

Uncovering Sentiments: A Big Data Analytic Framework for Twitter Data using Unsupervised Learning

A. A. Qadir¹, S. Iqbal², M. Qabulio³, H. Jamshed⁴, K. Abid⁵

^{1,5} Department of Computer Science, NFC Institute of Engineering and Technology Multan, Pakistan

² Department of Computer Engineering, UCET, Bahauddin Zakariya University, Multan, Pakistan

³ Department of Software Engineering, University of Sindh, Jamshoro, Pakistan

⁴ Department of Computer Science, DHA Suffa University, Karachi, Pakistan

¹ kamranabidhiraj@gmail.com

Abstract- Now days, Micro blogging websites are producing enormous unstructured data due to web 2.0 technologies. This unstructured data can be used to extract feelings or interest of people. However, there is limitation to extract feelings or interest of people from web 2.0. Therefore, the aim of this paper is to describe fine grain sentiment clustering (into strong positive, positive, neutral, negative and strong negative) study on 1.6 million tweets by applying Big Data Analytic framework using unsupervised machine learning “K-means” algorithm. The framework can work both for big data or non-big data environment. The framework consists of two stages. First stage consists of phases to manage and process social media text data to establish a Machine Learning Model (MLM) and work for non-big data environment. While, second stage described Big Data (BD) architecture and data analysis phases that used MLM model of first stage to get results using Big Data Analytics as well as other BD techniques. Study provides percentage, polarities of each sentiment group with web interface of the model to find the sentiment. Results shows that 351965, 208367, 342536, 159075, 538057 tweets are Strong Positive, Negative, Positive, Strong Negative and Neutral respectively.

Keywords- Sentiment Analysis, Twitter, Big Data Analytic Framework, Machine Learning

I. INTRODUCTION

As Web 2.0 continues to expand with features such as wikis and blogs, the incorporation of such technologies would prove to be very useful. More specifically, the global microblogging sites such as Twitter, Facebook, and Instagram are sources of unstructured information in the current society and presents great opportunity area for research interests on the areas of sentiment analysis, for understanding the buying behavior of customer, assessment of teaching performances, the trends in

Google search results for decision making purposes, measuring out broke health information, and examination of people’s attitude towards specific issues like the Coronavirus [1-2]. These platforms allow users to state their opinions on topics of interest, topics based on current events or politics, and personal attitudes towards a particular product or service and this makes it useful for sentiment analysis [3].

Sentiment analysis, which is defined as the process of extracting valuable information relevant to opinions from the raw text, is also highly relevant for the assessment of opinions. Since over 80% of the information existing on the web is unstructured, this analysis has gained credit across various discipline. Sentiment analysis is, for example, useful for companies who use this information to determine the general public opinion about their products [4], [5]. Supervised machine learning techniques, which are developed with a correlated dataset, are more used in the sentiment analysis. However, these techniques might be, the problem of getting large amounts of labeled data might be a major drawback[6]. On the other hand, there is no need for a data set of labeled instances in the case of unsupervised machine learning approaches [7]. However, their approach to sorting data is more rigid, collecting it in categories based on the similarity of its properties. K-means clustering is one of the leading algorithms used in unsupervised clustering, though literature reveals a lower presence of unsupervised learning methods used in sentiment analysis [8]. A few past works have formulated the problem of discovering related articles and creating potential classification models and a few recent research studies have suggested that it is possible to use both supervised and unsupervised learning for this purpose[9].

Based on these thoughts, this research hereby incorporates the unsupervised machine learning technique, k-means clustering, on sentiments analysis of tweets. To do so, we introduce a computational paradigm that is a Big Data Analytic

(BDA) model for sentiment clustering to sectors with both BIG DATA and sectors without BIG DATA[2]. It identifies two phases that form the structure of the framework. The first phase covers the areas of managing and processing social media text data in order to develop a machine learning model. The second stage explains further on the structure and analysis steps in the Big Data (BD) platform namely; Big Data architecture and analysis employing the MLM Model to develop results with the use of Big Data Analytics plus more techniques in BD.

