A Social Media Sentiment Analysis: Machine Intelligence Model For Worldwide Covid-19 Vaccination Using Twitter Data

Kalyan Kumar Jena
Parala Maharaja Engineering College (Govt.), Berhampur
Sourav Kumar Bhoi
Parala Maharaja Engineering College (Govt.), Berhampur
Satyajeet Behera
Parala Maharaja Engineering College (Govt.), Berhampur
Raghvendra Kumar
GIET University
Hoang Viet Long
Institute for Computational Science, Ton Duc Thang University
Nguyen Thi Kim Son (nguyenthikimson@tdtu.edu.vn)
Institute for Computational Science, Ton Duc Thang University

Research Article

Keywords: Sentiment Analysis, Interactive Tag Cloud, KNIME Platform, Covid-19 vaccination.

Posted Date: November 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-734770/v1

Abstract

Understanding human emotions is one of the crucial aspects when we are to take action. Our emotions dictate our apparent behaviors. In simple words, what we feel inside can predict things about what we would do. This creates a huge opportunity for government and businesses industry to understand and predict people's behaviors. There has been some really great research done on this with high accuracy. Recently, Covid-19 vaccination process is a challenging task going on all over the world and it is necessary to explore people's reaction over this for more effective vaccination process spread. In this paper, wetried to understand an event (Covid-19 vaccination) with a relatively simple model with decent accuracy compared to other sophisticated models. We use simple machine learning models to train and deploy it over the network. We have used KNIME Analytical Platform to design and implement our model as it provides end-to-end analytics. We have managed to get 88.67% accuracy and Cohen's kappa 0.789 with SVM model by tuning some parameters. The model is deployed on Twitter data. This paper shows our efforts trying to make a simple model to analyze an event (Covid-19 vaccination) and understand people's emotions towards the event. The methodology involves identifying important topics (terms) and finding out the sentiment (positive, negative, neutral). This paper tries to find a low-cost solution to analyze an event and provide data-driven insights from it without involving sophisticated algorithms.

1. Introduction

Human beings are one of the complex organisms in the known universe. Our emotions play a vital role in our decision making [1–16]. Keeping the external environment same, an angry person would act differently to a specific event than a happy person. That's why it's so crucial to understand people's emotions with higher accuracy. Let's assume we're a govt and we're trying to implement a new policy. To determine the effectiveness of the said policy, we'd have to implement it. But the results of that implementation could be catastrophic. It may lead to protests, hunger strikes. People might choose violence to revoke the said policy. So, we have to predict how people will react to something. And we can do that if we know the current emotion of the people. If we know that the people are angry, then we'd implement the said policy a little differently than if people were happy. Our current state can be an effective tool to predict certain behaviors. For a business, it could be a multi-million dollars of investment in a product. Those, who can predict that with higher accuracy, are at a state of significant profit. To capture people's sentiments about social events, political movements, new marketing campaigns gives a slight edge to those who can predict people's emotions and take decisions accordingly.

Social media provides a platform to people from different background and cultures to express their opinions about specific interests, raise awareness in the public, share facts and everything. And it's not just the youngsters who benefit from this advancement of a tight-knit digital world, old people participate in this too. They oppose their concerns about things, provide diverse opinions on things. More than 500 million people use social media on a regular basis. Analyzing people's emotion through social media is an effective way to understand it.

In this reality, people prefer to spend more time on their phones instead of their real lives. It gives them a voice and empowers them to pose their opinions, concerns, awareness or just simple fact sharing. Our lives have more interconnected than ever. We can reach someone at the other end of the world in literary seconds. This has created an opportunity for the govt, businesses, or any interested entity to collect people's emotions and make informed data-driven decisions. Uninformed decisions can lead to catastrophic consequences. But informed decisions can predict those severe consequences to a degree and we can make alternate plans to minimize the damages. To make an informed decision, we need understand those people's emotions. And this paper is about finding a quick overview of an event to make those informed decisions.

In this paper the major contributions are:

1. A model is proposed to get a quick overview of an event (Covid-19 vaccination) through interactive cloud to see the important terminologies and sentiment analysis to get the overall sentiment of the population.
2. We use simple machine learning models to train and deploy it over the network. The model is deployed on Twitter data. The methodology involves identifying important topics (terms) and finding out the sentiment (positive, negative, neutral).
3. We have used KNIME Analytical Platform [17] to design and implement our model as it provides end-to-end analytics.
4. This model gives an accuracy of 88.67% with SVM model when tested on 3K tweets and then deployed it over 150K tweets with 3 search filters.

The rest of the sections are discussed as follows. Section 2 describes the related work. Section 3 presents the problem statement and objective. Section 4 describes about the sentiment analysis model. Section 5 describes about the conclusion of the proposed work.

2. Related Works

Many research works have been done in the area of sentiment analysis [1–16]. Our paper focuses on two things; keywords (Interactive Tag Cloud) and Sentiment Analysis(Positive, Negative and Neutral). But before that, let’s dive into why it’s so important in the first place to do research anyway. To answer that, David P. Fan did a research on the topic “Sentiments from press” in 1993 to show a co-relation between the news produced at that time and the consumer behavior after that. Theyused NEXIS database to do an extensive research on media analysis for prediction [1].

The first part of our deployment (Interactive Tag Cloud) is immensely inspired by a research done by Minqing Hu and Bing Liu on “Mining Opinion Features in Customer Reviews” published in 2004 [2]. Their task was simple which involvedthe identification of the features of the products that the customers had opinions expressed on (opinion data feature) and rank those features according to their frequencies according to occurrence in the reviews. For each feature, they identified the number of customer reviews having positive or negative opinions.

Now the question arises about the efficacy of Twitter Social Media Platform and this was successfully demonstrated by Xue Zhang, Hauke Fuehres, and Peter A. Gloor [3] with their study on “Predicting Stock Market Indicators Through Twitter” published in 2011. Their paper describes the early works trying to predict stock market indicators like Dow Jones, NASDAQ and S&P 500 which was done by analyzing Twitter posts. Their approach involved collection of Twitter data over a period of six months and getting a randomized subsample which was 1% of the full volume of all the tweets. Many researchers started doing opinion mining after Twitter became the prominent hub for everyone to share their diverse opinions and feelings. Even county leaders started to rely on this platform to share important information, confront opposition. This raised the bar of Twitter as a social media platform into a must-have. Today over 500 million people use this incredible platform. A study done by Soo-Min Kim and Eduard Hovy [4] on “Determining the Sentiment of Opinions” which was published in 2004 describes a solution to the classification of sentiments problem by creating a Word Sentiment Classifier which entails named entity tagger, holder finder, pos tagger, and a wordnet.

During the next few years people started creating many rule-based filters to successfully automate the workflow of sentiment classification. Xiaowen Ding, Bing Liu, and Philip S. Yu [7] did a research on “A holistic lexicon-based approach to opinion mining” in 2008 which primarily emphases on customer reviews of products and the problem of determining the semantic orientations associated with them such as positive or negative or neutral. Another study based on Machine Learning for sentiment classification was done by S. Arafin Mahtab, N. Islam and M. Mahfuzur Rahaman [10] on "Sentiment Analysis on Bangladesh Cricket with Support Vector Machine" in 2018 shows 64.59% average accuracy. Tho, Cuk, Harco Leslie Hendric Spits Warnars, Benfano Soewito, and Ford Lumban Gaol [13] did a study on this topic "Code-Mixed Sentiment Analysis Using Machine Learning Approach–A Systematic Literature Review" in 2020 which is focused on studying the approaches used in code-mixed sentiments analysis.

