Sentiment Analysis of COVID-19 Tweets

Labeling tweets' sentiment based on content

Natural Language Processing and Text Analytics

KAN-CDSCO1002U

Submission Date: 30th of May 2022

Pages: 15

Character Count: 31.980

Authors:	**Professors:**
Max Ladegaard (110166)	Rajani Singh
Viktoria Sundelin (150664)	Daniel Hardt
Emilia Konopko (149895)	
Rebecca Emilia Trulsson (149868)	

Abstract

The project aims to find a successful approach to analyzing tweets published on twitter.com concerning COVID-19, during the year of 2020. The specific characteristics explored are the sentiment of the tweets and the topics discussed. The experiment has been executed by comparing four models, Naïve Bayes as the base model, a Logistic Regression model and two RNN models, LSTM and GRU. The models were chosen to examine if a more complex model achieves a better prediction for a sentiment analysis. The result displayed that the GRU outperformed the other models, though only slightly better than LSTM. Therefore, it can be stated that using a more complex model than the Naïve Bayes is preferable and that the RNN models are both good alternatives. The topics of the tweets were uncovered through topic modeling, and displays that the most concerned topics were supply, demand and shopping.

Keywords: Sentiment Analysis, COVID-19, Logistic Regression, Naïve Bayes, Long Short-Term Memory, Gated Recurrent Unit

1. Introduction

The commencement of COVID-19 was in December 2019, but it did not reach the general public's attention until the rapid spread in the start of 2020. The World Health Organization started recommending national lockdown to prevent the spread of the disease. The pandemic itself changed all peoples' everyday life for the last two years, some populations faced stricter restrictions than others, but everyone was affected. During this time, activity on social media platforms rapidly increased because of the limitation in engaging in other activities. One of the leading social media platforms is Twitter, the platform is projected to reach 329 million active monthly users in 2022 (Statista, 2022a). Twitter can be considered as a microblogging service on real-time data that allows its users to share their opinions and emotions in short text publications known as 'tweets'.

Along with the rise of COVID-19 and the following restrictions, many people's mental health was affected and it became an acknowledged societal problem (Talevi et al., 2020). Tweets related to the outbreak of COVID-19 can be categorized into different sentiments that are based on the subjective information of the tweets. By studying these tweets' text polarity, it is possible to assign them into different groups and study the expressed attitudes. With this in mind, this project strives to answer the question of how to build an effective tool with the most suitable model in order to perform an accurate sentiment analysis and uncover topics discussed in COVID-19 tweets.

2. Related Work

Prior research on the application of natural language processing methods to social media mainly dealt with sentiment analysis. Lwin et al., (2020), for instance, collected over 20 million Twitter posts to examine global trends of the emotions fear, anger, sadness and joy and derive their underlying narratives during the COVID-19 pandemic. Colnerič and Demšar's work (2020) aimed at extending previous studies of emotion recognition on Twitter, arguing that they mainly focused on the use of lexicons and simple classifiers on bag-of-words models. Instead, he proposed a deep learning approach based on several word- and character-based recurrent and convolutional neural networks and compared them to latent semantic indexing models as well as bag-of-words. His results suggest that the newly proposed training heuristic generated a model with performance comparable to that of three single models.

As the pandemic progressed, however, Twitter was one source of microblogging for people where they could release all thoughts and emotions. Tweets related to COVID-19 have been a popular data to perform sentiment analysis on and classify them into different sentiments such as positive, neutral and negative (Khan et al., 2020; Nemes & Kiss, 2020; Dubey, 2020). Khan et al. (2020) performed a sentiment analysis of COVID-19 tweets to be able to analyze people's reaction to decisions made by the government or authorities during the pandemic. The authors assigned each sentiment a number from 1, 0, -1 in order for it to be easier to perform classification through the library of Natural Language Toolkit (NLTK) or by pattern based. This paper also focused on the tweet's level of subjectivity, if the tweet contained facts or opinions, to facilitate the differentiation of the data. Nemes and Kiss (2020) also provides a study of the sentiment analysis on COVID-19 related tweets. Their work focuses on building a model with Recurrent Neural Network (RNN) to perform sentiment classification by examining correlations between words that determine the text polarity of the tweet. One of the authors’ focus areas was also to minimize the sentient class ’neutral’ as they considered it would improve the overview of the world’s opinions, if the classes contained solely different levels of positive and negative. This study also touches upon the subject of Twitter having mainly young people as active Twitter users. Hence, this could have a misleading effect on the result as young people were not as negatively affected by the pandemic compared to the older generation. Dubey (2020) performed a sentiment analysis through COVID-19 tweets in order to be able to study people’s attitude towards the pandemic on a country based level. To perform this sentiment analysis the author used The NRC Word-Emotion Association Lexicon which included 10 170 lexical items which could classify the tweet as positive and negative but also further organize them into eight different emotions. (Dubey, 2020)

3. Conceptual Framework

3.1 Sentiment Analysis

Sentiment analysis is a well used technique within Natural Language Processing (NLP) which is a process that identifies and extracts the text polarity of given data containing subjective information (Devika, Sunitha & Ganesh, 2016). Sentiment analysis has grown rapidly as subjective information has become more accessible with the growth of social media posts, another form of review data. There are many different methods and techniques used to perform sentiment analysis that can be placed in three main categories; knowledge based techniques, statistical methods and hybrid approaches (Cambria et al., 2017). The authors explain that knowledge based techniques involve classifying text into categories, depending on the level of unambiguity in the words, examples of these words are ‘happy’ and ‘angry’ etcetera. However, this category possesses a severe weakness as it fails to adapt to some linguistic rules. For example, the inability to detect when there is a negation of a verb versus when it is not, instead it solely focuses on the word that determines the sentiment. A concrete example of this is ‘yesterday was a good day’ versus ‘yesterday wasn’t a good day’. In this example, the knowledge based approach would fail to take into account the meaning of the verb and would only detect the word ‘good’ and from there appoint the sentiment.