II. LITERATURE REVIEW

The changing dynamics of creating and sharing content can be discussed before and after the appearance of Web 2. Earlier, internet was comprised only the technological sites, and the users were mere receivers of the information presented in the sites. Nevertheless, the deep integration of ‘Web 2.0, social media platforms emerged as people became active makers of content. These changes of the user roles in turn created tremendous amount of unstructured data containing from multiple social media such as the Twitter[3-4]. Among several social networks, the most important source in terms of contribution to this data pool is Twitter as it can provide essential information about people’s interests and opinions by the usage of methods like Sentiment Analysis. Sentiment analysis, which originated from the year 2003, aims at computing the polarity and strength of sentiment of textual feedbacks that have not been structured for opinions. For instance, Mining Hu of the University of Illinois at Chicago and Bing Liu have created positive and negative word lists for the sentiment analysis, which helped to contribute to the formation of theories [10-11].

The number of social media sites which include the fast-growing media outlet commonly known as twitter has grown exponentially and this has significantly contributed to growth in the masses of unstructured data that needs to be analyzed. This has prompted the creation of sound methods in order to make a lot of information useful, which has been made available through this flood of information. Conventional methods of sentiment analysis used widely fall under the domain of supervised learning and require data sets that has been labeled and vetted as well as presumptions on set categories, making them not so versatile in dynamic environments [11]. The latest progress in the unsupervised machine learning approach has presented more viable additional solutions that can identify patterns and sentiments on their own without using any labels[12]. Therefore, the unsupervised technique shall be embraced as they allow the identification of structures in the data that can demand less computational power than other techniques like

sentiment analysis. This shift is quite relevant for big data analysis, especially, when the amount and speed of data incoming necessitate working with large data sets which cannot be annotated manually. In the domain of Twitter data analysis, there has been proven practice of studying sentiments and trends by supervised learning frameworks. This has been done using Latent Dirichlet Allocation (LDA) and K-means clustering to determine the current topics and sentiment analysis making it easier to determine the general opinion of the people and some of the discussions made. For instance, LDA has been applied in the analysis of tweet corpora and other textual data by identifying areas of topic and theme, as well as underlying sentiment patterns that are associated with certain events or subjects. Closely to cauterization, there are also algorithms used in grouping of tweets based on the similarity of used sentiments, including the identification of emergent sentiments without prior classification [12]. Unlike the previous methodologies, these approaches do not only fine-tune the analysis, reaching the microlevel of sentiment analysis but also provide an ability to consider the new data[13]. While the social media is rapidly progressing the introduction of the unsupervised learning in the big data analytic frameworks is going to be more and more essential that will provide a modern set of tools for the sentiment’s identification in the logically consecutive expanded digital environment.

Thus, while in the recent past, organizations were just concerned with the accumulation of information, nowadays they need to analyze data to generate useful knowledge. To achieve this, various sources have developed frameworks that can help researchers to handle the issue. Some of the frameworks used lays emphasis on classification of data gotten from Twitter by using diverse models all in the interest of categorizing the sentiment [14], [15]. While some possess ML algorithms such as Naïve Bayes, Neural Net and others, they are not aimed at non big data situations. Other frameworks which have been used to categories sentiments in particular domains include the context-based framework, the domain-based framework, the culture-based framework, Malayalam tweets solely using supervised machine learning models including Naïve Bayes, Support Vector Machine and Random Forest [10], [11], [14]

Table 1: Literature Overview

Ref#	Dataset	Language	ML approach	Framework	Include Emoticon	Web Interface
[2]	twitter	English	supervised	Work for non-big data	no	no
[1]	twitter	Arabic	supervised	Work for non-big data	no	no
[4]	twitter	English	supervised	Work for non-big data	no	no
[4]	Yelp	English	supervised	Work for both big or non-big data environment	no	no
[16]	Twitter	English	supervised	Work for non-big data	yes	no