Twitter is riddled with people using misspelled words to express their emotions. Using “Sad” conveys less emotional connection than “Saaaaaaad”. The later word means to people apparently. Apart from that, when calculating sentiment, tons of data are given which holds little-to-no contextual meaning. That’s why it’s recommended to pre-process the data according to the task at hand. Yanwei Bao, Changqin Quan, Lijuan Wang and Fuji Ren [14] did a research on “The Role of Pre-processing in Twitter Sentiment Analysis” on 2014 and achieved a classification accuracy of 85.5%. Also, a study done by Saif Hassan, Fernández Miriam, He Yulan and Alani Harith [15] provided a comprehensive overview of “On Stop-words, Filtering and Data Sparsity for Sentiment Analysis of Twitter” in 2014. Many such works are done by the researchers in [16–27]. Above research works are based on sentiment analysis using different methods on social media data in different domains. However, as per our knowledge very less works has been done in the are of Covid-19 vaccination sentiment analysis. This will be a better area to explore; therefore we have proposed a model to using machine learning approach to analyze the sentiments of the people at the time of this global pandemic.

3. Problem Statement And Objective

Many researchers have been done on Twitter Sentiment Analysis involving hybrid rule-based and complex machine learning models. But our primary goal is to build a model that can analyze an event (Covid-19 vaccination) and provide a quick overview of that even in a simplest form. Our model doesn't involve any rule-based training. Nor does it have complex machine learning models. To analyze the event, we have taken a two-step approach; an Interactive Tag Cloud providing important terminologies and keywords repeated in a Tweet database and finding out the overall sentiment. We have KNIME Analytical Platform to design and implement our model as it provides end-to-end analytics. This platform does not require coding everything, instead it has built-in nodes which handles our tasks. It's a GUI based application.

4. Methodology

Natural Language Processing (NLP) is a task where we make the computer understand and interpret human languages. NLP Pipeline is an automated workflow which enables it to take Human language variants (speech, text etc.) and transform it in various stages to get desired output. Our model takes Tweets based on a specific keyword, hash tags or mentions and gives us a tag cloud along with the polarity of those tweets represented with a visual index. Our workflow is described in various steps as follows such as data collection phase, preprocessing, feature extraction, partitioning, training. Afterwards, the testing and deployment is shown in Sect. 5.

4.1 Data Collection

Data collection is a phenomenon of taking information on a specific variable in an already established system. The primary objective of this step is to get data which is more reliable and feature-rich. If we fail to find content-rich data, then our conclusion would be unreliable and our informed data-driven decision-making process will have low-accuracy predictions. That's why we have tried creating out own database instead of using an already built one. Twitter language changes according to trends, which is why it's crucial to train our model with recent data.

In this model, we have created data; 20K for model development (training and testing) and 150K for deployment from Twitter's public database. To collect the dataset, we require a Twitter Developer Account which can be created (for free) on dev.twitter.com and then we would be able to connect Twitter API. Twitter API lets us access to the whole Twitter Public Database and we can filter those based on our requirements. Twitter offers 4 keys (API Key, API Secret Key, Access Token Key, and Access Secret Token Key) which authenticate our use of the Twitter API. We have used Knime's Twitter API Connector node for this authentication. The next step is to use appropriate filters which would give us reliable and content-rich data. Here, we've used keywords as our primary filter mechanism. Our training part doesn't require any specific data, so we have used random, but distinct tweets. For our deployment phase, we have used keywords (Covid-19, Vaccine, Stay home) as filters to collect only relevant data. The later tweets contain other relevant features such as country, time, retweet count, favorite count, retweeted from which can be used for further analysis of the tweets if desired. This is done via Twitter Search node.

Before we could train our model with those distinct 20K tweets, we need to label it. So checked each tweet and labelled it manually with positive, negative or neutral polarity. Now, it's ready for preprocessing. Positive tweets are opinions of hope,joy, etc. while Negative tweets show fear, sadness, pessimism etc. and Neutral tweets arenot opinions, they are just facts. Figure 2 shows Tweet sample for negative, positive, and neutral.

4.2 Pre-processing

As we have discussed earlier, this step is crucial in NLP pipeline especially for Twitter. When we take data as unstructured source, most of our data is redundant, i.e., does not make much effect on the conclusion. So, we need only relevant data which can run in our model and provide valuable insights leading to data-driven decisions. But having those less relevant bits of data makes our model complex and the time complexity increases. Our goal is to minimize those time and space complexity. So, we need to transform our data into something that will make the model simple, yet effective. There are following steps to follow in preprocessing.

1. Normalization: This step filters out irrelevant information from the data and makes the data simpler. This stage consists of 4 nodes in its workflow.

- Punctuation Eraser: This node takes the input document removes all the punctuation from it.
  Example: Ah! What happened? =>Ah What happened.
- Case Converter: This takes the document from the previous node and transforms everything into lowercase letters.
  Example: John is STUPID = > john is stupid.
- N-Char Filter: This node takes the document and removes any special character in it.
  Example: i gave him $10 = > i gave him 10.
- Number Filter: This node removes the numerical values presented in the input document.
  Example: i donated 10 million = > i donated million.

2. Stop-words filter: This stage takes the document and filters out redundant words which does not provide much information. We have used the standard stop words library provided by OpenNLP.
Example: I was 10 inches wider before. => i inch wider before.
3. Tokenization: It's a phenomenon of separating a text into smaller units called 'tokens'. If we take a sentence as input, then we will get individual words as tokens. Our model uses the default node provided by Knime for this.
4. Stemming & Lemmatization: This process involves reducing individual words to its root form or 'stem'.
Example: run/runs/running = > run.

Lemmatization is a process where a pre-defined dictionary is used to reduce each token into its root form or 'lemma'.

Example: is/was/were = > be.

Our model uses Stanford Lemmatizer in this step.

4.3 Feature Extraction

Human language is complex for computers unlike numbers. In case of numbers machines are faster to process them and manipulate them into required form. A vector is used to perform calculations and find relevant conclusions. So, we need to feed the machines texts in numbers form so that they can understand it and transform it into meaningful insights for data-driven decisions. In this stage of our NLP pipeline, we convert out textual data into n-dimensional vector. For images and videos, there're pixel values which are numbers. But human language is much more complex. Feature extraction is a process where we extract and present a feature representation that we need for our specific NLP tasks. Features are the attributes that helps us understand our data better. In our model, we have used TF-IDF which is discussed as follows.

TF-IDF

In Bag of words, each word is treated equally and that's a mistake. Because some words have more relevance and feature-rich than others and we need to use that feature in our model. That's where TF-IDF comes in. It stands for Term Frequency-Inverse Document Frequency [32–35]. TF-IDF has a weighting factor which counts as the relevance for each word in a document. A word's relevance increases proportional to the number of times a word appears in the whole document. If a word appears less times, then its relevance is more.

