# CATEGORIZING 2019-N-COV TWITTER HASHTAG DATA BY CLUSTERING

Koffka Khan[1] and Emilie Ramsahai[2]

[1]Department of Computing and Information Technology, The University of the West Indies, St. Augustine Campus, Trinidad
[2]UWI School of Business & Applied Studies Ltd (UWI-ROYTEC), 136-138 Henry Street, 24105 Port of Spain, Trinidad and Tobago

***ABSTRACT***

*Unsupervised machine learning techniques such as clustering are widely gaining use with the recent increase in social communication platforms like Twitter and Facebook. Clustering enables the finding of patterns in these unstructured datasets. We collected tweets matching hashtags linked to COVID-19 from a Kaggle dataset. We compared the performance of nine clustering algorithms using this dataset. We evaluated the generalizability of these algorithms using a supervised learning model. Finally, using a selected unsupervised learning algorithm we categorized the clusters. The top five categories are Safety, Crime, Products, Countries and Health. This can prove helpful for bodies using large amount of Twitter data needing to quickly find key points in the data before going into further classification.*

***KEYWORDS***

*Unsupervised machine learning, clustering, Twitter, 2019-nCoV, hashtags, Kaggle, supervised, classification*

## 1. INTRODUCTION

Twitter [43] is a micro-blogging service which has millions of users from around the world. It enables users to post and exchange 140-character-long messages, also called tweets. Using a wide array of Web-based services, tweets can be published by sending e-mails, SMS text messages and directly from smartphones. Twitter thus enables the dissemination of information to a wide number of people in real time. This makes it an ideal environment for disseminating breaking news directly from the source of news and/or event location.

Before a related keyword or expression in their message, people use the hashtag [26] symbol (#) to categorize Tweets and make them display them more quickly in Twitter search. Clicking or tapping on a hashtagged word will show you other Tweets that have the hashtag. Hashtags can be included in a Tweet at any place. Thus, a hashtag is used on Twitter to index keywords or topics. Hashtagged words, which become very popular, are often trending topics.

Unsupervised learning algorithms [7] [20] has seen a recent spurt in usage with increasing advances in computing technology. It is a sub-field of machine learning. The machine simply receives inputs in unsupervised learning but does not obtain supervised target outputs or rewards from its environment. It is possible to establish a formal structure for unsupervised learning based on the notion that the purpose of the machine is to construct representations of the inputs that can be used for decision making, to anticipate potential inputs and to communicate inputs effectively to another system. In a way, unsupervised learning can be seen as identifying correlations over

and above what should be called mere unstructured noise in the results. Clustering [22] and dimensionality reduction are two very basic textbook examples of unsupervised learning.

Twitter has been used to monitor patterns and distribute health knowledge over the course of virus epidemics. The recent 2019-nCoV [41] is no exception. Researchers in [17] used Twitter and web news mining to predict COVID-19 outbreak. To quantify and understand early changes in Twitter activity, content, and sentiment [28], [33], [36] about the COVID-19 epidemic used a large volume of Twitter data [32]. To understand Twitter users' discussions and reactions about the COVID-19, researchers in [46] used various machine learning techniques. [1] aimed to identify the main topics posted by Twitter users related to the COVID-19 pandemic. Research offer insights based on theory to help explain and predict these behaviors and associated outcomes in order to inform future research and marketing practice using social media data including Twitter [21]. Researcher [16] call for collaboration amidst the growing mountain of daily data across PubMed, Twitter, Google Scholar and the World Health Organization.

We collected tweets matching hashtags linked to COVID-19 from a Kaggle dataset [42]. We compared the performance of nine clustering algorithms using this dataset. We evaluated the generalizability of these algorithms using a supervised learning model. Finally, using a selected unsupervised learning algorithm we categorized the clusters. This can prove helpful for bodies using large amount of Twitter data needing to quickly find key points in the data before going into further classification [3].

In section two we investigate previous work related to our study. Then we present a brief introduction to the nine unsupervised machine learning clustering models used to categorize 2019-nCoV hashtags in section three. In section four the method used to categorize 2019-nCoV hashtags is given. We show the results in section five. As part of our results we compare the performance of these methods as we believe it would better guide further research work in developing clustering techniques to combat future pandemics. In addition, our results can aid disaster relief bodies to quickly sift through huge amounts Twitter data to accurately capture meaningful categories ‘on-the-fly.’ Finally give our conclusion in section six.