In statistical methods, different machine learning and deep learning methods create algorithms that detect sentiments. Through these methods, the model can learn keywords but also learn how to detect random keywords. Within this category there are also limitations, the authors emphasize that these methods only reach a sufficient level of accuracy if the models are being trained on large text input, as the results are inadequate for smaller training sets. Lastly is the hybrid category, which are methods that draw techniques from both the knowledge based and statistical methods and exploit the advantages that both categories generate. Usually projects that include 'emotion recognition and polarity detection' (pp. 105) derive from the hybrid approach. (Cambria et al., 2017) In this project, we are primarily using statistical methods as we are executing the models of Naïve Bayes, Logistic Regression, Long Short-Term Memory and Gated Recurrent Unit.

3.2 Training Strategy

For this project, the data preprocessing is of great importance as it determines the functionality and outcome from the models we train on the data. The main steps of the data preparation are filtering and standardization, stop words removal and lemmatization as well as tokenizing. These steps were crucial as Twitter data is naturally very unstructured and needs to be converted to a more structured form. Thereafter, the clean data is used in four different models which constitute two substantial types, such as Text Classifiers and RNNs. The Naïve Bayes method was used as the base model to test the other models against, due to the simplicity of implementing it. All of these models will later be compared and see which one is the most suitable to build as a tool for sentiment analysis on COVID-19 Twitter data.

3.3 Logistic Regression

One of the most used techniques for sentiment analysis is Logistic Regression. Prabhat and Khullar (2017) describes Logistic Regression as a discriminative classifier that learns the features that are leading for distinguishing the different outcome classes. Furthermore, the features enter the prediction function and the output is the estimated sentiment. Prabhat and Khullar (2017) emphasizes that it is important that the different outcome classes are mutually exclusive in order for a Logistic Regression to be functional. For this project, the different outcome classes are the five sentiments; extremely negative, negative, neutral, positive and extremely positive. Depending on the amount of categories, the method is either binary or multinomial Logistic Regression. As stated this project has five different categories which is why we have performed a multinomial Logistic Regression. All of these are also mutually exclusive and suitable for performing a Logistic Regression because they contain certain features/words that distinguish a tweet from being negative and positive, or positive and extremely positive.

3.4 Naïve Bayes

The Naïve Bayes (NB) is another favorable method that is widely used within sentiment analysis due to its effectiveness and simple application (Abbas et al., 2019). The Naïve Bayes model belongs to the Text Classifier category of generative classifiers, which predicts the outcome based on conditional probabilities and does not take into account the words' relationship to each other into its computing (Prabhat & Khullar, 2017). Hence, it solely treats the words as a bag of words and computes the outcome class depending on the probability. Despite this, the method has been very successful for spam detection. This method is applicable for both supervised and unsupervised machine learning and can either appear as binary or multinomial (Prabhat & Khullar, 2017). For this project, we perform supervised machine learning as our labels, as known as sentiments, are known beforehand. Since the labels are also more than two, the method that is used in this project is the Multinomial Naïve Bayes.

3.5 Long Short-Term Memory

According to Géron (2019), the Recurrent Neural Network (RNN) suffers from one great disadvantage, its ability to store the earlier parts of the sequence and therefore the final output becomes incomplete. Sachin et al. (2020) propose that a gating mechanism to the RNN would be a solution to this weakness. The authors explain that the Long Short-Term Memory (LSTM) is a gated version of RNN that solves the short-term memory problem as it replaces every hidden unit with a LSTM cell that consists of new parameters that act as gates that determine the usage of the input. Each cell consists of three gates; input, forget and output that regulate the input depending if the word is important enough to be remembered for in order for the LSTM output to be accurate. (Sachin et al., 2020)

3.6 Gated Recurrent Unit

Another solution to the limitation of RNN is to create a Gated Recurrent Unit (GRU). This works as an added gated mechanism to a Recurrent Neural Network but does not contain the same level of functionality as the LSTM, making the GRU a lightweight version of it (Sachin et al., 2020). The GRU combines the long-term and short-term memory into its hidden state cell whereas it consists of two gates which are update or reset, which controls the movement of information throughout the different cells. Essentially, these gates overlook the information passed on from the previous cell and determine how much of that information should be retained and how much should be forgotten. (Sachin et al., 2020) Furthermore in this project, it has been relevant to add an extension to the GRU model, making it a Bidirectional Gated Recurrent Unit (Bi-GRU). This extension allows our model to exploit the context information from forward as well as backwards direction to compose the prediction in the current state (Sachin et al., 2020).

4. Methodology

4.1 Dataset Description

The dataset is a gathering of tweets regarding COVID-19 published on Twitter during the timeframe 4th of January 2020 to 4th of December 2020. It is structured in six columns; Username, ScreenName, Location; TweetAt, Tweet and Sentiment. The columns UserName and ScreenName have been replaced with numbers to maintain privacy of the account owners. The Tweets have been annotated manually and labeled in the Sentiment column to describe the sentiment of the Tweet. The five classes they have been divided in are; extremely negative, negative, neutral, positive and extremely positive. Each tweet has only one class labeled to it. In total, the dataset is 11.5 MB and contains 8.292 rows.

Table 1. Dataset description

Column Name	Explanation	Data Type
UserName	The user name of the user who has written and uploaded the tweet. Replaced by numbers.	int64

ScreenName	The screen name of the user who has written and uploaded the tweet. Replaced by numbers.	int64
Location	The location the user has chosen.	object
TweetAt	The date when the tweet was published.	object
Tweet	The tweet in text.	object
Sentiment	The manual label of sentiment that the Tweet has been annotated as.	object

4.2 Exploratory Data Analysis

To explore the full dataset, the first step was to combine the test and the train data to make an analysis possible for the full dataset. This is done by simply concatenating the two datasets.

The second part aims to look at the basic structure of the data and if there are any missing values to take into consideration. Missing values are found in the Location column (see Fig.2 in the Appendix). These will not be handled since the column is not one of the variables that will be used in the sentiment analysis.

To explore the distribution of the data, we are visualizing the number of tweets labeled to each category (see Fig. 3 in the Appendix). The dataset contains the most positive tweets followed by the negatives and then neutral. The extremely negative and extremely positive are the lowest two in count. What is interesting to note is that the positive categories are both higher in count than the negative ones. Implying there were more tweets published with a positive message than a negative during the time frame.

