International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

Real-Time Twitter Corpus Labelling Using Automatic Clustering Approach

Itisha Gupta[1] and Nisheeth Joshi[2]

[1,2]*Department of Computer Science, Banasthali Vidyapith, Tonk, Rajasthan, India*

E-mail address: itishagupta07@gmail.com, jnisheeth@banasthali.in

Abstract: In this paper, we present a novel automatic labelling approach for the large amount of unlabelled real-time twitter datasets for textual-based twitter sentiment analysis. The tweets are labelled or classified as Positive, Negative or Neutral using the novel automatic approach. The proposed approach applies an unsupervised clustering technique that would generate clusters based on the underlying patterns (finding similarities between tweets) in the collected twitter corpus. Twitter search API is used to collect real-time English tweets on several topics such as "#Demonetization", "#lockdown", and "#9pm9minutes" by the use of search operator. To analyse the sentiment from real-time tweets, labelling of the corpus is required. Manual annotation of large twitter corpus is time and labor-intensive. Moreover, domain experts are needed for labelling of tweets belonging to a particular domain. Thus, in this work, we propose the use of the K-mean clustering approach, which is an unsupervised way of labelling corpus, which could then be used for learning supervised models such as SVM for sentiment analysis. To make the corpus ready for clustering and to get quality clusters, we have applied some basic to advanced cleaning operations known as tweet normalization. Furthermore, we perform extensive feature engineering to generate different types of features including POS-based (Part-of-Speech), n-grams, twitter-specific, and lexicon-based features from our collected unlabelled twitter corpus. Those features act as input to the K-mean clustering algorithm and help it in identifying patterns from the data for cluster generation. In the end, cluster analysis is done manually to find out the sentiments expressing by tweets in a particular cluster. Accordingly, cluster classification is done and each cluster is assigned one class that is Positive, Negative, or Neutral. The main contribution of this work is the idea of amalgamation of extensive feature engineering with the unsupervised clustering approach for classification of large unlabelled twitter corpus.
.

Keywords: Feature Engineering, Cluster Analysis, Corpus Labelling, Real-Time Tweets, Twitter Sentiment Analysis, Pre-processing

1. INTRODUCTION

This Microblogging is one of the popular and widespread broadcasting media amidst the internet world. People frequently use microblogging websites such as Twitter (created in 2006) for sharing their views, opinions, emotions, etc. on any event, product, services, and idea. Thus, the enormous amount of opinionated data is available in digital form on different platforms such as blogs and discussions which is very useful for decision making or feedback. Day to day basis huge amount of tweet gets generated on hot and latest topics. Automatic analysis and reasoning of such data help in deriving meaningful visions, which carries opportunities for users, consumers, and businesses i.e. analysis of such data provides insight into people's opinion and inclination. One of the effective techniques for analysis of such opinionated tweets is Twitter sentiment analysis (TSA). In most of the earlier works on TSA, a supervised learning approach has been used that needs labelled corpus for training and sentiment prediction. However, real-time tweets gathered from twitter using API are not labelled or classified readily. That is it is necessary to label the collected twitter corpus before performing supervised sentiment analysis on them. Corpus can be labelled or classified either into sentiment classes (such as Positive or Negative) or into emotions such as joy, fear, sadness, and many more. Manual annotation of the large unlabelled corpus is labor-intensive and requires the knowledge of domain expert. Thus, our solution is to assign a sentiment label (Positive, Negative or Neutral) to the real-time tweet (that is classifying a tweet according to the sentiment expressed by it) through the automatic clustering approach.

There exist various automatic approaches in the literature for corpus labelling [12], [17], [20], [23], [28]. One of the famous approaches is the distant supervision approach for corpus classification based on the presence

E-mail:author's email

of emoticons in a tweet. That is, if a tweet is containing only positive emoticons, then that tweet would be classified as positive. Few works used positive and negative hashtag words such as #happy for corpus labelling. Rule-based classifier (based on the count of positive and negative sentiment words in a tweet) is also an automatic way of corpus classification. As an illustration, if a tweet contains a minimum two positive sentiment words and no negative sentiment words then that tweets is labelled as positive. Nevertheless, such automatic approaches have shown considerable performance, but ignore the context in which a word appears due to classifying the corpus based on the presence of positive or negative counts of emoticons and sentiment words. Moreover, there is a possibility that a classifier learns on the corpus labelled through above-mentioned approaches will not generalize. It would simply mimic the rule-based classifier on the labelled training corpus.