## 2. RELATED WORK

Twitter data has been used extensively since the start of the 2019-nCoV pandemic, for example in predicting the onset [18] and tracking social media discourse [9]. Another area of research impacted by the use of Twitter data is sentiment analysis [6]. Research using sentiment analysis on tweets indicated that while majority of the people throughout the world are taking a positive and hopeful approach, there are instances of fear, sadness and disgust exhibited worldwide [11]. Also, in [31] 126,049 tweets from 53,196 unique users were evaluated. Researchers note that the hourly number of COVID-19-related tweets starkly increased from January 21, 2020 onward and that nearly half (49.5%) of all tweets expressed fear and nearly 30% expressed surprise. Researchers [45] showed that trust for the authorities remained a prevalent emotion using Twitter data analysed consisting of about 1.8 million Tweets messages related to coronavirus collected from January 20th to March 7th, 2020. A total of salient 11 topics are identified and then categorized into 10 themes, such as "cases outside China (worldwide)," "COVID-19 outbreak in South Korea," "early signs of the outbreak in New York," "Diamond Princess cruise," "economic impact," "Preventive/Protective measures," "authorities," and "supply chain" using unsupervised machine learning techniques. Researchers [35] collected tweets from the users who shared their location as ‘Nepal’ between 21st May 2020 and 31st May 2020. They observed that while majority of the people of Nepal are taking a positive and hopeful approach, there are instances of fear, sadness and disgust exhibited too. Researchers [24] performed a Twitter search using 14 separate common hashtags and keywords associated with the COVID-19 outbreak. Then, in

comparison to checked and peer-reviewed tools they analyzed and tested individual tweets for misinformation. Finally, they utilized descriptive statistics to compare words and hashtags, and describe the characteristics of individual tweets and account. They found that the keyword “COVID-19” had the lowest rate of misinformation and unverifiable information, while the keywords “#2019_ncov” and “Corona” were associated with the highest amount of misinformation and unverifiable content respectively. Researchers found that two-thirds (66.1%) of the Instagram users use "COVID-19", and "coronavirus" hashtags to disperse the information related to COVID-19 [38]. Other work in [12] investigated the temporal tweeting dynamic and the Twitter users involved in the online discussions around COVID-19-related research. They observed that throughout the course of time, a shift in the direction of the Twitter discussions can be noted, from initial exposure to virological and scientific science to more realistic issues such as new therapies, policy countermeasures, welfare measures, and more recent effects on the economy and culture. Though many researchers attempt to categorize the actual twitter data, we did not find any work on finding patterns in the hashtags to elicit an initial overview of the existing tweets which can be then classified further if desired. Our approach does this in an attempt to overcome the huge sizes of these social media datasets.

## 3. CLUSTERING MODELS

Clustering generally uses iterative techniques to group cases into clusters in a dataset which contain similar characteristics. These groupings are useful to explore data, identify anomalies in the data, and ultimately make predictions. Models of clustering can also help you identify relationships in a dataset that you may not logically derive from browsing or simply observing. For these reasons, clustering is often used to explore the data in the early stages of machine learning tasks, and to discover unexpected correlation.

### 3.1. K-means

The K-means algorithm [25] starts with an initial set of randomly selected centroids, which serve as starting points for each cluster, when processing the training data, and applies the Lloyd 's algorithm to iteratively refine the centroid locations. Cluster assignment is effected by calculating the distance between each cluster 's new case and the centroid. Every new case is allocated with the nearest centroid to the cluster.

### 3.2. DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) [5] is a non-parametric density-based clustering algorithm: given a set of points in some space, it aggregates points that are closely packed together (points with many nearby neighbours), marking outliers that lie in low-density regions (whose closest neighbors are too far away).

### 3.3. Spectral clustering

Spectral clustering (SC) models [44] use the data similarity matrix spectrum (eigenvalues) to obtain a decrease in dimensionality before clustering in smaller dimensions. The similarity matrix is provided as an input and consists of a quantitative evaluation of the relative similarity of each pair of points within the dataset.

### 3.4. Agglomerative clustering

Agglomerative clustering (AC) [14] is a "bottom-up" approach: each observation starts in its own cluster, merging pairs of clusters as one moves the hierarchy upwards. A measure of the dissimilarity between sets of observations is necessary in order to determine which clusters should be combined. As a function of the pairwise distances between observations, a linkage criterion determines the distance between sets of observations.