Term frequency provides us the information about how frequently a word appears in the entire document. It tells us the probability of a word within that document. Term frequency tf(x,y) = Nb(x,y), where Nb(x,y) is the number of times a word x appears in document y. Inverse document frequency is denoted as:

idf(x,y) = log[Nb/df(x)] [1]

Where Nb is the total number of documents and df(x) is the number of documents that contain the word x.

tf-idf(x,y) = tf(x,y)*idf(x,y) [2]

When a term appears more frequently in the whole document, then the log value approaches to 0 making the term less relevant. This step ends the feature extraction part. Now, it's time to actually train the model.

4.4 Partitioning

To build an effective model which can predict new data with high accuracy, a model should be trained on a diverse dataset. Partitioning is a process where we split our data into 2 sets; train data and test data. Typically, we need to use more data to train our model, but our testing data should be enough diverse to successfully test the validity of our data. That's why we used 85 – 15 split on our dataset; 85% is used for

training and 15% for testing. Our dataset consists of 20K manually labelled by us and we found that 3K tweets is enough diverse to test the validity. We have infused the sentiments with distinct colors for better visual representation purposes; Green for Positive Sentiment, Red for Negative Sentiment and Orange for Neutral.

4.5 Training

We have a labelled dataset of 20K tweets and we're using 17K tweets to train our model. The tweets are classified into 3 polarity; positive, negative, neutral sentiment. Our model consists of results from 3 types of Machine Learning models [17–31]; Decision Tree, SVM, and Logistic Regression. The simulation is performed better under this environment with 88.67% accuracy in SVM. We have discussed the machine learning models as follows.

Decision Tree

Decision Tree (DT) is a type of supervised machine learning algorithm where it's represented in a tree-based model with each node representing a test on the attribute, and each branch represents the results those tests. A Decision Tree model learns by splitting the input dataset into subsets based on an attribute value test. This step is repeated multiple times till all data is classified or the splitting doesn't add any value. This type of classification doesn't require any domain-specific knowledge to classify the data. In other words, it starts with root node at the node and asks a question every time a data tries to pass through. The answer decides which direction it would go next. Then the process continues till the data reaches the terminal node (leaf nodes). In our model, we've used gini-index as quality measure technique and no pruning is applied. Table 1 and Table 2 shows the results for DT.

Table 1
Accuracy of Decision Tree for Positive, Negative and Neutral

	Positive[predicted]	Negative[predicted]	Neutral[predicted]
Positive[actual]	1750	10	50
Negative[predicted]	60	350	50
Neutral [predicted]	240	30	460
	85.37%	89.74%	82.14%

Table 2
Accuracy, Error, Cohens Kappa of Decision Tree

Accuracy	Error	Cohen's kappa	Correct	Incorrect
85.33%	14.67%	0.719	2560	440

SVM: Support Vector Machine (SVM) is a machine learning algorithm where we try to find a hyperplane to successfully classify our n-dimensional dataset. To find the best possible hyperplane, we use a plane which separates the data with maximum margin. The dimension of the hyperplane depends on the number of features used. For two features, it's a line and a plane if it's 3. Kernels are a mathematical function which is used to transform the input data to desired form. Here, we've tried the model with Polynomial, hypertangent, RBF(Radial Basis Function) and found that polynomial kernel performed best with bias, gamma and power set to 1.0 and 1.2 overlapping penalty. Table 3 and Table 4 shows the results for SVM.

Table 3
Accuracy of SVM for Positive, Negative and Neutral

	Positive[predicted]	**Negative[predicted]**	**Neutral[predicted]**
Positive[actual]	1740	10	60
Negative[predicted]	50	360	50
Neutral [predicted]	140	30	560
	90.16%	90.00%	83.53%

Table 4
Accuracy, Error, Cohens Kappa of SVM

Accuracy	**Error**	**Cohen's kappa**	**Correct**	**Incorrect**
88.67%	11.33%	0.789	2660	340

Logistic Regression

Logistic Regression (LR) is a classification machine learning algorithm used to classify binary dependent variables (pass/fail), but it can be used to classify more values (multinomial). Stochastic gradient solver algorithm makes a prediction for a training instance and the model is updated based on the error on that training instance. The model repeats till error are reduced or a specific epoch is mentioned. Here, we've used 100 epochs with fixed learning rate strategy and 0.1 step size. To regularize 0.1 variance is used on gauss prior. Table 5 and Table 6 shows the results for LR.

Table 5
Accuracy of Logistic Regression for Positive, Negative and Neutral

	Positive[predicted]	Negative[predicted]	Neutral[predicted]
Positive[actual]	1565	0	245
Negative[predicted]	20	320	120
Neutral [predicted]	45	0	685
	96.01%	100.00%	65.23%

Table 6
Accuracy, Error, Cohens Kappa of Logistic Regression

Accuracy	Error	Cohen's kappa	Correct	Incorrect
85.67%	14.33%	0.695	2570	430

From Fig. 3 it is observed that SVM shows high accuracy than DT and LR. From Fig. 4 it is observed that SVM has less error than DT and LR. Figure 5 shows SVM has high Cohens Kappa than DT and LR. From the results it is concluded that it is better to prefer SVM for our sentiment analysis model to analyze the Covid-19 vaccination tweets.

5 Results And Discussion

The simulation is performed in machine of 8 Gb RAM, 2.4 GHz processor, and 64-bit Windows 10 operating system platform. The KNIME platform [17] is used all data analysis. We have successfully tested our machine learning models over 3K tweets and found that SVM algorithm performed the best with 88.67% accuracy where polynomial kernel was used with bias, gamma and power set to 1.0 and 1.2 overlapping penalty.

5.1 Deployment

The objective is to create a model which can provide us an overview of an event with relatively high accuracy. We deployed this in two steps; Interactive Tag Cloud and Sentiment Analysis. The former gives us an idea about the key terminologies meanwhile the latter gives us insights about people's sentiment based on a given event. Here, we've chosen Covid-19 vaccination as our event and we've managed to deploy 150,000 tweets on this event spanning from January 2021 to May 2021. The tweets were collected through Twitter API connector and Twitter Search nodes on Knime platform.

5.2 Interactive Tag Cloud

This phase produces a Word cloud with high relevant terms from the tweets. The size of each term describes how frequently that term has been used in the whole tweet dataset. To classify the tweets, we've used the time of posting as the deciding factor. The timeline is divided into 4 classes;

morning(6AM-12PM), afternoon(12PM-6PM), Evening(6PM-12AM), Night(12AM-6AM). To represent the tag cloud, we have assigned colors to each time parameters; Morning (Orange), Afternoon (Red), Evening (Green), Night (Yellow). This is achieved via Knime's default node for word cloud. If we click on each term in the cloud, then we can see the associated tweets which would be helpful for us to understand the importance of the term. From the Fig. 7, we can safely conclude that most tweets come at night.