One way to combat this problem is to use as many features for the corpus labelling rather than using only the count features (such as no. of positive emoticons, no of positive words, and many more). Thus, in this paper, we aim to present an unsupervised clustering approach for classification or labelling of real-time unlabelled twitter corpus. Clustering is an unsupervised approach used to find sub-groups (clusters) within a dataset based on the underlying patterns [9], [16]. Objects in the same cluster are more similar to each other than the other cluster. Put simply, clustering needs unlabelled data as input and gives clusters as output. It is widely used in several applications such as market segmentation, image segmentation, and many more. Nevertheless, the main aim of clustering is to find structures in the data, but clusters generated by it could be considered as labels for the unlabelled corpus. Then one can train a classifier using those labels as the target. Hence, clustering can be used for classification that is labelling of unlabelled data. Furthermore, clustering has the power to reveal the unforeseen groups in a large dataset that may convey significant information.

Since our main goal is labelling or classification of real-time twitter corpus, we have collected real-time tweets on the various topics including “#Demonetization”, “#Lockdown”, and “#9pm9minutes” through the use of Twitter search API. First of all, the collected corpus is cleaned and normalized by the use of some basic to advanced cleaning tasks from our previous work [14]. Then, we extracted various types of features such as Pos-based features, lexicon-based features, morphological features, and Tf-Idf features (n-grams) from the cleaned twitter corpus. All the extracted features are concatenated to get a final feature vector, which is then given as input to one of the popular clustering algorithm K-mean, often used for labelling of unlabelled data [18]. Finally, K-mean clustering approach is applied to classify the twitter corpus into three clusters based on the various syntactic and semantic features. Generated clusters are manually analysed for inspecting the sentiment expressed by tweets in clusters. This way helps in assigning each cluster one sentiment class that is Positive, Negative or Neutral. To be more specific, we now have twitter corpus classified into three classes namely positive, negative or neutral. We empirically evaluate the combination of different features in cluster generation. We have observed the best result when all features are given as input to the clustering algorithm. This shows that features play an important role in identifying underlying patterns from the unlabelled corpus through the K-mean clustering method.

Following are the main contributions of this paper:

- We present an automatic cascade approach which is an amalgamation of extensive feature engineering and unsupervised approach for the labelling or classification of the large unlabelled data.
- We use the Twitter search API to collect real-time tweets on various topics including “#Demonetization”, “#Lockdown”, and “#9pm9minutes.
- We perform extensive feature engineering to generate various syntactic and semantic features such as Pos-based, twitter-specific, lexicon-based, and many more that would be used to identify structures within the dataset.
- We demonstrate the use of one of the popular K-mean clustering approach for evaluation of large unlabelled twitter datasets in a fast and objective way to generate clusters (equals to the number of classes or labels that we want for our unlabelled corpus) by the use of various extracted features.
- We inspect each cluster manually outputted by the K-mean for the assignment of sentiment class based on the sentiment expressed by tweets in each cluster. Labelled corpus could be used to train a classifier so that new instances can be predicted.

The rest of paper is organized as follows: section 2 presents the literature review of earlier work on clustering with classification, section 3 describes the framework of our proposed approach for labelling of twitter corpus, section 4 provides the labelling result and the last section concludes this paper with possible future directions.

2. LITERATURE REVIEW

There exist a large amount of opinionated text in digital form which can provide informative knowledge for strategic decision. There are many sources of such data including blogs, newspapers, social media platforms, etc. Raw data available from such sources are in unformatted form. In order to analyse opinionated data, one needs to

classify data either into classes (groups) or need to find hidden structure from it. In the existing literature, several methods have been used for the automatic classification or labelling of the large amount of unlabelled corpus specifically twitter corpus. The most popular and widely used approach is the distant supervision approach, in which based on the presence of positive (":-), :)") and negative emoticon (":-(, :("), the label is assigned to each training tweets. It is an automatic approach for assigning class labels to text. Reference [25] firstly presented this approach, which used this approach for labelling of data from the Usenet newsgroup. Reference [12] was the first to use the distant supervision approach for the labelling of training tweets containing emoticons. They queried the Twitter API with positive emoticon (";)") as well as with negative emoticon (":(") and collected 1.6 million tweets classified into positive and negative classes. Thus, tweet having only positive emoticons is classified as positive and vice-versa. They trained three machine learning models such as SVM, NB, and MaxEnt on the corpus labelled by the distant supervision approach. Nevertheless, they obtained significant performance but observed that there is a negative impact on SVM and MaxEnt due to the presence of emoticons in the training set. The reason is the use of emoticons for training set labelling. Several researchers depend on a distant supervision approach for the collection of training tweets (labelling of tweets) [5], [21], [23], [28].

Another important approach for the creation of the training set is the use of positive and negative hashtag words such as #joy, #disappointed, and many more because people often use the hashtag in a tweet for expressing their sentiment. This technique has been used by several researchers in the past [7], [17], [20]. Reference [17], for instance made use of frequent hashtagged words (e.g., #epicfail, #love) (which are an indication of positive, negative and neutral sentiment) for the training set collection that is labelling of training tweets into positive, negative or neutral classes. This approach was even used by [7] for training data creation but their experiments was limited to sentiment/non-sentiment classification. Reference [20] collected and labelled 15214 tweets based on positive and negative hashtag words.