### 3.5. Gaussian mixtures

Gaussian mixture (GM) models [29] are a probabilistic model within an overall population to describe naturally distributed subpopulations. Two types of measurements, the weights of the mixture component and the means and variances/covariances of the component are parameterized by a Gaussian mixture model.

### 3.6. BIRCH

The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [47] algorithm takes a set of data points as its input, represented as real-valued vectors, and a necessary number of clusters. Due to its ability to find a good clustering solution with only one data scan, BIRCH is particularly suitable for very large data sets or for streaming data.

### 3.7. Mini Batch

The key principle of the Mini Batch K-means algorithm [34] is to use small random batches of fixed size data, so that they can be stored in memory. Each iteration obtains and uses a new random sample from the dataset to update the clusters, and this is repeated until convergence.

### 3.8. Mean-Shift

Mean-Shift [2] is a procedure to locate the maxima of a density function given the discrete data sampled from that function. It functions by placing a kernel in the data set at each point. Mean shift takes advantage of this idea by considering what the points would do if they all climbed up hill to the closest kernel density estimate (KDE) surface top.

### 3.9. OPTICS

Ordering points to identify the clustering structure (OPTICS) [4] is an algorithm for the identification of clusters based on density in spatial data. OPTICS therefore outputs the points in a particular order, annotated at their smallest distance of reachability.

## 4. METHODS

We used the Python version 3.8 programming language to run experiments. Text data requires special preparation before use for modeling can begin. To remove words, the text must be parsed, called tokenization. Then the words must be encoded as integer or floating point values for use as input to a machine learning algorithm, called extraction (or vectorization). We use the scikit-learn library to perform both tokenization and hashtag data extraction functionality.

One difficulty with simple counts is that certain words like "the" will occur several times, and their large counts in the encoded vectors will not be very important. An alternative is to calculate

the word frequencies, and TF-IDF (Term Frequency–Inverse Document) [15] is by far the most popular method. The TF-IDF are the components of each word assigned to the resulting scores. The term frequency describes the frequency at which a given word appears inside a text. Inverse Document Frequency scales down words that occur a great deal across files. Thus TF-IDF are word frequency scores that attempt to highlight more interesting words, e.g. frequent in a file but not across files. The Tfidf Vectorizer will tokenize documents, learn the vocabulary and reverse weighting of document frequencies, and allow you to encode new documents/files. A Tfidf Transformer is used for measuring reverse text frequencies and starting text encoding. The scores are normalized to values between 0 and 1 and, as with most machine learning algorithms, the encoded document vectors can then be used directly.

We employ three metrics for evaluating the performance of clustering models: Silhouette [37], Calinski-Harabasz [30] and Davies-Bouldin [30]. Silhouette refers to a method for interpreting and validating consistency within data clusters. It is a function of how close an entity is to its own (cohesion) cluster relative to other (separation) clusters. The score varies from −1 to +1, where a high value means that the object is well aligned with its own cluster and poorly suited to neighboring clusters. If most objects have a high value, then the setup for the clustering is correct. However, if many points have a low or negative value then there may be too many or too few clusters in the clustering configuration.

The Calinski-Harabasz score [30] is defined as the ratio between the dispersion within a cluster and the dispersion between a cluster. This procedure ensures that the number of potential splits is effectively reduced. The approach can be generalized to a dichotomous division but is well suited to any number of clusters and for a global division. The Calinski-Harabasz Index should be greatest at the optimal clustering size.

The Davies-Bouldin score [30] is an internal assessment scheme where the analysis of how good the clustering was performed is achieved using the underlying quantities and characteristics of the dataset. A lower value will mean that the clustering is better because of the way it is defined, as a function of the ratio of the cluster scatter within, to the separation between clusters. It happens to be the average similarity, averaged over all clusters, between each cluster and its most similar one. This affirms the idea that no cluster has to be similar to another, and thus the best clustering scheme minimizes the Davies–Bouldin index.

Gradient Descent is a common technique of optimization in machine learning and deep learning, which can be used for most, if not all, learning algorithms. The slope of a function is a gradient. It measures the degree of variability of a variable in response to other variable changes. Gradient Descent is a convex function whose output is the part derivative of a set of its input parameters. The steeper the slope the greater the gradient. Stochastic gradient descent (SGD) [8] is an iterative method to optimize an objective function with proper smoothness (e.g. differentiable or sub-differentiable) properties. It can be considered as a stochastic approximation of gradient descent optimization, since it substitutes the true gradient (calculated from the entire data set) with an average of it (calculated from a randomly chosen data subset). For each iteration, a few samples are chosen randomly in Stochastic Gradient Descent, instead of the entire data set. The sample is shuffled at random and selected to perform the iteration. The cost function gradient of a single sample is calculated at each iteration. The Stochastic Gradient Descent (SGD) classifier [19] essentially incorporates a simple SGD learning routine.