5.3 Sentiment Analysis

After getting an understanding of the keywords, we tried sentiment analysis. In our training phase, we have found that SVM performed best with polynomial kernel with 1.2 overlapping penalty. So, we have deployed our model with that parameters. Keywords used in the data collection are Covid-19, Vaccine, Stay home. 50K Tweets are collected for each keyword mentioned to showcase the sentiment polarity.

- This first deployment of our model takes 50K Tweets as its input and the tweets are based on the key word 'Vaccine' and also shown in Fig. 8.

POSITIVE: 4.1% [2050 Tweets]

NEGATIVE: 39.1% [19550 Tweets]

NEUTRAL: 56.8% [28400 Tweets]

This conclusion from this set of Tweets is that people aren't necessarily happy about the vaccines, but they're complying anyway.

- The second iteration of Tweets is based on the keyword 'Covid-19'and also shown in Fig. 9.

POSITIVE: 10.26 % [5130 Tweets]

NEGATIVE: 75.36 % [37680 Tweets]

NEUTRAL: 14.38 % [7190 Tweets]

This shows us that people are angry towards Covid-19 and there are very few willing to look at the upsides.

- The final iteration of the Tweets is based on the keyword 'Stay home'and also shown in Fig. 10.

POSITIVE: 83.34 % [41670 Tweets]

NEGATIVE: 3.1 % [1550 Tweets]

NEUTRAL: 13.56 % [6780 Tweets]

From this result we can say with reasonable certainty that people are very happy about staying home.

6 Conclusions

In this paper, we evaluated simple machine learning algorithms when used for sentiment analysis task and we yielded 88.67% accuracy. We also deployed our model on an interactive tag cloud to showcase important terminologies. There is tons of research on this problem where complex algorithms, rule-based filters are used to get higher accuracy. Our model showed relatively decent accuracy compared to those, but it's simpler. This model lacks comprehensiveness and higher accuracy. Our future works include implementing a name-entity model to better understand the key terminologies along with making our sentiment analysis part hybrid for better accuracy while keeping the model simple. The model isnot equipped to handle political campaigns or million-dollar product launch yet. They are extensive tasks and demand more insights than our model is currently capable of producing. Those problems demand more accuracy and a comprehensive analysis as the failure to meet them could cause catastrophic consequences (riots, loot, ban etc.). Still, our model produced quite impressive results with 88.67% accuracy which is better suited for regional shop owners to gather insights from their customer reviews, or something that have less series consequences upon failure. As our model showed that people are quite happy to stay at home, this can be an opportunity for companies to market their products in a way suited for those people. Having meaningful insights help make better data-driven decisions and our model tries to deliver that in a low-cost solution.

Declarations

Ethics declarations

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Authorship contributions

All authors contributed equally to this work.

Data availability

Data can be shared if needed.

References

1. Fan DP "Predictions of consumer confidence/sentiment from the press." In *Proceedings of the Section on Survey Research Methods*, vol. 2, pp. 1154–1159. American Statistical Association Westport, CT, 1993
2. Hu M, Liu B. "Mining opinion features in customer reviews." *AAAI*. Vol. 4. No. 4. 2004

3. Zhang X, Fuehres H, Peter A (2011) Gloor. "Predicting stock market indicators through twitter “I hope it is not as bad as I fear”. Procedia-Social Behavioral Sciences 26:55–62
4. Kim S-M, Hovy E. "Determining the sentiment of opinions." In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1367–1373 (2004) Kim, Soo-Min, and Eduard Hovy. "Determining the sentiment of opinions." In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1367–1373. 2004
5. Kouloumpis E, Wilson T, Moore J (2011) Twitter Sentiment Analysis: The Good the Bad and the OMG!. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1)
6. Shelar A, and Ching-Yu Huang. "Sentiment analysis of twitter data." *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2018
7. Ding X, Liu B, and Philip S. Yu. "A holistic lexicon-based approach to opinion mining." *Proceedings of the 2008 international conference on web search and data mining*. 2008
8. Prabowo R, Thelwall M (2009) Sentiment analysis: A combined approach. Journal of Informetrics 3(2):143–157
9. Go A, Huang L, Bhayani R (2009) "Twitter sentiment analysis" *Entropy* 17:252
10. Arafin Mahtab S, Islam N, Mahfuzur Rahaman M, "Sentiment Analysis on Bangladesh Cricket with Support Vector Machine," 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018, pp. 1–4, doi: 10.1109/ICBSLP.2018.8554585
11. Makhmudah U, Bukhori S, Putra JA, Yudha BAB, "Sentiment Analysis Of Indonesian Homosexual Tweets Using Support Vector Machine Method," 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), 2019, pp. 183–186, doi: 10.1109/ICOMITEE.2019.8920940
12. Muhammad A, Nizar S, Bukhori, Pandunata P. "Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes–Support Vector Machine (NBSVM) Classifier." *2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*. IEEE, 2019
13. Tho C et al ("Code-Mixed Sentiment Analysis Using Machine Learning Approach–A Systematic Literature Review." *2020*) *4th International Conference on Informatics and Computational Sciences (ICICoS)*. IEEE, 2020
14. Bao Y et al. "The role of pre-processing in twitter sentiment analysis." *International conference on intelligent computing*. Springer, Cham, 2014
15. Saif H, Fernandez M, He Y, Alani H (2014) On stopwords, filtering and data sparsity for sentiment analysis of twitter
16. Jianqiang Z (2017) and Gui Xiaolin. "Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access 5:2870–2879
17. https://www.knime.com/knime-analytics platform?gclid = CjwKCAjwzMeFBhBwEiwAzwS8zFADhci0x9BunHSf7AoJ2etXv96vM1hhMc4cAp3G3gE_tR5a-NlYDxoCGh4QAvD_BwE, accessed on April 2021

18. Ritonga M, Ihsan MAliA, Agus Anjar, and Fauziah Hanum Rambe. "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm." In *IOP Conference Series: Materials Science and Engineering*, vol. 1088, no. 1, p. 012045. IOP Publishing, 2021
19. Villavicencio C, Macrohon JJ, Alphonse Inbaraj X, Jeng J-H, Jer-Guang H. "Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes." *Information* 12, no. 5 (2021) 204
20. Ghasiya P, Okamura K (2021) Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach. IEEE Access 9:36645–36656
21. Dubey AD. "Public Sentiment Analysis of COVID-19 Vaccination Drive in India." *Available at SSRN 3772401* (2021)
22. Chowdhury A, Ahmed A, Das SK, Saha, Mahfujur Rahman, and Khandaker Tabin Hasan. "Sentiment Analysis of COVID-19 Vaccination from Survey Responses in Bangladesh." (2021)
23. Nurdeni D, Ade I, Budi, Aris Budi S. "Sentiment Analysis on Covid19 Vaccines in Indonesia: From The Perspective of Sinovac and Pfizer." In (2021) *3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, pp. 122–127. IEEE, 2021
24. Hussain A, Sheikh A. "Opportunities for artificial intelligence–enabled social media analysis of public attitudes toward Covid-19 vaccines." *NEJM Catalyst Innovations in Care Delivery* 2, no. 1 (2021)
25. Hung M, Lauren E, Hon ES, Birmingham WC, Xu J, Su S, Hon SD, Park J, Dang P, Martin S (2020) Lipsky. "Social network analysis of COVID-19 Sentiments: application of artificial intelligence. J Med Internet Res 22(8):e22590
26. Kwok SW, Hang SK, Vadde, Wang G (2021) Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: Machine learning analysis. Journal of Medical Internet Research 23(5):e26953
27. Abdulkareem NM, Abdulazeez AM, Zeebaree DQ, Dathar A, Hasan (2021) COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms. Qubahan Academic Journal 1(2):100–105
28. Bhoi SK (2021) "Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach". Turkish Journal of Computer Mathematics Education (TURCOMAT) 12(10):3074–3084
29. Bhoi S, Kumar KK, Jena SK, Panda HV, Kumar LR (2021) P. Subbulakshmi, and Haifa Bin Jebreen. "An Internet of Things assisted Unmanned Aerial Vehicle based artificial intelligence model for rice pest detection. Microprocess Microsyst 80:103607
30. Jena K, Kumar SK, Bhoi SR, Nayak, Mallick C. "Machine Learning-Based Virus Type Classification Using Transmission Electron Microscopy Virus Images." *Machine Vision Inspection Systems, Volume 2: Machine Learning-Based Approaches* (2021) 1–22
31. Apoorva A, Mishra GK, Sahoo RR, Bhoi SK (2021) and Chittaranjan Mallick. "Deep Learning-Based Ship Detection in Remote Sensing Imagery Using TensorFlow". In: Advances in Machine Learning and Computational Intelligence. Springer, Singapore, pp 165–177