To explore the distribution based on time, a time line was created with the number of tweets per day for the full dataset (see Fig. 4 in the Appendix). A peak can be noticed in the middle/late portion of march.

The 20 most common words in the dataset, before text prepossessing, are printed and shown in Table 2. The most occuring words are not meaningful and do not bring a lot of information for a sentiment analysis of the tweet. Therefore, text preprocessing will be done in order to capture the most meaningful words for the sentiment analysis. The process will be described in the next chapter.

Table 2. The 20 most occuring words before text preprocessing.

Word	Occurring	Word	Occurring	Word	Occurring
the	44210	**#coronavirus**	1326	**with**	6542
to	40974	**is**	12833	**have**	6517
and	25423	**are**	11855	**that**	6468
of	23275	**on**	9893	**prices**	6415
a	19654	**I**	9412	**this**	6412
in	19363	**you**	8490	**food**	6026
for	14590	**at**	8183		

4.3 Data Preprocessing

Data preprocessing is a crucial step in sentiment analysis, as selecting appropriate preprocessing methods allows for an increase in the number of correctly classified cases (Krouska et al., 2016). For the purpose of this study, we followed the Pak & Paroubek (2010) approach by selecting different preprocessing methods, which are described below.

a) Filtering and Standardization

Starting with filtering, unnecessary elements such as URL links, emails, user mentions, orphans, pure numbers and white spaces were removed as they were not relevant to the classification task. In addition, all characters were converted to lowercase.

b) Stop Words Removal and Lemmatization

Frequent words that were lacking contextual meaning were removed at this stage. The reason for this is to allow the model to learn more relevant features. In addition, the tweets were lemmatized to transform all words to their root word. The WordNetlemmatizer() function available in NLTK was used for this purpose.

c) Tokenization

We segmented the tweets by dividing them by spaces and punctuation marks, and created a bag of words.

After applying preprocessing to the dataset, we wanted to see which words are most commonly used in tokenized tweets. The set of these words is shown in Fig. 5 in the Appendix. It is clearly visible that the most frequent words are: covid, grocery, store, supermarket, consumer, food and people. A detailed analysis of topics and words will be described in the topic modeling section.

4.4 Topic Modeling

Topic modeling is one of the most powerful text mining techniques for discovering hidden data and finding relationships between data and text documents. (Jelodar et al., 2018). Blei (2012) described it as a set of algorithms that reveal, discover and annotate the thematic structure in a set of documents. According to Kherwa & Poonam (2019), the most widely used topic modeling algorithm is Latent Dirichlet Allocation (LDA). LDA represents topics as word probabilities and enables the discovery of hidden topics by clustering words based on their co-occurrence in a document (Chuang et al., 2012).

We built LDA based on the gensim package because it has a simple implementation in Python. However, the biggest challenge was to extract good quality topics that are clear and meaningful. To cope with this, we placed great emphasis on proper preprocessing and finding the optimal number of topics.

Apart from the number of topics, the main inputs to LDA are the corpus and the dictionary. Gensim generates a unique identifier for each word in the document, and the corpus maps each identifier with the word frequency. In addition, chunksize, passes, and alpha parameters are needed. By choosing default settings for all these parameters and a number of topics equal to 5 (chosen by trial and error), we obtained a coherence score of 0.3956.

Röder et al. (2015) in their paper 'Exploring the space of topic coherence measures' explained what coherence is and how to measure it. The coherence score measures the suspension and matching of individual words or subsets of words. In the case of the LDA Gensim model, the coherence score is the average coherence score of all n topics created by the model.

Having obtained a reasonably satisfactory consistency score, we checked the dominant theme for each sentence together with the weighting of the topic and keywords (Fig. 6 in the Appendix) and the most representative sentence for each topic (Fig. 7 in the Appendix). To better understand the keywords, a word cloud is created (Fig. 8 in the Appendix) and pyLDAVis is used to visualize the information contained in the topic model (Fig. 9 in the Appendix).

PyLDAvis provides two panels for visualization (Fig. 9 in the Appendix): the panel on the left shows a global view of the topic along with a map of distances between topics, while the right side provides bar charts of terms (Hidayatullah & Ma'arif, 2017). In our case, the pyLDAvis results represent five bubbles, where the larger the bubble, the more popular the topic. To prove that the model is good, the bubbles should be scattered throughout the graph and not overlap, which we have achieved.

4.5 Data Analytics

4.5.1 Gated Recurrent Unit

The GRU Bidirectional model's first layer is an embedding level, it takes the integer encoded tweet and finds an embedding vector for each word-index. As the model trains these vectors are learned and adds another dimension to the output. The Bidirectional layer connects two hidden layers from opposite directions to the same output. The Global average pooling layer passes a fixed-length vector for each example to the dense layer with activation 'ReLu', which is a very common activation function in deep learning. A dropout layer is defined to hinder overfitting and the value chosen is 0.5. The last layer is a softmax layer where the five possible classes are defined.

4.5.2 Long Short-Term Memory

The LSTM-model consists of an embedding layer, a pooling layer, a dense-layer with a ReLU activation function, a dropout layer with a value of 0.2, and lastly a dense layer with a softmax activation function to categorize the output into one of the five classes.

4.5.3 Logistic Regression and Multinomial Naïve Bayes

For both Logistic Regression and Multinomial Naïve Bayes we use vectorization and future extraction by CountVectorizer or TfidfTransformer. While the CountVectorizer() function converts texts to a token count, the TfidfTransformer() function converts to a weighted representation. To deal with class imbalance, the upsampling technique SMOTE() is used.

Models are specified and trained by executing a gridsearch on defined pipelines containing: Countvectorizer(), TfidfTransformer(), SMOTE() as well as MultinomialNB() or LogisticRegression(). For model selection, the GridSearchCV() function finds model parameters by cross-validated grid-search. This result is used to select the best model and then the performance is measured on the test set.

5. Results

For our results, we will mainly focus on the F1-score. There is no especially detrimental effect to either having a false positive or a false negative, therefore it makes sense to choose an accuracy metric that strikes a balance between precision and recall. Nonetheless, the scores are very similar for all three accuracy metrics.