There are four Outcomes of Binary Classification [39]:

- True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is positive and the value of predicted class is positive.
- True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is negative and value of predicted class is negative.

- False Positives (FP) – When actual class is negative and predicted class is positive.

- False Negatives (FN) – When actual class is positive but predicted class in negative.

Accuracy is simply a ratio of correctly predicted observation to the total observations (Accuracy = TP+TN/TP+FP+FN+TN). The number of true positives divided by the number of true positives plus the number of false positives is known as precision (Precision = TP/TP+FP). It measures the ability of a classification model to identify only the relevant data points. Recall (Recall = TP/TP+FN) is the ability of a classification model to identify all relevant instances. It is the ratio of correctly predicted positive observations to the all observations in actual class. The F1 score is a single metric that combines recall and precision using the harmonic mean (F1 Score = 2*(Recall * Precision) / (Recall + Precision)).

The procedure we carried out followed these steps:

1. Read hashtags
2. Convert to string, vectorize and transform hashtags
3. Select clustering model
4. Fit model to hashtag data
5. Predict and obtain labels from specified model
6. Evaluate model using Silhouette, Calinski-Harabasz and Davies-Bouldin metrics
7. Test and train model using a Stochastic Gradient Descent (SGD) Classifier
8. Print the accuracy, precision, recall and F1 score SGD Classification Report
9. Categorize the cluster centroids by obtaining top terms per cluster

The last and very important step in our method is to categorize the clusters by obtaining the top terms per cluster. We used the TF-IDF vectorizer, thus "features" are the words in a given hashtag document (and each document is its own vector). Thus, when the document vectors are clustered, each "feature" of the centroid represents the relevance of that word to it. The "word" (in vocabulary) is equal to the "feature" (in the vector space) which is equal to the "column" (in the centroid of the matrix). We get the mapping of column index to the word it represents and convert each centroid into a sorted (descending) list of the columns most "relevant" (highly valued) in it, and hence the words most relevant since words are equal to columns. Thus, in essence we are sorting each centroid in descending order of the features/words most valued in it, then mapping those columns back to their original words.

## 5. RESULTS

We use a 2019-nCoV dataset from Kaggle [42] which was obtained from the Johns Hopkins University [27]. The dataset was made available from January 22nd, 2020 but was accessed on March 30th 2020. This dataset has regular level details from 2019-nCoV on the number of infected cases, deaths and recovery. The dataset has 1086 cases with nineteen (19) features. Some features were:

- Observation Date - Date of the observation in MM/DD/YYYY
- Province/State - Province or state of the observation (Could be empty when missing)
- Country/Region - Country of observation
- Last Update - Time in UTC at which the row is updated for the given province or country. (Not standardised and so please clean before using it)
- Confirmed - Cumulative number of confirmed cases till that date
- Deaths - Cumulative number of deaths till that date
- Recovered - Cumulative number of recovered cases till that date

The dataset also contained a hashtags.CSV file which contained 26,833,314 hashtags collected over the period starting on March 13th 2020 and ending on March 28th 2020. There were two fields status_id and the hashtag. An example of the hashtag was 'SocialDistancing.' We observed that in a dataset of this size it would take huge amounts of computing power to gather the main points in the Twitter dataset. Thus, we focused on the hashtags as a starting point which is still considerably large but we experimented to explore the possibility of this 'smaller' sample of the dataset (that is, not containing the actual tweets) giving insights to a good representation or categorization of what the main elements of the tweets contained.

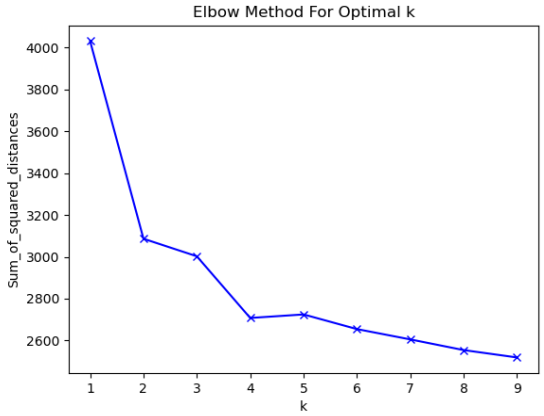


Figure 1. Elbow method using K-means.