III. METHODOLOGY

The extensive use of social media especially the twitter has accelerated the rate at which a large amount of unstructured data is generated. This growth has called for the need to come up with better ways of analyzing this big heap of data that has been generated. Conventional approaches to sentiment analysis that mainly use supervised learning methods are highly dependent on training data and categories; therefore, they may not be very helpful in developing and constantly changing environments. Modern studies in the area of unsupervised machine learning have offered more viable solutions that can learn patterns and sentiments without requiring any labeled data. In contrast to the supervised learning techniques, the clustering and topic modeling do not require defined classes and provide more adequate and more diverse solution for sentiment analysis. This change is especially relevant when it comes to the big data analytics since the nature and the amount of data make manual processes of annotating them impossible.

The main acknowledgment is that generic pre-processing qualitatively prepared the data for analysis to ensure a better consistency of a dataset. Some of the steps that followed include; deduplication, filtering out non-English tweets, conversion of all the words into lower cases, removal of highly non-informative words known as ‘stop words,’ URLs, and special characters. All text transformations were preprocessed first for tokenization – splitting the text into words or phrases; then, stemming or lemmatization into base forms was done. Although the paper discusses these data pre-processing steps in a general way, there is no clear identification of the used techniques and justifications for their application. More details in relation to the manner in which the data was cleaned and preprocessed is lacking, which affects the sample’s reliability and its detail, these additional inputs should help in ensuring that other researcher replicate the sample in an accurate manner.

Data Integration and Big Data Analytics Framework:

The framework consists of two stages. A data integration stage and a Big Data Analytics stage as shown in figure 1.

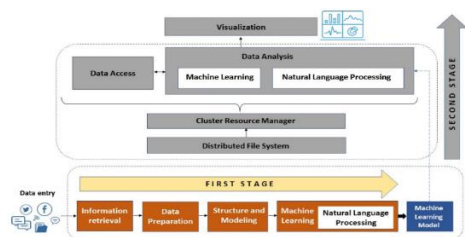


Figure 1: Data integration and big data analytic framework

The first subsidiary step involves all stages between data input and the creation of the Machine Learning Model (MLM) as described in section 3 as illustrated in figure3 above. It encompasses several phases, in which it seeks to address the challenges posed by Big Data by isolating a BSD and work on this instead; which may be obtained by sampling the Big Data sources for better results. Specifically designed for text data analysis, this framework undergoes several phases: Specifically designed for text data analysis, this framework undergoes several phases:

Phase 1: Data Gathering: The first process is of data collection where information is obtained from a given source which can be original and normally in textual form such as, microblogging services or social media platforms. However, as mentioned earlier in this paper, Twitter data is the lens through which the aforementioned dynamics are viewed and analyzed.

Phase 2: Information Retrieval: Information access stage involves processes to obtain information where API is used commonly. Collection of data must also be current and up to date, the following issues may however arise; Data unavailability or data noises which are likely to shift the time frame of the project.

Phase 3: Pre-processing: In this step, the data collected is pre-processed with an aim of eliminating any unwanted information that may end up compromising the results obtained. Data cleaning, a process included in the project and covering approximately 70 percent of the overall project, encompasses dealing with outliers. If this phase is not tackled, the outcome will shift from the anticipated result thus the need of this phase.

Phase 4: Structural and modelling: In the course of this phase, a data model is established to mirror the output developed in the former phase. This model defines the structure and the contents within the systems mentioned above. An initial data model is easier to create when one is dealing with only one source of data but it is a challenge when working with many sources of data. Algorithmic methods are then applied to the data model for further processing.

Phase 5: Machine Learning and Natural Language Processing: In this step, text containing unstructured data prepared in the earlier step are analyzed by incorporating both Machine Learning (ML) and Natural Language Processing (NLP) techniques. These two fields are slowly but surely changing the face of numerous industries while improving the ability to process data. As recognized by the scientific community, artificial intelligence, especially ML, as well as NLP have the uses in recommender systems, sentiment analysis, and user reviews.