Table 3. Performance of the models

Model	F1-Score	Precision	Recall
GRU	0.69	0.71	0.69
LSTM	0.69	0.69	0.68
Logistic Regression	0.62	0.62	0.62
Multinomial Naïve Bayes	0.47	0.48	0.46

When looking at the accuracy scores of our different models, it is apparent that the GRU and LSTM models perform better than the Logistic Regression and Multinomial Naive Bayes models. The two RNN models both achieve an overall F1-score of 0.69 and they are better at predicting every single text sentiment tabel (Figure 10). The GRU and LSTM models especially excel when classifying neutral tweets, while the Logistic Regression model and the Multinomial Naive Bayes models are best at predicting the neutral and extremely positive tweets respectively.

While the weighted average F1-scores are similar between the GRU and the LSTM model, Fig. 10 reveals that the GRU model is superior when predicting neutral and extremely positive tweets, while the LSTM model only outperforms the GRU model when classifying positive tweets.

5.1 Valuable Outcomes

The models can be used to determine the public sentiment of COVID-19. By analyzing Tweets, a politician or law-maker can quickly grasp the public opinion of new laws or restrictions related to COVID-19. This is likely to be a much faster option than waiting for the results of an opinion poll and is therefore highly useful in a constantly changing setting like a pandemic. Another useful case is to obtain a general picture of mental health among the younger generation. Around 62% of Twitter's users are under the age of 34 (Statista, 2022b), which makes it a perfect environment to take the temperature of the younger generations. The catch is, however, that the model has been trained on COVID-19 related tweets and therefore might only be relevant for a situation similar to this recent pandemic.

The models might also be useful in determining sentiment for future diseases as many of the words used to label the COVID-19 Tweets are also likely to be relevant for other diseases. Training and developing a model that is capable of correctly labeling Tweets related to a global disease might therefore prove to be a useful tool to have in the future for governments.

5.2 Topic Modelling

To better understand the results of Topic modeling, a word cloud with 4 main themes was created (Fig. 8 in the Appendix). From these, it can be seen that people focused on supply, demand, and shopping topics in their tweets. This can be seen in the first topic with keywords such as groceries, supermarket and shop, and the third topic with keywords such as food, buy and stock. In addition, theme number 0 refers to protective measures such as hand sanitizer and masks, and the second theme reveals people's concerns about the crisis and businesses.

6. Discussion

6.1 Comparison of the Models

The different Text Classifiers, Naïve Bayes and Logistic Regression, are some of the most common statistical models to apply in machine learning due to its easy implementation and adequate results. But examining all the evaluation scores of this dataset, it is apparent that Logistic Regression outperforms the Naïve Bayes model, as its score is significantly higher. Arguably, this depends on the dataset's size. As seen in the work of Prabhat and Khullar (2017), the Multinomial Naïve Bayes method also performed worse than the Logistic Regression model on a dataset that was even smaller than ours. The performance gap between them was smaller, however, which indicates that Naïve Bayes is a more appropriate method on a smaller dataset and is therefore not achieving good evaluation scores with our dataset. On the other hand, Logistic Regression is not applicable on small amounts of data as it becomes a more generalized model, which could explain the better performance of Logistic Regression for our selected data.

The Naïve Bayes is proven to be not as suitable for sentiment analysis on these tweets compared to Logistic Regression because the method does not take into account the words' relationship and solely view them as a bag of words. This makes the method weak to the emergence of new words that the method did not spot in the training data. Since tweets are very unstructured and written in different personal styles, the Naïve Bayes method is therefore not suitable for performing sentiment analysis in tweets, as the tweets' content can not be guaranteed to be included in the training data. As the Logistic Regression generated a good F1-score and does not possess the limitation as stated above, as well as its appropriateness to data analysis, the Logistic Regression is a better performing technique for this type of sentiment analysis.

The F1-score indicates that the Bi-GRU model is slightly outperforming the LSTM model on the COVID-19 Twitter dataset. This entails that this model is the optimal one on the specific data. But only with very little difference to the LSTM. The LSTM is a more complex model then the GRU with the differences in structure of having three gates compared to the GRUs two gates and therefore less training parameters. Arguably, the relatively small dataset, used in this project, could be a better fit for the GRU model since it is a simpler RNN. However, a LSTM model would be prefered on a larger dataset than the one used in this project to achieve high accuracy scores, although it would require more time to execute the training.

Another factor to the results that should be considered is the risk of overfitting. Overfitting the model could generate a better accuracy score than what the model can actually accomplish. Meaning that the model generates a good prediction on the training data but not on new data.

To summarize, the RNN models outperform the less complex Naïve Bayes and Logistic Regression when executing a sentiment analysis on the COVID-19 Twitter data. Depending on the size of the selected dataset, either LSTM or GRU should be used to achieve good evaluation scores.

6.2 Topic Modelling

There is no unique way to determine whether a coherence score is good or poor. The score and its value depend on the data on which it was computed. The higher the coherence score, the better (Mohammed and Al-augby, 2020). Usually the higher number of subjects increases this score, but this is not always the case. In our project, the consistency score is almost 0.4, which can be considered a good result, but there is still a lot of room for improvement. Definitely keywords in topics should not be repeated. If we could achieve this, the coherence would increase. Moreover, with a larger dataset and higher number of topics we would probably also see the improvement.

6.3 Areas of Improvements & Future Work

To achieve models with higher accuracy, a larger dataset would be preferable to use. An LSTM model or the GRU model could possibly perform even better predictions but would also require more computing power and simply more time to execute the training of the model. As the complexity is more extensive and layers are increased, the LSTM model could possibly outperform the GRU on a larger dataset. This comparison would be interesting to execute in possible future work.

As mentioned the risk of overfitting is a limitation in the project that could be remedied by using methods such as less epochs or cross-validation. This could also provide information to be used for fine tuning the parameters and achieve better predictions through better fitted models.

The dataset entails more information that could be uncovered in future projects. The Location column in the set is one that could give a lot of insights to the different sentiments of the tweets if grouped. The problem to overcome is finding a solution to assort the non-standardized locations the users themselves have chosen. If achieved, this opens up the possibility of comparing countries, states, or cities and provides insights to the connection between the general sentiment and societal structures or characteristics.