For a range of values for k (say from 1-10) the elbow method runs k-means clustering on the dataset and then for each value of k calculates an average score for all clusters. In order to determine the optimum number of clusters, we must select the value of k at the "elbow" i.e. the point after which the distortion/inertia starts to decrease linearly. We employ the elbow method on the Kaggle dataset using k-means clustering, see Figure 1. We observe that the optimal number of clusters is 5.

Next we run evaluate the performance of the nine clustering models with the following defined parameters:

- KMeans: n_clusters=5, max_iter=10000000, n_init=42
- DBSCAN: eps=0.078
- Spectral Clustering: n_clusters = 5
- Agglomerative Clustering: n_clusters = 5
- Gaussian Mixture: n_components=3, covariance_type='full'
- BIRCH: n_clusters = 5
- Mini Batch K-Means: n_clusters = 5
- Mean Shift: quantile=0.2, bin_seeding=True
- OPTICS: min_samples=5

The results are shown on Table 1. The best Silhouette score is obtained using the Mean-Shift model (0.724). OPTICS (0.530) and DBSCAN (0.526) also had high Silhouette scores. BIRCH gave the worst Silhouette score (0.108). Gaussian Mixture (738.158) produced the highest Calinski-Harabasz Score. Agglomerative Clustering (581.066), k-means (561.455) and Mini-Bach k-means (545.470) also did well considering their high Calinski-Harabasz Scores. BIRCH (19.214) performed the worst. Mean-Shift (0.655) had the best Davies-Bouldin Score while Gaussian Mixture (0.981), k-means (0.992) and Spectral Clustering (0.992) performed better than the other models. BIRCH (1.840) performed the worst.

Table 1. Heading and text fonts.

| **Clustering Algorithm** | **Silhouette Score** | **Calinski-Harabasz Score** | **Davies-Bouldin Score** |
|---|---|---|---|
| k-Means | 0.397 | 561.455 | 0.992 |
| DBSCAN | 0.526 | 69.922 | 1.585 |
| SC | 0.404 | 580.535 | 0.992 |
| AC | 0.403 | 581.066 | 1.218 |
| GM | 0.320 | 738.158 | 0.981 |
| BIRCH | 0.108 | 19.214 | 1.840 |
| Mini Batch | 0.390 | 545.470 | 1.217 |
| Mean-Shift | 0.724 | 82.744 | 0.655 |
| OPTICS | 0.530 | 69.176 | 1.653 |

Though arbitrary, the data is now 'labelled' after running the clustering models. This means we can use supervised learning now to see how good the clustering is generalizing. It is just one way of testing the clustering. If the clustering model could find a meaningful split in the data it should be possible to train a classifier to predict which cluster should belong to a given instance. All models performed excellently having good classification metric values using the SGD classifier except Mean-Shift and BIRCH, see Table 2.

Table 2. Heading and text fonts.

| **Clustering Algorithm** | **Accuracy Score** | **Precision** | **Recall** | **F1 score** |
|---|---|---|---|---|
| k-Means | 100.000 % | 100.000 % | 100.000 % | 100.000 % |
| DBSCAN | 99.412 % | 97.503 % | 97.455 % | 97.448 % |
| SC | 100.000 % | 100.000 % | 100.000 % | 100.000 % |
| AC | 100.000 % | 100.000 % | 100.000 % | 100.000 % |
| GM | 100.000 % | 100.000 % | 100.000 % | 100.000 % |
| Birch | 99.804 % | 78.347 % | 77.314 % | 77.816 % |
| Mini Batch | 100.000 % | 100.000 % | 100.000 % | 100.000 % |
| Mean Shift | 74.063 % | 20.383 % | 20.622 % | 20.369 % |
| OPTICS | 99.524 % | 98.339 % | 97.918 % | 98.071 % |

For the final part of our procedure we illustrate using k-means. In K-means, the centroid is the mean of the documents in the cluster, and in TF-IDF all values are non-negative, so every word in every document in the cluster will be represented in its centroid. Thus, the terms significant in the centroid are those that are most significant across all the documents in that cluster. No word gets left out, but a lot become insignificant. The highest TF-IDF values in a document vector are those words most significant to that document; likewise, those highest valued words in the centroid are those most significant to the cluster as a whole

.