32. https://en.wikipedia.org/wiki/Tf%E2%80%93idf, accessed on April 2021
33. Ahuja R, Chug A, Kohli S, Gupta S, Pratyush Ahuja. "The impact of features extraction on the sentiment analysis." Procedia Computer Science 152 (2019): 341–348
34. Avinash M, Sivasankar E (2019) "A study of feature extraction techniques for sentiment analysis". In: Emerging Technologies in Data Mining and Information Security. Springer, Singapore, pp 475–486
35. Agarwal B, Mittal N (2016) Prominent feature extraction for sentiment analysis. Springer International Publishing, Berlin

Figures

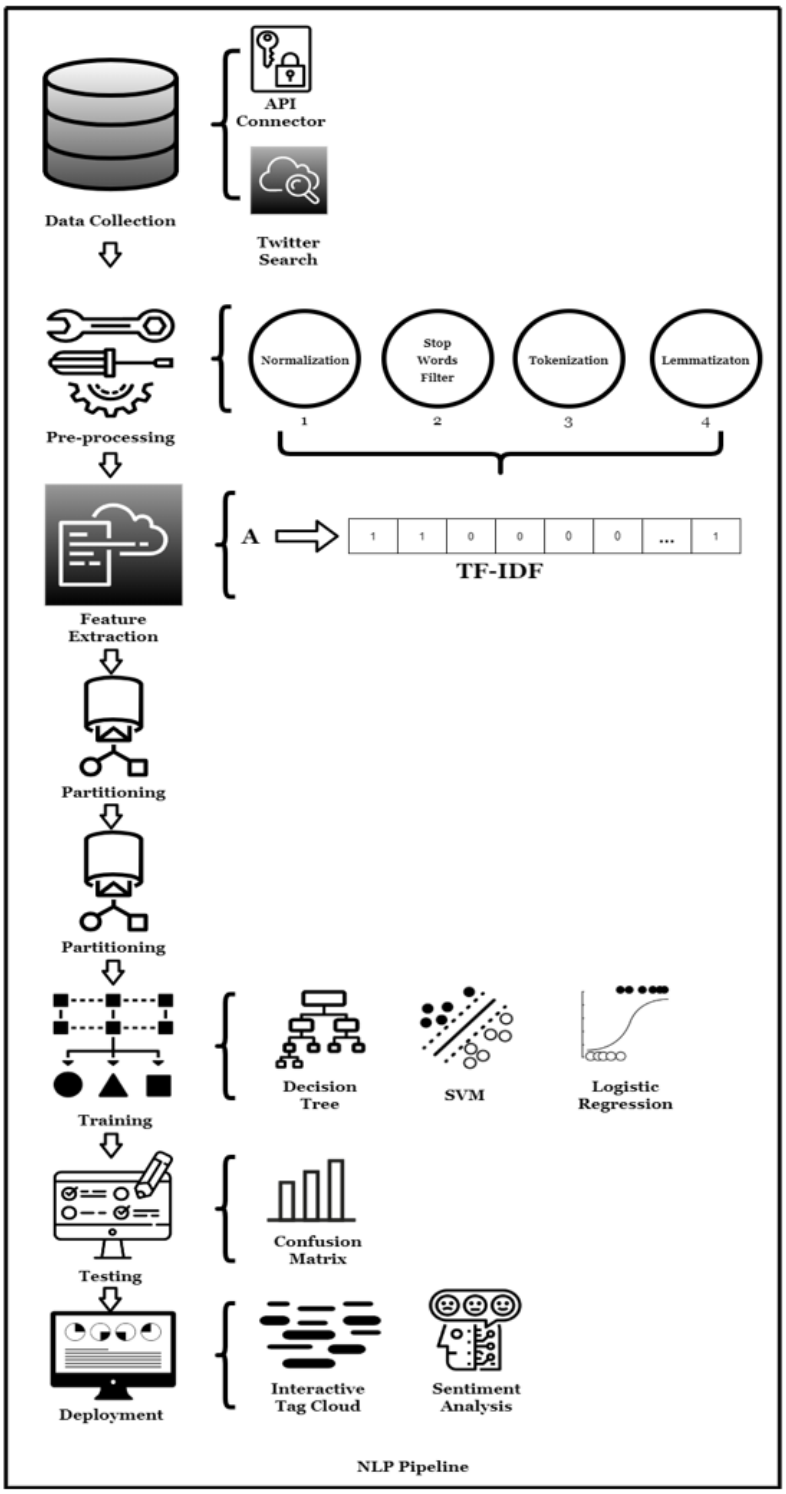

Figure 1

Workflow of the sentiment analysis model.

Row 1	Covid-19 has ruined all the stuff that. . .	Negative
Row 2	I think this pandemic has taught us to. . .	Positive
Row 3	50k died last night in NY City Hospit. . .	Neutral