To improve topic modeling and obtain a higher coherence value, other topic modeling methods such as Machine Learning for Language Toolkit (MALLET) or Latent Semantic Analysis (LSA) could be applied. MALLET is considered to be a model that often provides better quality subjects and can also be easily built using the Gensim package. Furthermore, the model could be enhanced with an optimisation algorithm that would find the optimal number of topics and pick the one with the highest coherence score.

7. Conclusion

This project performs four models from the statistical category of sentiment analysis and a topic modeling of the tweets. All models are well known and functional methods to use when performing sentiment analysis on subjective data, such as tweets. But they are suitable for different types of data scope and features. After comparing the models, the conclusion drawn is that the Text Classifiers are easier to implement, however we can see that RNN models are more accurate and contribute better performance. Depending on the data size, LSTM is more appropriate for larger and complex sets and GRU is sufficient enough to be used on smaller ones. These models achieve good evaluation scores for classifying the tweet's sentiments. The topic modeling that has been done in the project is displaying the most discussed topics during 2020.

To conclude, the RNN models performance within sentiment analysis and topic modeling provides a good interpretation of the Twitter data as we could take part of the general sentiment and discussion topics on COVID-19 related tweets in 2020. The level of success of this project gives us an optimistic mindset that we have built a well functioning tool to perform sentiment analysis and topic modeling on COVID-19 Twitter data.

References

Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. *IJCSNS Int. J. Comput. Sci. Netw. Secur*, 19(3), 62.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In *A practical guide to sentiment analysis* (pp. 1-10). Springer, Cham.

Chuang, J., Manning, C. D., & Heer, J. (2012, May). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces* (pp. 74-77).

Colnerič, N., & Demšar, J. (2018). Emotion recognition on twitter: Comparative study and training a unison model. IEEE transactions on affective computing, 11(3), 433-446.

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: a comparative study on different approaches. Procedia Computer Science, 87, 44-49.

Dubey, A. D. (2020). Twitter sentiment analysis during COVID-19 outbreak. Available at SSRN 3572023.

Géron, A. (2019). H*ands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, Inc.

Hidayatullah, A. F., & Ma'arif, M. R. (2017, November). Road traffic topic modeling on Twitter using latent dirichlet allocation. In 2017 international conference on sustainable information engineering and technology (SIET) (pp. 47-52). IEEE

Khan, R., Shrivastava, P., Kapoor, A., Tiwari, A., & Mittal, A. (2020). Social media analysis with AI: sentiment analysis techniques for the analysis of twitter covid-19 data. Critical Rev, 7(9), 2761-2774.

Krouska, A., Troussas, C., & Virvou, M. (2016, July). The effect of preprocessing techniques on Twitter sentiment analysis. In 2016 7th international conference on information, intelligence, systems & applications (IISA) (pp. 1-5). IEEE.

Lwin, M. O., Lu, J., Sheldenkar, A., Schulz, P. J., Shin, W., Gupta, R., & Yang, Y. (2020). Global sentiments surrounding the COVID-19 pandemic on Twitter: analysis of Twitter trends. JMIR public health and surveillance, 6(2), e19447.

Mohammed, S. H., & Al-augby, S. (2020). Lsa & lda topic modeling classification: Comparison study on e-books. Indonesian Journal of Electrical Engineering and Computer Science, 19(1), 353-362.

Nemes, L., & Kiss, A. (2021). Social media sentiment analysis based on COVID-19. Journal of Information and Telecommunication, 5(1), 1-15.

Prabhat, A., & Khullar, V. (2017). Sentiment classification on big data using Naïve Bayes and logistic regression. In International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-5). IEEE.

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408).

Sachin, S., Tripathi, A., Mahajan, N., Aggarwal, S., & Nagrath, P. (2020). Sentiment analysis using gated recurrent neural networks. SN Computer Science, 1(2), 1-13.

Statista. (2022a). Number of Twitter users worldwide from 2019 to 2024. Available Online: https://www.statista.com/statistics/303681/twitter-users-worldwide/ [accessed: 2022-05-18]

Statista. (2022b). Distribution of Twitter users worldwide as of April 2021, by age group. Available Online:

https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/ [accessed: 2022-05-23]

Talevi, D., Socci, V., Carai, M., Carnaghi, G., Faleri, S., Trebbi, E., & Pacitti, F. (2020). Mental health outcomes of the CoViD-19 pandemic. Rivista di psichiatria, 55(3), 137-144.

Appendix

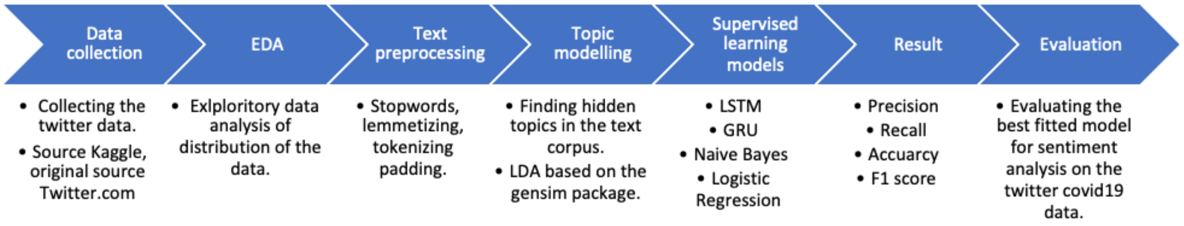
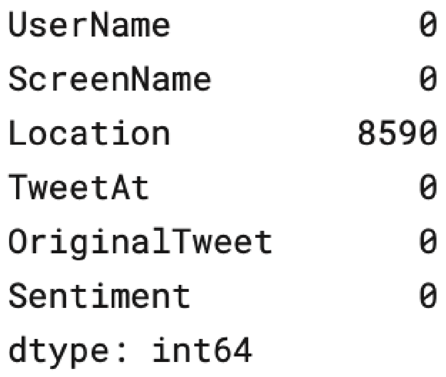
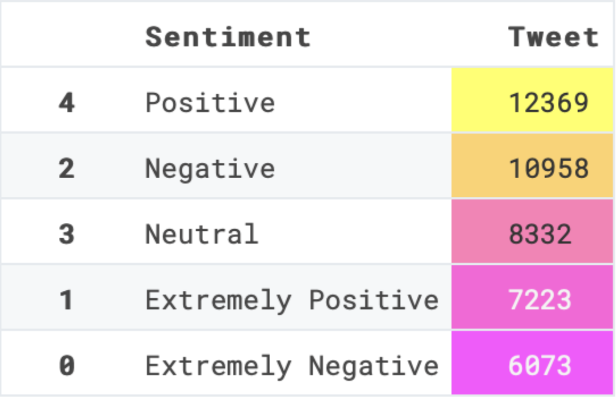