Using k-means we are able to distinguish five top cluster categories: (1) Safety, (2) Crime, (3) Products, (4) Countries and (5) Health, see Table 3. Note that in each category contain hashtags corresponding to it. This becomes relevant in today's world where there are millions of Twitter tweets daily on 2019-nCoV. This makes it very processing intensive to go through all the tweets to decipher meaning. Our procedure shows how governments and world bodies like the World Health Organization can quickly elicit meaningful categories from Twitter hashtags. This will give them a general overview of what the tweet data contains with the chance of placing more elaborate methods to selected tweets of interest, if any were found from looking at the different categories produced by the clustering model.

Table 3. Hashtag Categories using K-means.

| **Safety** | **Crime** | **Products** | **Countries** | **Health** |
|---|---|---|---|---|
| confinement | crimestatistics | Dental products | covid_19de | deliverhappiness |
| lockdown | domestic abuse | Hand sanitizers | covid_19india | difficltybreathing |
| quarantine life | staysafe | Dettol | covid_19italia | digitalhealth |
| social distancing | stayhomesavelives | Soap | covid_19sa | diagnosing |
| curfew threat | precautions | Droplet | covidespana | deathcareindustry |

## 5.1. Limitations

In general, the main limitations of our approach are:

1. A lack of a formal validation of the results using an independent Twitter dataset.

2. The study is presented just for one Twitter dataset, somehow limiting the potential generality of the proposed approach.

3. The use of supervised learning as a method of verifying the classification requires experts to validate the classifications made.

## 5.2. Improvements to Existing System

To improve our current system, we need to include in our testing additional 2019-nCoV datasets. In addition to further validation for our system, this can also show that our system can generalize to different Twitter datasets. These datasets can first include other 2019-nCoV data and then other past disease pandemics such as H1N1 [40] and SARS [10]. We can also hire experts to analyse and categorize Twitter data. This would give us a better indication on the exactness of our chosen classification algorithm: Stochastic Gradient Descent (SGD) Classifier. Based on this performance compared to the expert we can also implement other classifiers which may give better results.

### 5.3. Future aid in 2019-nCoV

In the present fight against 2019-nCoV our system can be used by government and private agencies to categorize Twitter data. This will give them the ability to quickly find tweets of interest instead of employing machine learning algorithms on the actual tweets which can take days to process. By using these categories government bodies can set up media briefs or meetings more targeted to the individuals in a geographic region or country. This will give them a more targeted approach in dealing with the virus. In addition, other machine learning algorithms can be employed on the results by first selecting tweets represented by a certain category. This is done by searching for one or more hashtags present within that category. This data reduction technique will now result in a faster processing of the now smaller number of tweets. For instance, many text-based machine learning algorithms can now be employed to find the sentiment of the individuals on this particular category or topic of interest. As machine learning sentiment analysis techniques improve emotions such as fear and panic can be discovered, and this can advise certain medical institutions of attending to certain mental states and affects within the population caused by 2019-nCoV first and even second infection spikes.

## 6. CONCLUSIONS

Unsupervised machine learning techniques such as clustering are widely gaining use with the recent increase in social communication platforms like Twitter and Facebook. Clustering enables the finding of patterns in these unstructured datasets. We collected tweets matching hashtags linked to COVID-19 from a Kaggle dataset. We compared the performance of nine clustering algorithms using this dataset. We evaluated the generalizability of these algorithms using a supervised learning model. Finally, using a selected unsupervised learning algorithm we categorized the clusters. The top five categories are Safety, Crime, Products, Countries and Health. This can prove helpful for bodies using large amount of Twitter data needing to quickly find key points in the data before going into further classification.

## REFERENCES


[1] Abd-Alrazaq, Alaa, Dari Alhuwail, Mowafa Househ, Mounir Hamdi, and Zubair Shah. "Top concerns of Tweeters during the COVID-19 pandemic: infoveillance study." *Journal of medical Internet research* vol. 22, no. 4 (2020): e19016.

[2] Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A. and Zoppetti, C., 2013. Nonparametric change detection in multitemporal SAR images based on mean-shift clustering. *IEEE transactions on geoscience and remote sensing*, *51*(4), pp.2022-2031.

[3] Allan, K., 1977. Classifiers. *Language*, *53*(2), pp.285-311.

[4] Ankerst, M., Breunig, M.M., Kriegel, H.P. and Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. *ACM Sigmod record*, *28*(2), pp.49-60.