In contrary with automatic approaches, several researchers used the crowdsourcing method [1], [2], [19] such as Amazon Mechanical Turk or third-party service (Alchemy API) for the labelling of twitter corpus, while few labelled the twitter corpus by themselves [4], [8], [20]. For instance, Obama-McCain Debate (OMD) dataset [27] contains 3238 tweets on U.S presidential election (Sept. 2008), labelled as positive, negative, mixed or others by Amazon Mechanical Turk. Nevertheless, manual labelling of corpus provides more accurate results, but it is very time consuming and labor-intensive.

Recently, the unsupervised clustering approach is being used for the labelling of the corpus. Reference [6] presented a comparative analysis of various unsupervised clustering approaches that have been used in the past for analyzing Twitter data. They presented a comparison of several clustering algorithms, dataset size, clustering features, no. of clusters, and evaluation metrics. Though main aim of unsupervised clustering approach is to find the hidden patterns among dataset, it is being widely used for the classification of unlabelled corpus [9], [10], [13], [16], [18], [22], [24], [26].

Reference [16], for instance, proposed the use of hybrid technique in order to improve SVM performance. They used the K-mean clustering approach for the training subset selection and then hyperparameter tuning was done to optimize the effectiveness of classifier. They evaluated their result on Stanford Twitter Dataset (STS) [12] and the Amazon customer review dataset. Reference [13] presented a comparison of two clustering approach K-mean and Non-Negative Matrix factorization (NMF) on 30000 tweets containing the term "world cup to find topics". Reference [18] presented a hybrid framework for sentiment analysis unlabelled Email data. They presented a comparison of three clustering approaches including K-mean, sentiment clustering, and polarity labelling for labelling of unlabelled Email data and several supervised models for sentiment analysis such as SVM, NB, etc. Results showed that K-mean outperformed the other two approaches in the clustering of Email data and SVM performed best in sentiment analysis.

Reference [26], more recently presented a combination of two clustering approach that is K-mean and DENCLUE for twitter sentiment analysis. They observed that a combination of those two algorithms provided effective results than the state-of-the-art methods (e.g., DBSCAN, K-mean) in terms of clustering performance, run time and no. of clusters. In another work [22], authors used the combination of various techniques such as Tf-Idf, Singular Value Decomposition (SDF) (for dimensionality reduction), and artificial bee colony (ABC) (an algorithm used to detect the best initial state of centroids for K-mean) for improving the K-mean performance (41% than normal K-mean). They applied K-mean to generate clusters which were then scored by SentiWordNet [3] for class labelling.

Most of the above-aforementioned works used the Bag-of-Word approach (BOW) to generate the feature vector for the clustering algorithm K-mean. However, considerable performances have been reported by them, but a classifier learn on the training set generated through a simple BOW approach will mimic the word look-up based distribution and might not generalize. In this paper, we too use the K-mean clustering for labelling of unlabelled data but extensive feature engineering (several syntactic and semantic features) is also performed that

will help K-mean in identifying the hidden patterns in the unlabelled corpus.

3. METHODOLOGY

This section presents a detailed description of our proposed novel automatic approach for the labelling of real-time twitter corpus. We start with the collection of real-time tweets on various topics such as "#Demonetization", "#Lockdown", and "#9pm9minutes (through the use of twitter search API), followed by the pos tagging and tokenization of generated twitter corpus using CMU Pos tagger [11], designed especially for twitter. CMU tagger is able to identify linguistic peculiarities of tweets such as usernames, URLs, hashtags, emoticons, and many more as discrete entities. Thus, for each input tweet, we have a list of tweet tokens and their corresponding POS tokens. Those tokens would be very useful for extensive feature engineering. Our proposed framework incorporates several phases including data (tweet) pre-processing, feature engineering, clustering for classification, and finally cluster analysis. Fig. 1 epitomizes the workflow of proposed framework.

A. Data Pre-processing (Tweet Normalization)

Tweets are user-generated short messages and often have oddities and quirks. Thus, the tweet is highly unstructured containing a lot of misspelled words, acronyms, and domain-specific entities. It is necessary to clean the tweet by the removal of unnecessary symbols and words which don't have any semantic orientation such as digits, URLs, whitespaces, stopwords, and many more. The indispensable task of noise removal and normalizing the out-of-vocabulary and Non-English words to their canonical forms is known as data pre-processing. It would prepare the real-time twitter corpus for further analysis and helps in the reduction of feature space too by the removal of unnecessary elements from the tweet. Several early works highlights the significance of data pre-processing before clustering [13], [16], [22], [26].