Figure 2

Tweet sample for negative, positive, and neutral

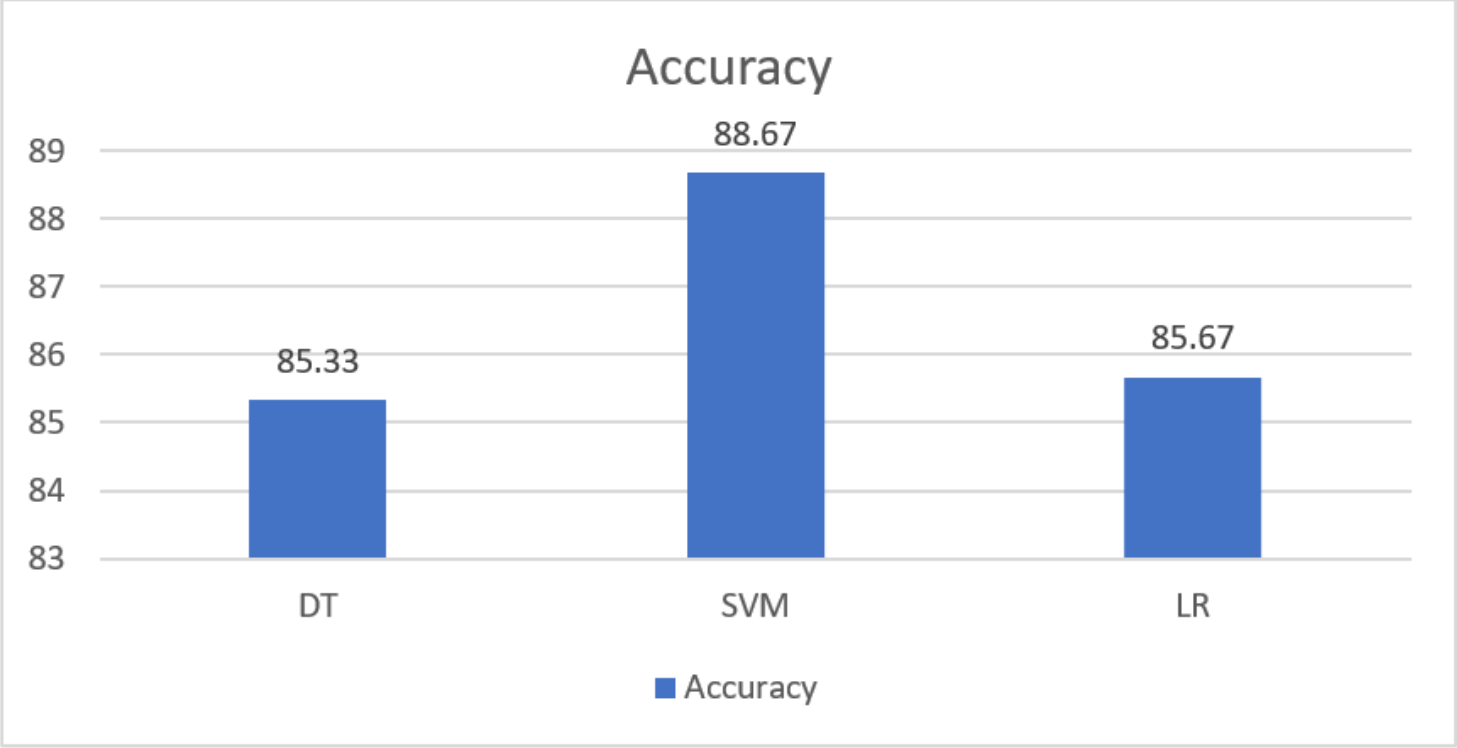

Figure 3

Accuracy of DT, SVM, and LR.