[5] Arlia, D. and Coppola, M., 2001, August. Experiments in parallel clustering with DBSCAN. In *European Conference on Parallel Processing* (pp. 326-331). Springer, Berlin, Heidelberg.

[6] Bakshi, R.K., Kaur, N., Kaur, R. and Kaur, G., 2016, March. Opinion mining and sentiment analysis. In 2016 3rd IEEE International Conference on Computing for Sustainable Global Development (INDIACom), pp. 452-455.

[7] Bao, W., Lianju, N. and Yue, K., 2019. Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, *128*, pp.301-315.

[8] Bottou, L., 2012. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade* (pp. 421-436). Springer, Berlin, Heidelberg.

[9] Chen, E., Lerman, K. and Ferrara, E., 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. JMIR Public Health and Surveillance, 6(2), p.e19273.

[10] De Wit, E., Van Doremalen, N., Falzarano, D. and Munster, V.J., 2016. SARS and MERS: recent insights into emerging coronaviruses. Nature Reviews Microbiology, 14(8), p.523.
[11] Dubey, A.D., 2020. Twitter Sentiment Analysis during COVID19 Outbreak. Available at SSRN 3572023.
[12] Fang, Z. and Costas, R., 2020. Tracking the Twitter attention around the research efforts on the COVID-19 pandemic. arXiv preprint arXiv:2006.05783.
[13] Feldman, Maryann, and Pierre Desrochers. "Research universities and local economic development: Lessons from the history of the Johns Hopkins University." *Industry and Innovation* vol. 10, no. 1 (2003): 5-24.
[14] Gowda, K.C. and Krishna, G., 1978. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*, *10*(2), pp.105-112.
[15] Havrlant, L. and Kreinovich, V., 2017. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *International Journal of General Systems*, *46*(1), pp.27-36.
[16] Hechenbleikner, Elizabeth M., Daniel V. Samarov, and Ed Lin. "Data explosion during COVID-19: A call for collaboration with the tech industry & data scrutiny." *EClinicalMedicine* (2020).
[17] Jahanbin, K. and Rahmanian, V., 2020. Using Twitter and web news mining to predict COVID-19 outbreak. *Asian Pacific Journal of Tropical Medicine*, vol. 13.
[18] Jahanbin, K. and Rahmanian, V., 2020. Using Twitter and web news mining to predict COVID-19 outbreak. Asian Pacific Journal of Tropical Medicine, 13.
[19] Kabir, F., Siddique, S., Kotwal, M.R.A. and Huda, M.N., 2015, March. Bangla text document categorization using stochastic gradient descent (sgd) classifier. In *2015 International Conference on Cognitive Computing and Information Processing (CCIP)* (pp. 1-4). IEEE.
[20] Khan, K., Nikov, A. and Sahai, A., 2011. A fuzzy bat clustering method for ergonomic screening of office workplaces. In *Third International Conference on Software, Services and Semantic Technologies S3T 2011* (pp. 59-66). Springer, Berlin, Heidelberg.
[21] Kirk, Colleen P., and Laura S. Rifkin. "I'll Trade You Diamonds for Toilet Paper: Consumer Reacting, Coping and Adapting Behaviors in the COVID-19 Pandemic." *Journal of Business Research* (2020).
[22] Kiselev, V.Y., Andrews, T.S. and Hemberg, M., 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, *20*(5), pp.273-282.
[23] Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M.B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E.W. and Baddour, K., 2020. Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. Cureus, 12(3).
[24] Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M.B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E.W. and Baddour, K., 2020. Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. Cureus, 12(3).
[25] Krishna, K. and Murty, M.N., 1999. Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *29*(3), pp.433-439.
[26] Kywe, S.M., Hoang, T.A., Lim, E.P. and Zhu, F., 2012, December. On recommending hashtags in twitter networks. In *International conference on social informatics* (pp. 337-350). Springer, Berlin, Heidelberg.
[27] Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC *Transactions on ECE*, Vol. 10, No. 5, pp120-122.
[28] Lima, M.L., Nascimento, T.P., Labidi, S., Timbó, N.S., Batista, M.V., Neto, G.N., Costa, E.A. and Sousa, S.R., 2016. Using sentiment analysis for stock exchange prediction. International Journal of Artificial Intelligence & Applications (IJAIA), 7(1), pp.59-67.
[29] Maugis, C., Celeux, G. and Martin-Magniette, M.L., 2009. Variable selection for clustering with Gaussian mixture models. *Biometrics*, *65*(3), pp.701-709.
[30] Maulik, U. and Bandyopadhyay, S., 2002. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(12), pp.1650-1654.
[31] Medford, R.J., Saleh, S.N., Sumarsono, A., Perl, T.M. and Lehmann, C.U., 2020. An" Infodemic": Leveraging High-Volume Twitter Data to Understand Public Sentiment for the COVID-19 Outbreak. medRxiv.