Data collection
• Collecting the twitter data.
• Source Kaggle, original source Twitter.com
EDA
• Exlploritory data analysis of distribution of the data.
Text preprocessing
• Stopwords, lemmetizing, tokenizing padding.
Topic modelling
• Finding hidden topics in the text corpus.
• LDA based on the gensim package.
Supervised learning models
• LSTM
• GRU
• Naive Bayes
• Logistic Regression
Result
• Precision
• Recall
• Accuarcy
• F1 score
Evaluation
• Evaluating the best fitted model for sentiment analysis on the twitter covid19 data.

Fig. 1 Overview of project process



```
UserName            0
ScreenName          0
Location         8590
TweetAt             0
OriginalTweet       0
Sentiment           0
dtype: int64
```

Fig. 2 Missing values



	Sentiment	Tweet
4	Positive	12369
2	Negative	10958
3	Neutral	8332
1	Extremely Positive	7223
0	Extremely Negative	6073

Fig. 3 Data distribution

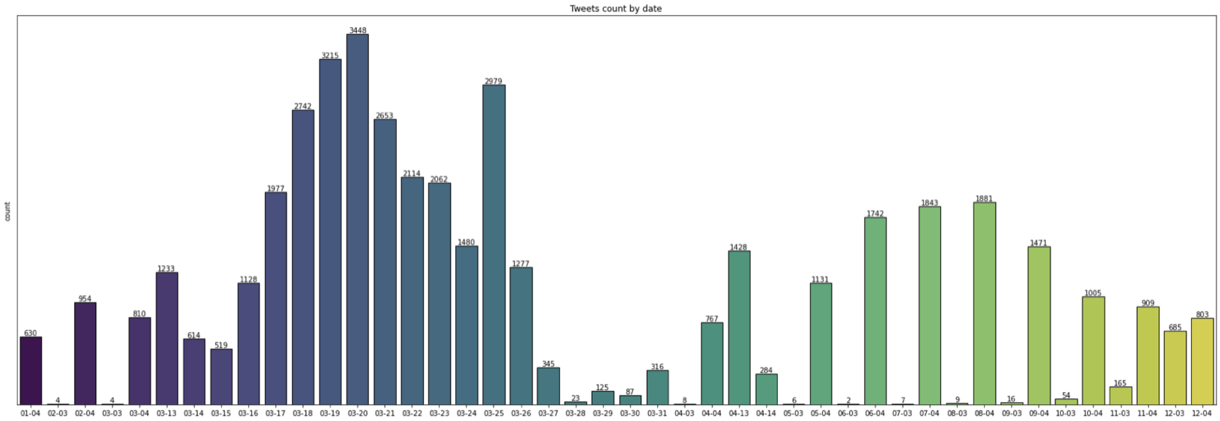


Fig. 4 Number of Tweets by date

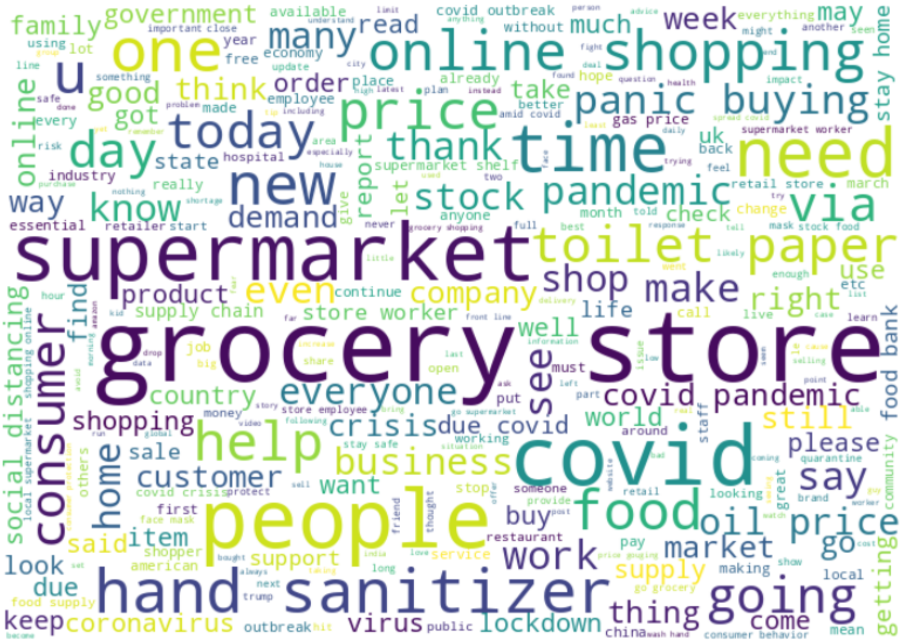
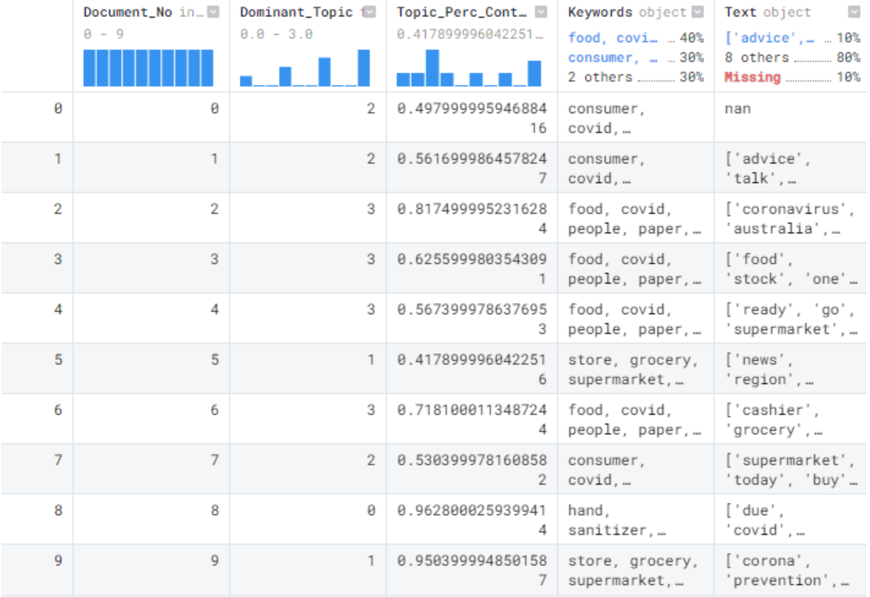
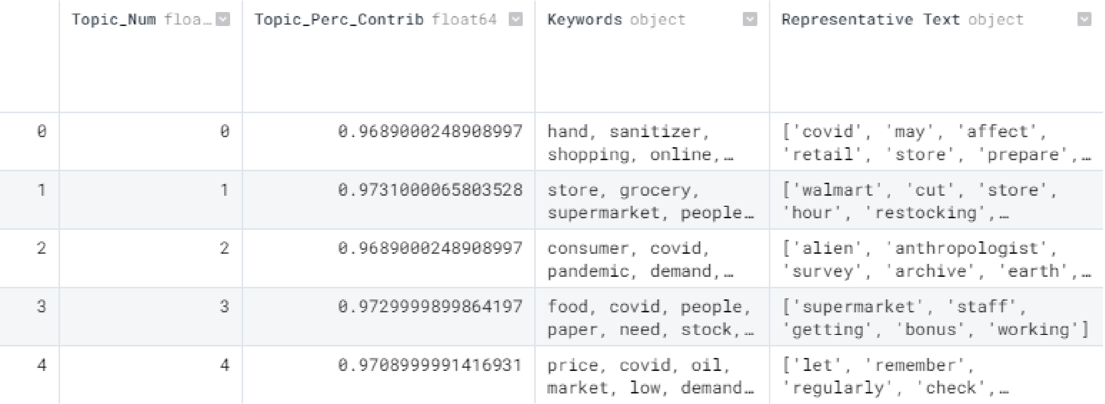