In our previous work [14], we have implemented a pre-processing framework containing two phases: basic cleaning and tweet normalization. Operations or tasks for the noise removal from tweets come under the basic cleaning phase such as removal of whitespaces, punctuations, stopwords, URLs, numbers, and many more. Tweet normalization phase includes the task of replacing the ill-formed and non-standard words to their canonical forms such as replacement of acronym "lol" by "laughing out loud". Thus, we are able to get a clean and normalized real-time twitter datasets which is ready for the next phase that is feature engineering.

It is important to mention an important linguistic element namely "negation" that we have handled during the tweet normalization phase. Negation has the ability to change the entire semantic orientation of the text. In addition to replace negation cues with tag "negation", we have determined the scope of negation too i.e. words affected by negation are suffixed with tag "_NEG". Furthermore, we exclude few negation tweets from the scope determination procedure which are having explicit negation words but literally, there is no sense of negation [15] as in tweet "Now isn't this lovely ! Hazards of #DeMonetization https://t.co/mBwXVxFYKN". In that negation tweet, the word "isn't" not affecting the semantic orientation of opinionated word "lovely". Hence, negation is ignored in this tweet.

B. Feature Engineering

The main aim of this work is to use the K-mean unsupervised clustering approach for the automatic classification of unlabelled twitter corpus. It is worth to mention that, most of the algorithms need input in the form of nd-array that is "no. of observations * no. of features". Thus, there is a need for numerical representation of input twitter corpus. In this fold, we have performed extensive feature engineering which led to the generation of several syntactic and semantic features as shown in below table I.

C. Clustering for Classification

. In this phase, we have used a popular unsupervised feature-based clustering technique known as K-mean for the labelling of unlabelled real-time twitter corpus. The purpose of using the clustering approach here is segmentation as well as classification of unlabelled corpus. There exist several other automatic approaches (describe above in the literature review section) (such as distant supervision) for labelling of unlabelled corpus but such approaches label the corpus based on some linguistic elements of a tweet such as an emoticon and hashtag. This might led to the biased result when a classifier would be trained on that labelled corpus (classifier would not generalize).

Moreover, manual annotation of the unlabelled corpus is expensive and time-consuming too. Thus, we adopt a clustering approach which is often used for classification purpose. K-mean is a good option because it is capable of handling high dimensional data. It is a distance or centroid based algorithm in which we calculate the distances to assign a data point to a cluster. Each cluster is associated with a centroid.

The main objective of K-mean is to minimize the sum of distances between data points and their respective centroids. It deals in determining structure in the unlabelled corpus. The input to K-mean is a large feature vector which is the concatenation of various syntactic and semantic features, generated in above sub-section B. The aim of using many feature groups apart from Tf-Idf (most of the early works, Tf-Idf is the only feature given as input to K-Mean) is to help the K-mean in finding hidden patterns more accurately. Additionally, classifier trains on the clusters created by K-mean through the use of many feature groups will generalize rather than mimicking.

International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

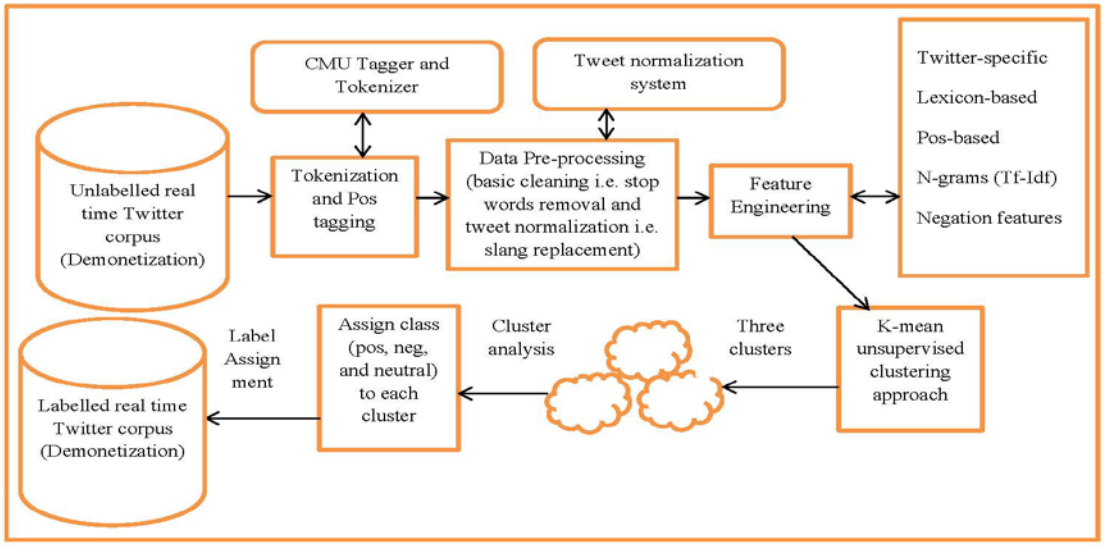

Figure 1. Proposed framework for automatic labelling of real-time twitter corpus (e.g., #Demonetization)

