE International Conference on Big Data, Silicon Valley, CA, 2013, pp. 99-104, doi: 10.1109/BigData.2013.6691740.

Blei, D.M. & Laerty, J.D. (2006). Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. pp. 113-120

Calendar, T. (2020, March 15). *Trending words on 14th March 2020*. Trend Calendar. https://us.trend-calendar.com/trend/2020-03-14.html

Coronavirus Update (Live): 68,561,815 Cases and 1,562,895 Deaths from COVID-19 Virus Pandemic - Worldometer. (n.d.). Corona Virus Update (Live). Retrieved December 7, 2020, from https://www.worldometers.info/coronavirus/

COVID19 Tweets. (2020, August 30). Kaggle. https://www.kaggle.com/gpreda/covid19-tweets

Huang, B., & Carley, K.M. (2020). Disinformation and misinformation on twitter during the novel coronavirus outbreak. arXiv preprint arXiv:2006.04278

Hutto, C.J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media

Ihsan Zul, Muhammad & Yulia, Feoni & Nurmalasari, Dini. (2018). Social Media Sentiment Analysis Using K-Means and Naïve Bayes Algorithm. 10.1109/ICon-EEI.2018.8784326.

Joachims, Thorsten. (1997). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.143-151.

Jordan, J. (2020, December 1). *Organizing machine learning projects: project management guidelines.* Jeremy Jordan. https://www.jeremyjordan.me/ml-projects-guide/

Manasson, A. (2020, February 25). *Why using CRISP-DM will make you a better Data Scientist*. Medium. https://towardsdatascience.com/why-using-crisp-dm-will-make-you-a-better-data-scientist-66efe5b72686

P. D. Turney & M. L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. CoRR, cs.LG/0212012, 2002.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Project Schedule: The Ultimate Guide (Example Included). (2020, December 3). ProjectManager.Com. https://www.projectmanager.com/project-scheduling

Pulung, H. P., Amin, S. S., Ade, W. D., & Adhistya, E. P. (2020, October 28). Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel. *Journal of Information Systems Engineering and Business Intelligence*. Http://E-Journal.Unair.Ac.Id/Index.Php/JISEBI.

Samuel, J., Ali, G. G. M. N., Rahman, M. M., Esawi, E., & Samuel, Y. (2020). COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. *Information*, *11*(6), 314. https://doi.org/10.3390/info11060314

Zhou, J., Yang, S., Xiao, C. & Chen, F. (2020): Examination of community sentiment dynamics due to covid-19 pandemic: a case study from australia. CoRR abs/2006.12185, https://arxiv.org/abs/2006.12185