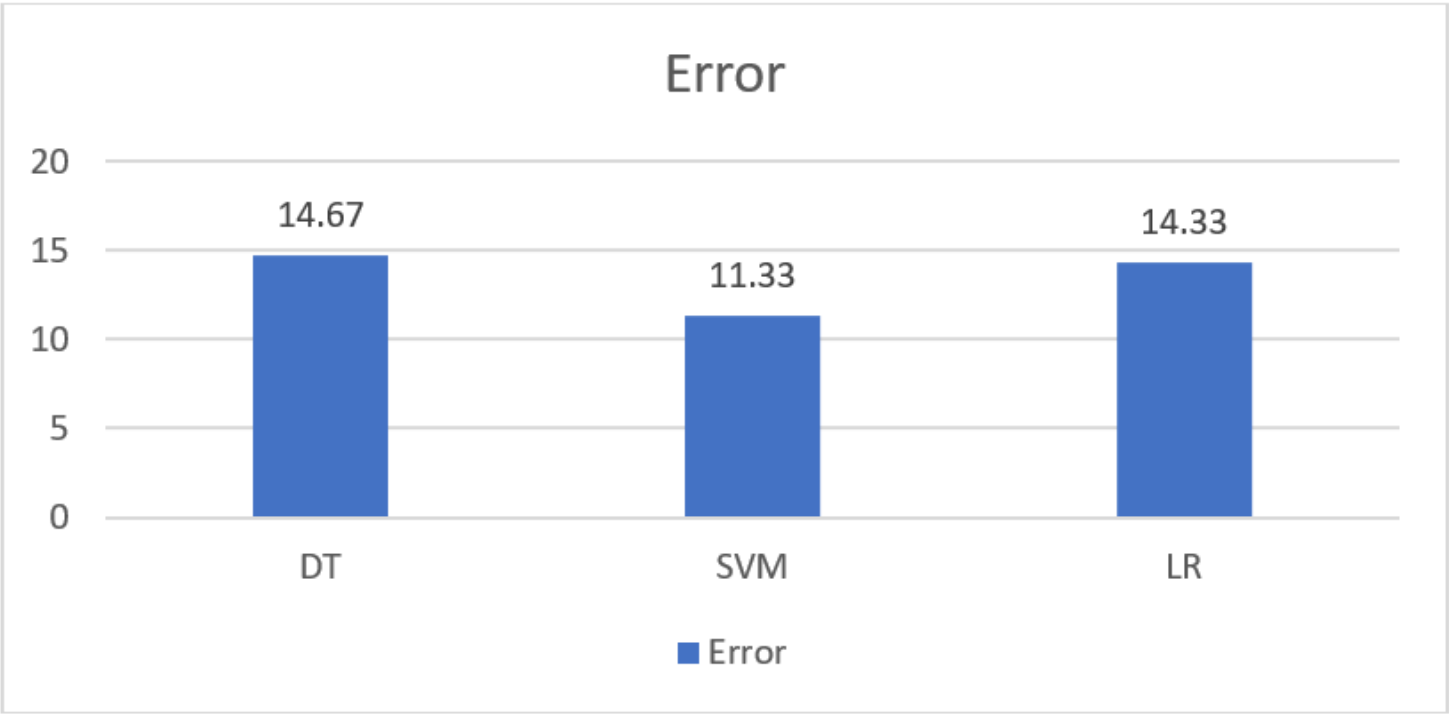

Figure 4

Error of DT, SVM, and LR

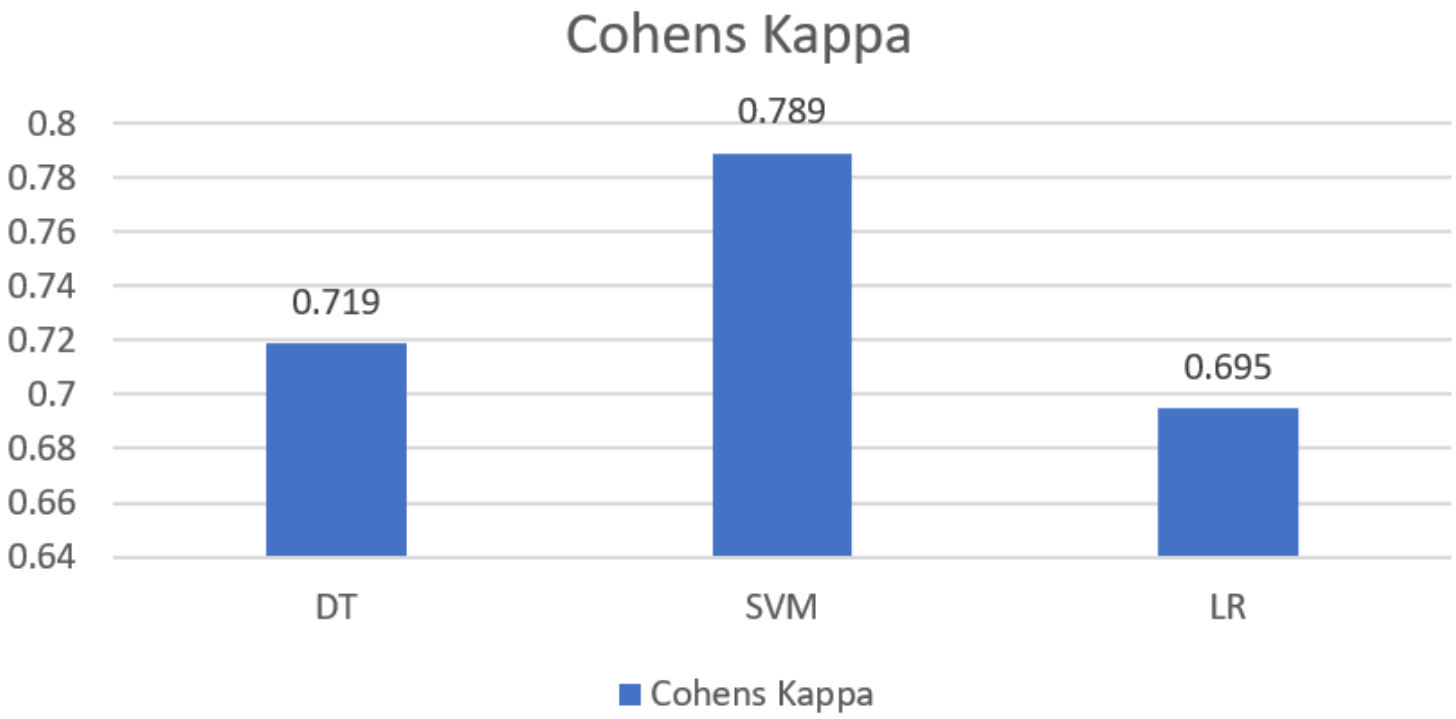

Figure 5

Cohens Kappa of DT, SVM, and LR

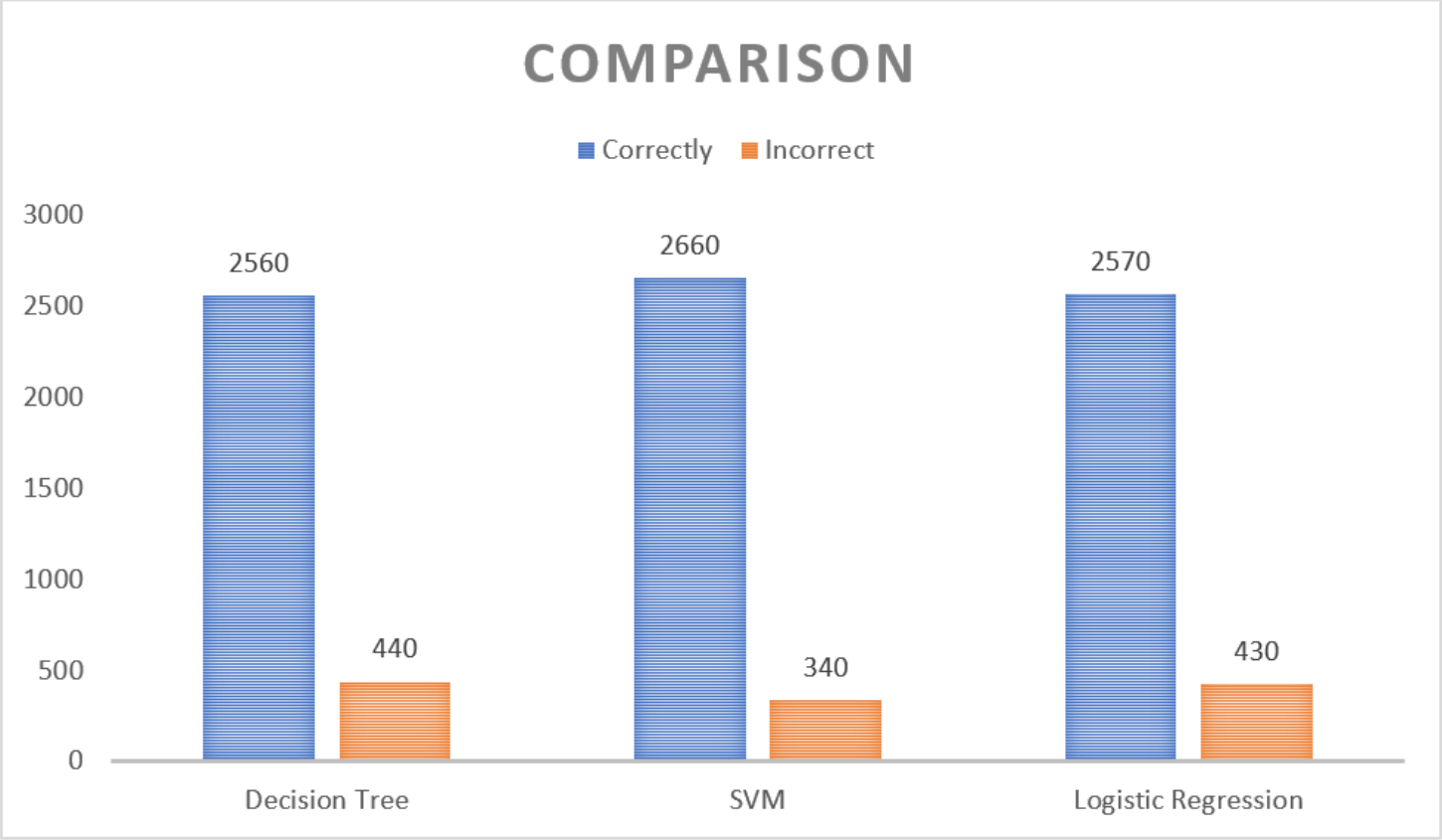

Figure 6

Tweetstested for DT, SVM, and LR

Figure 7

Tag cloud for tweets Morning (Orange), Afternoon (Red), Evening (Green), Night (Yellow).