TABLE I. FEATURE VECTOR GENERATED FROM THE UNLABELLED TWITTER CORPUS

Feature Group	Features	Description
N-grams (unigrams + bigrams)	Tf-Idf feature vector (drop the terms that occur less than 2 times and more than 50%).	Normalize the count of a token based on the no. of document in which it appears. It penalizes the most occurring term and reward the rare term.
Morphological features	No. of hashtag words Count and presence of elongated words Count of emoticons Count of exclamation, question mark, no. of tokens having only exclamation, question mark, count and existence of exclamation, question at the tweet end No. of capitalized words Presence of slang 21-dimensional FV is generated	Twitter-specific and generated based on the linguistic peculiarities of a tweet i.e. hashtag, emoticons, specific punctuation like exclamation and question, all capitalized words, and many more.
Pos-based features	No. of existences of each unique pos tag.	Part of speech features such as noun, adjective, adverb, etc. Helps in context identification.
Negation features	Count and no. of negated context	Affects lexicon-based and n-gram features. Help the clustering technique in finding the negated context patterns like "not good_NEG" so that good would not be considered as positive sentiment bearing word if tagged with "_neg".
Lexicon-based features Twitter-specific automatic lexicon-based features (S140 and NRC-Hashtag)	• Count of tokens with non-zero sentiment score • Sum of score • Maximum of score • Score of the last token Generated for all positive, negative and all tokens (12-dimensional feature vector)	Automatic lexicons are specifically created to provide the score to word under negated context based on the fact that negation doesn't reverse polarity every time. They are having real-valued scores for unigrams and bigrams. Each n-gram is given two scores: one in affirmative context and another in negated context.
Manual lexicon-based features (Bing-Liu, NRC-Emoticon, and MPQA)	Sum of positive score of words and sum of negative score of words in negated context Sum of positive score of words and sum of negative score of words in affirmative context Above 4 features are repeated for hashtag words. Repeat features for all—caps, lowercase and unique pos tags which led to generation of 104 dimensional FV.	There is no real valued score for NRC-Emoticon and Bing-Liu. Put simply, they indicate a word as positive or negative. We used +1 for positive word and -1 for negative. MPQA indicates strength of polarity too so we used +1/-1 for weak intensity and +2/-2 for strong.

E-mail:author's email

The output of the K-mean approach is the three clusters, each having tweets that are more similar to each other. The reason for getting only three clusters is that we want to label the real-time tweets as Positive, Negative or Neutral.

D. Cluster Analysis for Class Assignment

This is the last and final phase of our framework in which generated 3 clusters are inspected manually so that each cluster can be assigned to one of three classes namely Positive, Negative, or Neutral. We have analysed the sentiments expressed by a few tweets belonging to each cluster. There is no need to analyse each tweet of a cluster because tweets in a cluster are more similar to each other i.e. they will express the same kind of sentiment. Analysis of a few tweets respective to a cluster will give us the idea of whether a tweet is positive, negative or neutral. Fig. 2 epitomizes the manual analysis procedure for clusters. Accordingly, each cluster is assigned to one of three classes. For instance, if some of the tweets of a cluster are expressing the positive sentiment, then that cluster is assigned class "Positive". Put simply, all the tweets of a positively assigned cluster will be labelled as "Positive". In the end, we get a labelled real-time twitter corpus with each tweet labelled as Positive, Negative or Neutral.

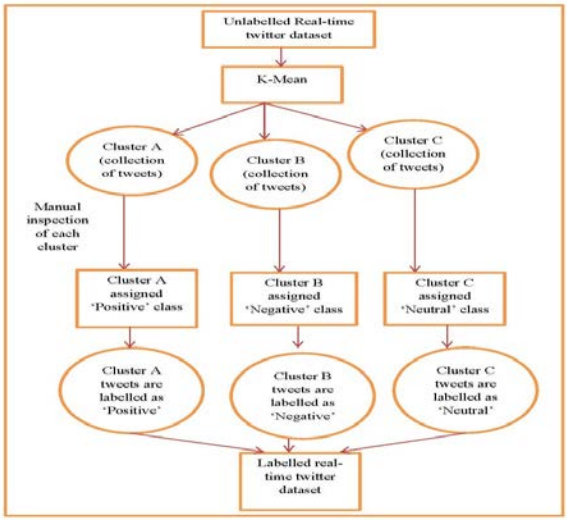

Figure 2. Procedure for manual analysis of clusters

4. EVALUATION

Several experiments are undertaken in this work in order to show the effectiveness of our proposed automatic clustering approach cascaded with extensive feature extraction for classification of unlabelled twitter corpus.