[32] Medford, Richard J., Sameh N. Saleh, Andrew Sumarsono, Trish M. Perl, and Christoph U. Lehmann. "An" Infodemic": Leveraging High-Volume Twitter Data to Understand Public Sentiment for the COVID-19 Outbreak." *medRxiv* (2020).
[33] Mustafa, H.H., Mohamed, A. and Elzanfaly, D.S., 2017. An enhanced approach for arabic sentiment analysis. International Journal of Artificial Intelligence and Applications (IJAIA), 8(5), pp.1-14.
[34] Peng, K., Leung, V.C. and Huang, Q., 2018. Clustering approach based on mini batch kmeans for intrusion detection system over big data. *IEEE Access*, *6*, pp.11897-11906.
[35] Pokharel, B.P., 2020. Twitter Sentiment Analysis During Covid-19 Outbreak in Nepal. Available at SSRN 3624719.
[36] Romanyshyn, M., 2013. Rule-based sentiment analysis of ukrainian reviews. International Journal of Artificial Intelligence & Applications, 4(4), p.103.
[37] Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, pp.53-65.
[38] Rovetta, A. and Bhagavathula, A.S., 2020. Global Infodemiology of COVID-19: Focus on Google web searches and Instagram hashtags. medRxiv.
[39] Saito, T. and Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3).
[40] Schulze, M., Nitsche, A., Schweiger, B. and Biere, B., 2010. Diagnostic approach for the differentiation of the pandemic influenza A (H1N1) v virus from recent human influenza viruses by real-time PCR. PloS one, 5(4), p.e9966.
[41] Song, F., Shi, N., Shan, F., Zhang, Z., Shen, J., Lu, H., Ling, Y., Jiang, Y. and Shi, Y., 2020. Emerging 2019 novel coronavirus (2019-nCoV) pneumonia. *Radiology*, *295*(1), pp.210-217.
[42] SudalaiRajkumar: Novel Corona Virus 2019 Dataset. data retrieved March 30, 2020 from Kaggle, https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset (2020)
[43] Toshniwal, A., Taneja, S., Shukla, A., Ramasamy, K., Patel, J.M., Kulkarni, S., Jackson, J., Gade, K., Fu, M., Donham, J. and Bhagat, N., 2014, June. Storm@ twitter. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 147-156).
[44] Von Luxburg, U., 2007. A tutorial on spectral clustering. *Statistics and computing*, *17*(4), pp.395-416.
[45] Xue, J., Chen, J., Chen, C., Zheng, C. and Zhu, T., 2020. Machine learning on Big Data from Twitter to understand public reactions to COVID-19. arXiv preprint arXiv:2005.08817.
[46] Xue, Jia, Junxiang Chen, Chen Chen, ChengDa Zheng, and Tingshao Zhu. "Machine learning on Big Data from Twitter to understand public reactions to COVID-19." *arXiv preprint arXiv:2005.08817* (2020).
[47] Zhang, T., Ramakrishnan, R. and Livny, M., 1996. BIRCH: an efficient data clustering method for very large databases. *ACM Sigmod Record*, *25*(2), pp.103-114.

## AUTHORS

**Koffka Khan** received the M.Sc., M.Phil., DPhil. degrees from the University of the West Indies. He is currently an Assistant Lecturer and has up-to-date, published numerous papers in journals & proceedings of international repute. His research areas are computational intelligence, routing protocols, wireless communications, information security and adaptive streaming controllers.

**Emilie Ramsahai** is a consulting Data Scientist, with more than 20 years industry experience. She is currently working with UWI-Roytec in programme development and course writing. She completed her PhD in Statistics and a Masters in Computer Science, both at the University of the West Indies, where she has also lectured the Big Data and Visualisation course from the Masters in Data Science, offered by the Department of Computing and Information Technology, St Augustine Campus. She also completed her fellowship at the International Centre for Genetic Engineering and Biotechnology (ICGEB) in New Delhi, India and continues to publish and collaborate with a number of researchers in this area.