Figure 8

Sentiment analysis for the keyword “vaccine”

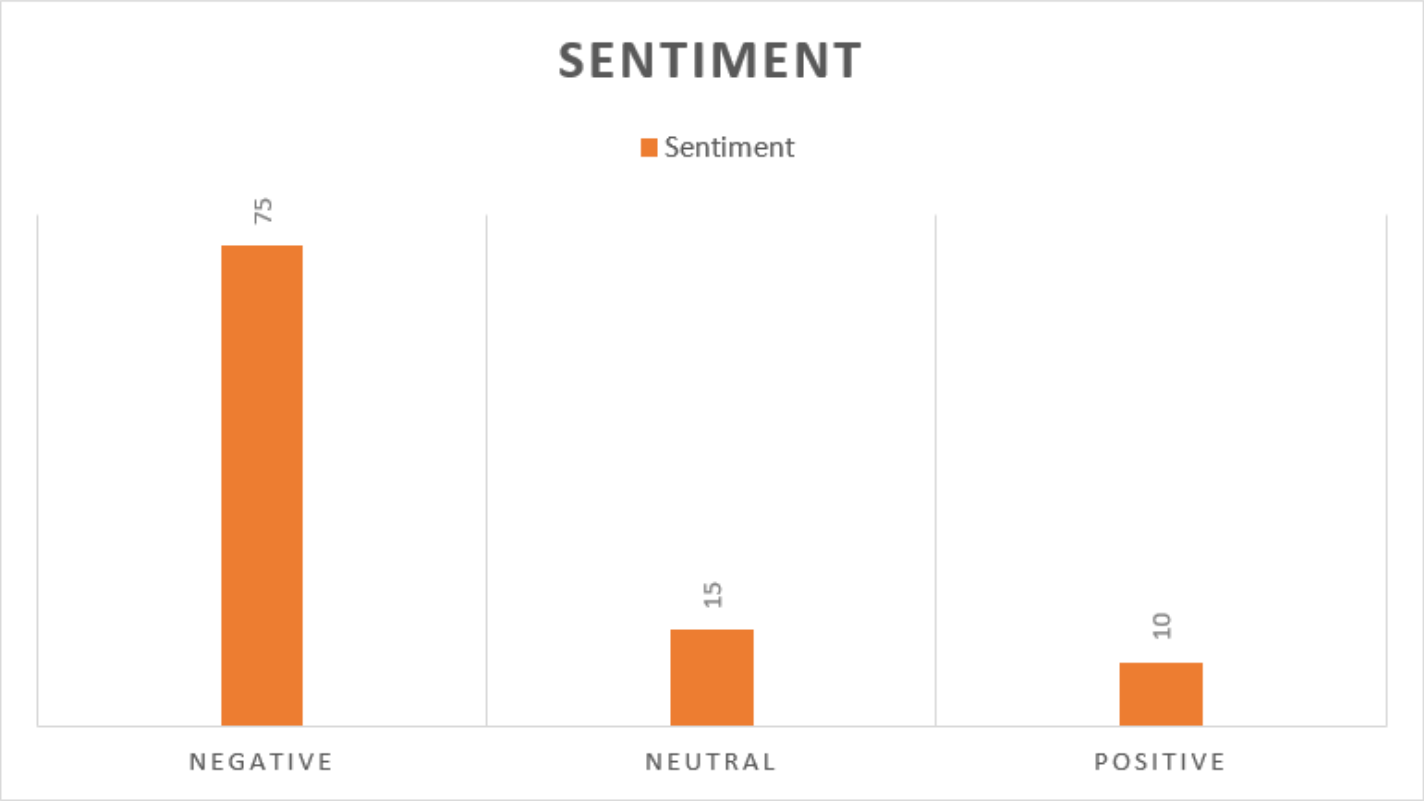

Figure 9

Sentiment analysis for the keyword “Covid-19”

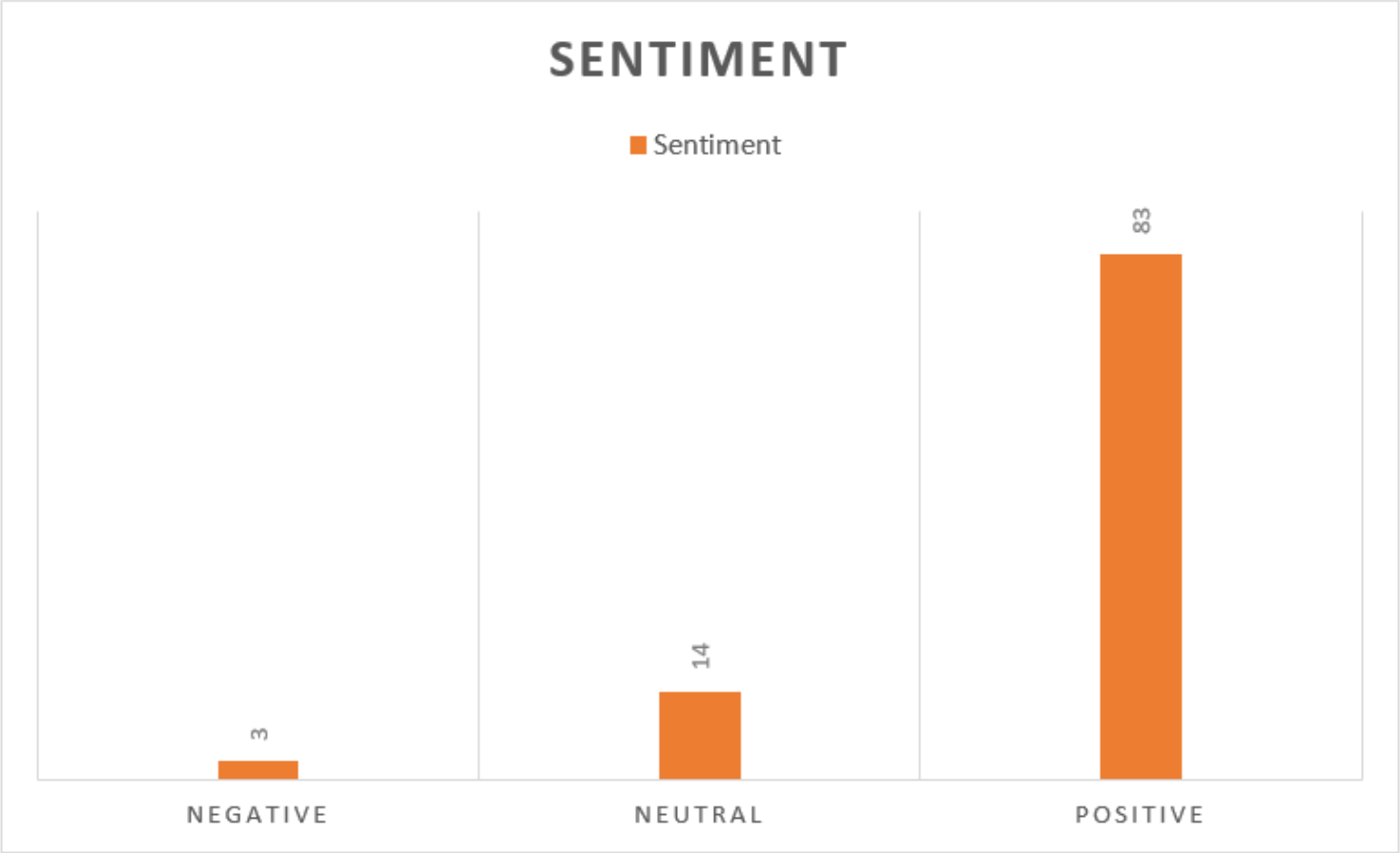

Figure 10

Sentiment analysis for the keyword “Stay home”