A. Corpus

Since this work aims in labelling of real-time twitter dataset, we need to collect the real-time tweets from twitter. First of all Twitter API authentication is required to generate authentication credentials. Authentication keys (API keys) are generated upon logging in the Twitter account. To establish a connection with twitter stream API keys are needed. Twitter API allows pulling each and every tweet on a certain topic. We used python library "Tweepy" to make a connection with Twitter API. Tweepy provides a convenient way of accessing API with language Python. It contains several classes and functions that epitomize API endpoints and it handles various low-level details such as HTTP request, rate limit, encoding, decoding, and many more. There are two types of Twitter API: streaming API and search API.

Twitter streaming API lets you collect tweets in near real-time i.e. push of tweets by twitter (forward). However, tweets collected by streaming API form a small segment of actual tweets. We chose the search API, which led to the collection of tweets that have happened already (search in back). Past 7 days tweet can be collected with search API. Search API provides a powerful set of operators for filtering tweets based on language, sender location, and many more. Furthermore, using search API more number of tweets can be collected because we can make 15 API requests per minute.

We used Twitter search API for downloading real-time English tweets related to keyword "#Demonetization", '#Lockdown', and '#9pm9minutes' with English language filtering operator. Tweets on demonetization were collected from 31/12/2016 to 20/03/2017. We collected total of 19615 tweets on the topic "#Demonetization". We also collected tweets on current hot trending topics such as tweets on "#Lockdown" and "#9pm9minutes" (i.e. light candle or torch at 9 pm for 9 minutes to show unity in India). Tweets on "#Lockdown" were collected from 27/03/2020 to 6/04/2020 and tweets on "#9pm9minutes" were collected from 4/04/2020 to 6/04/2020. We gathered 18365 tweets on #Lockdown and 6358 tweets on #9pm9minutes. Search API results into a JSON object containing tweet text and several associated metadata (its data about a tweet like data, time, user, etc.). We have extracted only tweet text from the JSON object and saved it into three different text files, one for each real-time dataset. Table II presents the statistics of unlabelled real-time twitter corpus that we have generated and fig. 3 portrays the real-time tweet datasets generation procedure.

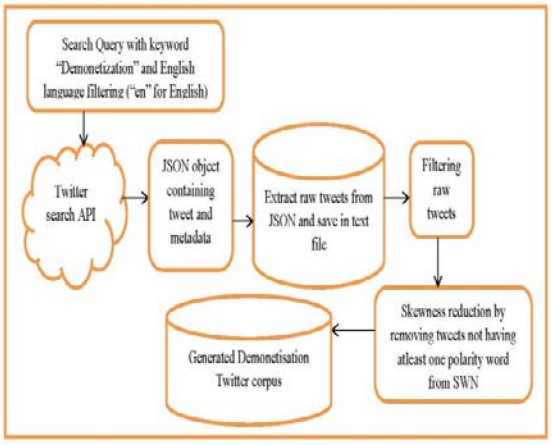

Figure 3. Real-time twitter corpus generation

TABLE II. STATISTICS OF UNLABELLED REAL-TIME TWITTER DATASETS

Dataset	Size (No. of tweets)
Demonetization	19615
Lockdown	18365
9pm9minutes	6358

B. *Experiments*

In this section, experimentation results are presented showing the cluster generation from the real-time twitter corpus collected in above sub-section 4.A. We performed a series of experiments with the K-mean clustering approach and combinations of several feature groups. We observed the best clusters when all features are given as input to K-mean, showing the significance of using more features with K-mean. It is important to notice that, we set the "n_clusters" hyperparameter value of K-mean to 3 because our goal is to classify the real-time twitter dataset into three classes namely "Positive", "Negative", or Neutral". Table III shows the statistics of three clusters (number of tweets per cluster) generated by K-mean for each real-time twitter dataset.

TABLE III. POPULATION OF EACH CLUSTER (NO. OF TWEETS PER CLUSTER) FOR REAL-TIME TWITTER CORPUS

Dataset	First cluster (Cluster 0)	Second cluster (Cluster 1)	Third cluster (Cluster 2)	Total
Demonetization	4976	8865	5774	19615
Lockdown	4924	8252	5189	18365
9pm9minutes	1614	2339	2405	6358

Finally, cluster analysis is done manually for the assignment of class to each cluster. Table IV shows the result of a class assignment to each cluster for all the three twitter datasets. For instance, from table IV we observed that for the "Demonetization" dataset cluster 0 (first cluster) is assigned "Positive" class, cluster 1 (second cluster) is assigned "Neutral class, and cluster 2 (third cluster) is assigned "Negative" class. Put simply, all the tweets of cluster 0 in "demonetization" corpus are labelled as "Positive", cluster 1 tweets are labelled as "Neutral", and cluster 2 tweets are labelled as "Negative". Table V presents the final labelling of unlabelled twitter corpus based on the classes assigned to clusters and fig. 4 portrays the graphical representation of labelled real-time corpora. It is worth noting that all three real-time twitter datasets are unbalanced, so balancing of datasets is required before using it for further analysis i.e. sentiment analysis.