Fig. 5 Word Cloud of Tokenized Tweets



	Document_No in…	Dominant_Topic f…	Topic_Perc_Cont…	Keywords object	Text object
	0 - 9	0.0 - 3.0	0.417899996042251…	food, covi… … 40% consumer, … … 30% 2 others … 30%	['advice',… … 10% 8 others … 80% Missing … 10%
0	0	2	0.49799999594688416	consumer, covid,…	nan
1	1	2	0.5616999864578247	consumer, covid,…	['advice', 'talk',…
2	2	3	0.8174999952316284	food, covid, people, paper,…	['coronavirus', 'australia',…
3	3	3	0.6255999803543091	food, covid, people, paper,…	['food', 'stock', 'one'…
4	4	3	0.5673999786376953	food, covid, people, paper,…	['ready', 'go', 'supermarket',…
5	5	1	0.4178999960422516	store, grocery, supermarket,…	['news', 'region',…
6	6	3	0.7181000113487244	food, covid, people, paper,…	['cashier', 'grocery',…
7	7	2	0.5303999781608582	consumer, covid,…	['supermarket', 'today', 'buy'…
8	8	0	0.9628000259399414	hand, sanitizer,…	['due', 'covid',…
9	9	1	0.9503999948501587	store, grocery, supermarket,…	['corona', 'prevention',…

Fig. 6 Dominant topic and its percentage contribution in each document



	Topic_Num floa…	Topic_Perc_Contrib float64	Keywords object	Representative Text object
0	0	0.9689000248908997	hand, sanitizer, shopping, online,…	['covid', 'may', 'affect', 'retail', 'store', 'prepare',…
1	1	0.9731000065803528	store, grocery, supermarket, people…	['walmart', 'cut', 'store', 'hour', 'restocking',…
2	2	0.9689000248908997	consumer, covid, pandemic, demand,…	['alien', 'anthropologist', 'survey', 'archive', 'earth',…
3	3	0.9729999899864197	food, covid, people, paper, need, stock,…	['supermarket', 'staff', 'getting', 'bonus', 'working']
4	4	0.9708999991416931	price, covid, oil, market, low, demand…	['let', 'remember', 'regularly', 'check',…