TABLE IV. MANUAL CLASS ASSIGNMENT TO EACH CLUSTER FOR THE REAL-TIME TWITTER DATASETS

Dataset	Cluster	Classes
Demonetization	Cluster 0 (4976)	Positive
	Cluster 1 (8865)	Neutral
	Cluster 2 (5774)	Negative
Lockdown	Cluster 0 (4924)	Negative
	Cluster 1 (8252)	Neutral
	Cluster 2 (5189)	Positive
9pm9minutes	Cluster 0 (1614)	Negative
	Cluster 1 (2339)	Neutral
	Cluster 2 (2405)	Positive

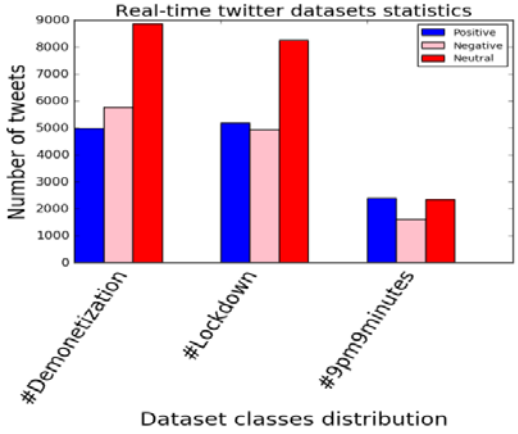

Figure 4. Labelled real-time twitter corpus statistics

TABLE V. CLASS (LABEL) ASSIGNMENT TO TWEETS (STATISTICS OF LABELLED REAL-TIME TWITTER DATASET)

Dataset	# Positive tweets	# Negative tweets	# Neutral tweets	Total
Demonetization	4976 (25.37%)	5774 (29.44%)	8865 (45.19%)	19615
Lockdown	5189 (28.25%)	4924 (26.81%)	8252 (44.94%)	18365
9pm9minutes	2405 (37.83%)	1614 (25.38%)	2339 (36.79%)	6358

We also provided an interesting visualization of generated clusters in the word cloud form. Word cloud is an interesting technique for textual data representation such that each token (word) size is directly proportional to its importance or frequency. Word cloud helps for analysis of data for the microblogs and other social media sites. We have generated word clouds for each cluster with respect to real-time twitter dataset. As an illustration, fig. 5 shows the word cloud for positive cluster of "#Demonetization" twitter corpus. We have observed big size words such as success, positive, great, thanks demonetisation, good, benefit, etc. in the below fig. 5, which make sense to be in positive tweets.

Figure 5. Positive cluster word cloud for "Demonetization" corpus

5. CONCLUSION AND FUTURE WORK

In this work, we have described an automatic framework for the labelling of real-time twitter datasets into three classes namely Positive, Negative, or Neutral, collected through twitter search API. We have used the most popular unsupervised K-mean clustering approach in cascading with extensive feature engineering for classification of unlabelled twitter datasets. We managed to generate 3 clusters equal to no. of classes. Each cluster is manually inspected for a class assignment to get a label for each tweet.

In the future, we shall aim for the optimization of feature vectors for dimensionally reduction and cut down of computational cost. Moreover, we would explore the process of twitter sentiment analysis on labelled real-time twitter dataset in order to show the effectiveness of labels generated by K-mean.

REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R.J. Passonneau, "Sentiment analysis of twitter data," in Proceedings of the Workshop on Language in Social Media (LSM 2011), June 2011, pp. 30-38.

[2] M.Z. Asghar, A. Khan, S. Ahmad, M. Qasim, and I.A. Khan, "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme," PloS one, vol. 12, no. 2.

[3] S. Baccianella, A. Esuli, and F. Sebastiani, F, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in proceedings of the Lrec, May 17-23, 2010, vol. 10, no. 2010, pp. 2200-2204, Valletta, Malta.

[4] A. Bakliwal, J. Foster, J. van der Puil, R. O'Brien, L. Tounsi, and M. Hughes, "Sentiment analysis of political tweets: Towards an accurate classifier," in NAACL workshop on language analysis in social media, Association for Computational Linguistics, June 13, 2013, Atlanta, GA.

[5] K.Z., Bertrand, M. Bialik, K. Virdee, A. Gros, and Y. Bar-Yam, "Sentiment in new york city: A high resolution spatial and temporal view," arXiv preprint arXiv:1308.5010, August 22, 2013.

[6] K.A. Crockett, D. Mclean, A. Latham, and N. Alnajran, "Cluster Analysis of twitter data: a review of algorithms," in Proceedings of the 9th International Conference on Agents and Artificial Intelligence (Vol. 2, pp. 239-249). Science and Technology Publications (SCITEPRESS)/Springer Books, 2017.

[7] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in Proceedings of the 23rd international conference on computational linguistics: posters, August 23-27, 2010, pp. 241-249, Beijing, China

[8] I. El. Alaoui, Y. Gahi, R. Messoussi, Y. Chaabi, A. Todoskoff, and A. Kobi, "A novel adaptable approach for sentiment analysis on big social data," Journal of Big Data, vol. 5, no. 1, pp. 12, 2018.

[9] D. Ferraretti, G. Gamberoni, and E. Lamma, E, "Unsupervised and supervised learning in cascade for petroleum geology," Expert Systems with Applications, vol. 39, no. 10, pp. 9504-9514, 2012.

[10] V. Friedemann, "Clustering A Customer Base Using Twitter Data," CS-229, 2015.

[11] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, et al., "Part-of-speech tagging for twitter: Annotation, features, and experiments," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, June 2011, pp. 42-47. Association for Computational Linguistics.

[12] A. Go, R. Bhayani, and L. Huang, L, "Twitter sentiment classification using distant supervision," CS224N project report, Stanford, vol. 1, no. 12, 2009.

[13] D. Godfrey, C. Johns, C. Meyer, S. Race, and C. Sadek, C, "A case study in text mining: Interpreting twitter data from world cup tweets," arXiv preprint arXiv:1408.5427, 2014.

[14] I. Gupta and N. Joshi, N, "Tweet normalization: A knowledge based approach," in 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), December 2017, pp. 157-162, Dubai, UAE.

[15] I. Gupta and N. Joshi, "Enhanced Twitter Sentiment Analysis Using Hybrid Approach and by Accounting Local Contextual Semantic," Journal of intelligent systems, vol. 29, no. 1, pp. 1611-1625, 2019.

[16] K. Korovkinas, P. Danėnas, and G. Garšva, G, " SVM and k-Means Hybrid Method for Textual Data Sentiment Analysis," Baltic Journal of Modern Computing, vol. 7, no. 1, pp. 47-60, 2019.

[17] E. Kouloumpis, T. Wilson, and J. Moore, J, "Twitter sentiment analysis: The good the bad and the omg!," in Fifth International

AAAI conference on weblogs and social media, July 2011, Barcelona, Spain.

[18] S. Liu and I. Lee, I, "Email sentiment analysis through k-means labeling and support vector machine classification. Cybernetics and Systems, vol. 49, no. 3, pp. 181-199, 2018.

[19] S.M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin, J, "Sentiment, emotion, purpose, and style in electoral tweets. Information Processing & Management," vol. 51, no. 4, pp. 480-499, 2015.

[20] S. Mukherjee and P. Bhattacharyya, "Sentiment analysis in twitter with lightweight discourse analysis," in Proceedings of COLING 2012, December 2012, pp. 1847-1864.

[21] A. Muhammad, N. Wiratunga, and R. Lothian, "Contextual sentiment analysis for social media genres," Knowledge-based systems, vol. 108, no. 92-101, 2016.

[22] K. Orkphol and W. Yang, "Sentiment Analysis on Microblogging with K-Means Clustering and Artificial Bee Colony," International Journal of Computational Intelligence and Applications, vol. 18, no. 03, 1950017, 2019.

[23] A. Pak and P. Paroubek, P, "Twitter as a corpus for sentiment analysis and opinion mining," in proceedings of the LREc, May 17-23, 2010, vol. 10, no. 2010, pp. 1320-1326, Valletta, Malta.

[24] R.H. Patil and S.P. Algur, S.P, "Classification Connection of Twitter Data using K-Means Clustering," IJITEE, vol. 8, 2019.

[25] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in Proceedings of the ACL student research workshop, June 2005, pp. 43-48.

[26] H. Rehioui and A. Idrissi, A, "New Clustering Algorithms for Twitter Sentiment Analysis," IEEE Systems Journal, vol. 14, no. 1, pp. 530-537, 2019.

[27] D.A. Shamma, L. Kennedy, and E.F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources," in Proceedings of the first SIGMM workshop on Social media, October, 2019, pp. 3-10.

[28] J. Spencer and G. Uchyigit, "Sentimentor: Sentiment analysis of twitter data," in SDAD@ ECML/PKDD, pp. 56-66, 2012.

Itisha Gupta is a research scholar in the department of computer science at Banasthali Vidyapith, Rajasthan, India. She did MCA (Masters of computer applications) from Gurgaon Institute of Technology and Management, Gurgaon, India. Her areas of interests are Data analysis, Natural Language Processing and machine learning

Nisheeth Joshi is an Associate Professor in the department of computer science at Banasthali Vidyapith, Rajasthan, India. He primarily works in Machine Translation, Information Retrieval and Cognitive Computing. He has over 12 years of teaching experience.