Fig. 7 The most representative sentence for each topic

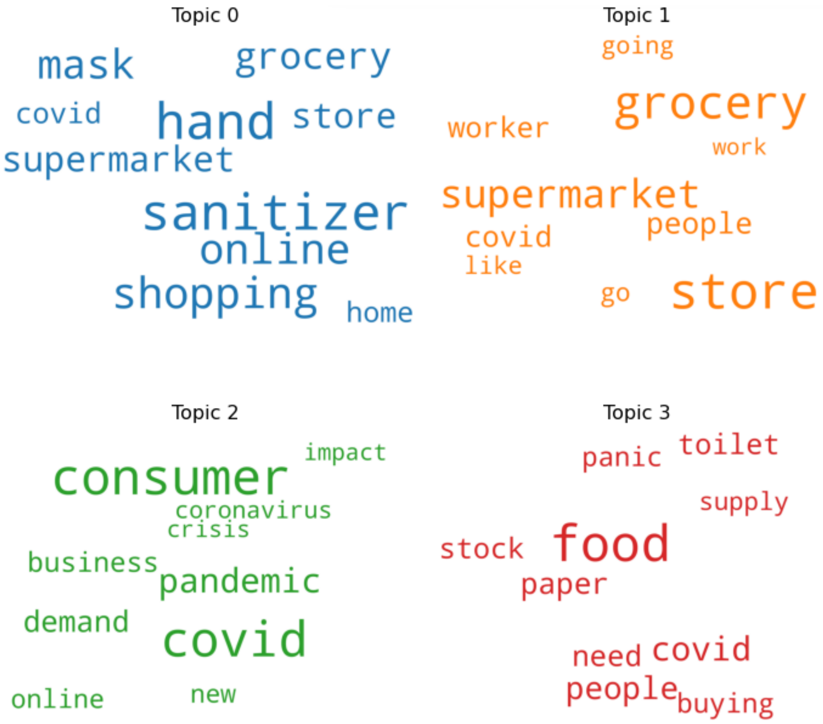
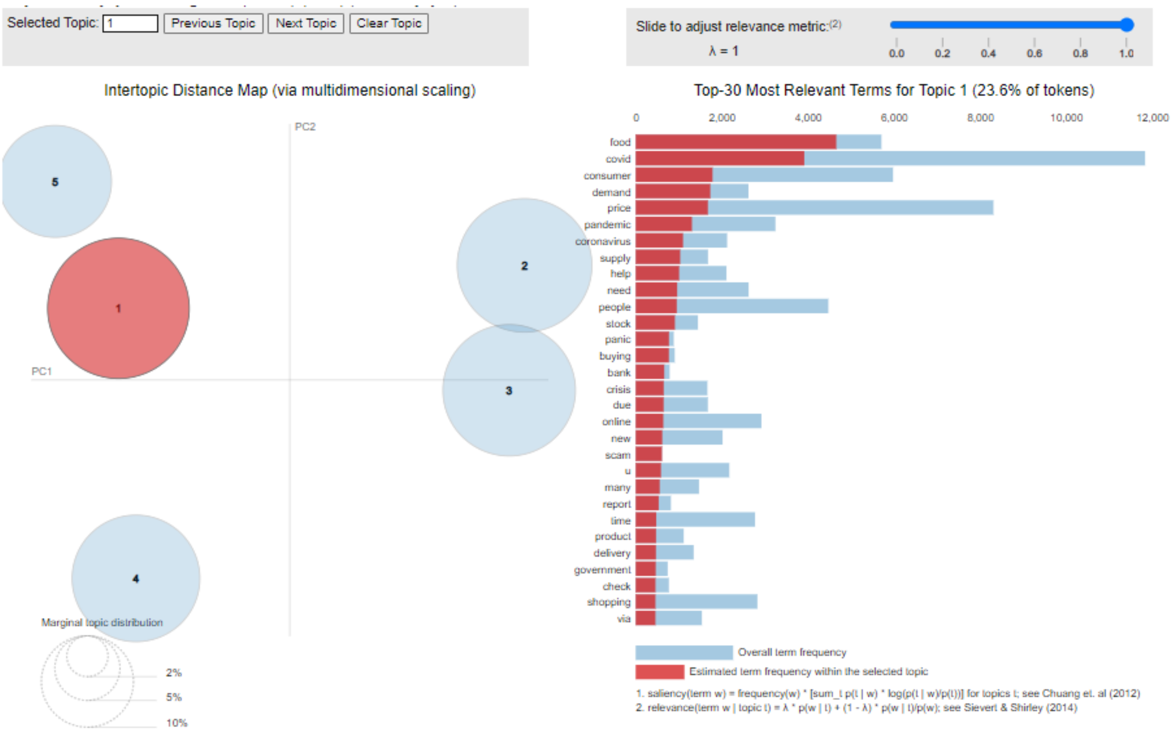
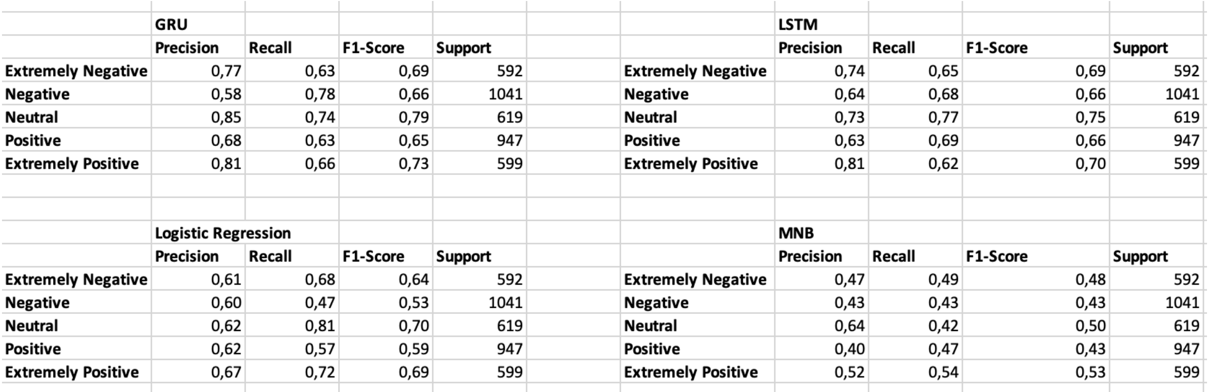


Fig. 8 Word cloud visualization for n=4 topics



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 9 pyLDAVis



	GRU						LSTM			
	Precision	Recall	F1-Score	Support			Precision	Recall	F1-Score	Support
Extremely Negative	0,77	0,63	0,69	592		**Extremely Negative**	0,74	0,65	0,69	592
Negative	0,58	0,78	0,66	1041		**Negative**	0,64	0,68	0,66	1041
Neutral	0,85	0,74	0,79	619		**Neutral**	0,73	0,77	0,75	619
Positive	0,68	0,63	0,65	947		**Positive**	0,63	0,69	0,66	947
Extremely Positive	0,81	0,66	0,73	599		**Extremely Positive**	0,81	0,62	0,70	599
	Logistic Regression						**MNB**			
	Precision	**Recall**	**F1-Score**	**Support**			**Precision**	**Recall**	**F1-Score**	**Support**
Extremely Negative	0,61	0,68	0,64	592		**Extremely Negative**	0,47	0,49	0,48	592
Negative	0,60	0,47	0,53	1041		**Negative**	0,43	0,43	0,43	1041
Neutral	0,62	0,81	0,70	619		**Neutral**	0,64	0,42	0,50	619
Positive	0,62	0,57	0,59	947		**Positive**	0,40	0,47	0,43	947
Extremely Positive	0,67	0,72	0,69	599		**Extremely Positive**	0,52	0,54	0,53	599

Fig. 10 Scores of the different models.