Phase 6: Machine Learning Model: The output of the ML and NLP phase to creates the substantiation

for the Machine Learning Model (MLM). It is a great tool because with the help of this model, one can work with existing datasets and perform feature extraction as well as sentiment analysis. Sophistication of ML algorithms and the inclusion of other strategies help to increase the model's effectiveness. Hence, it is compatible with non-big data systems, and delivers enhanced outcomes without complex processing systems. In addition, the MLM plays the role of supporting Big Data analysis as a base in the Machine Learning phase within the second stage.

IV. RESULTS

This is the reason we have formulated the following research questions; – What are the different positive sentiments expressed by the users of the social media platforms in relation to the 1. Tomorrow, they plan to present their analysis of 6 million tweets collected from randomly selected Twitter users. The research seeks answers to several key questions: 1) To discuss a concrete methodology on how to attain a Big Data classification that would be fine-grained in regard to the sentiment to be identified. 2) How does each sentiment group relate to the overall polarized sentiment for the message being conveyed? 3) What other words are typically associated with each positive, negative or mixed sentiment set? Therefore, the research objectives of the study are as follows: To apply a Big Data Framework for sentiment clustering the study will use a big data analytic framework while using the k-means algorithm. Drawn here are the findings that were realized following the steps of the first stage of the research. The following unveils all the packages and source materials used by the researchers in the study:

Phase 1: Data Entry

The source whereby this data was obtained was downloaded from the Kaggle website, and specifically the Twitter data set. A study revealed that both Python and R languages can support ML but due to prior acquaintance with the Python language, researchers chose Python. Python is composed of more than seven thousand interfaces and libraries, which helps solve research problems. Based on what was highlighted above as the thesis of the study, the following Python libraries have been used in this study: Pandas for data manipulation and analysis, neat text for cleaning the text, modules (for preprocessing and using NLP tools), Vader Sentiment (for the identification of sentiment polarity), sklearn (for importing clusters), matplotlib pyplot. The tools utilized during analysis included; pandas (for data manipulation), NumPy (as a numerical computation tool), seaborn and matplotlib (for data visualization) and word cloud (for word-cloud visualization).

Phase 2: Information Gathering Phase

Information Retrieval Phase Carefully and thoroughly read all types of materials that are related to the subject of the proposal. The goal of this phase is to make sure that data is available for processing purposes Data availability implies that data exists in its original form and can be easily accessed by the system. Therefore, the given dataset was uploaded to “Google Colab” as well as the “pandas” library in Python was employed to the reading of the data. The data from Twitter analysis was large and constituted of 227 MB; nonetheless, the entire data was uploaded in Google Colab.

Phase 3: Data Preparation

Specifically, in their study, the authors used the following theoretical and methodological approaches: The dataset comprised 1. 6 million tweets in the form of the CSV that does not indicate the column named. Therefore, the researchers gave some aliases to the columns where they filtered the data from as Sentiment, id, Date, Query, User, and Tweet.

Phase 4: Structure and modeling

As the research is on Clustering sentiments of tweets, there was no need of Sentiment, id, Date, Query and User column. So, it was necessary to delete that column from the dataset and the first fifteen tweets of the final dataset with the remaining column is shown below.

	Tweet
0	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	is upset that he can't update his Facebook by ...
2	@Kenichan I dived many times for the ball. Man...
3	my whole body feels itchy and like its on fire
4	@nationwideclass no, it's not behaving at all...
5	@Kwesidei not the whole crew
6	Need a hug
7	@LOLTrish hey long time no see! Yes.. Rains a...
8	@Tatiana_K nope they didn't have it
9	@twittera que me muera ?
10	spring break in plain city... it's snowing
11	I just re-pierced my ears
12	@caregiving I couldn't bear to watch it. And ...
13	@octolinz16 It it counts, idk why I did either...
14	@smarrison i would've been the first, but i di...
15	@iamjazzyfizzle I wish I got to watch it with ...

Figure 2: Final Dataset

Phase 5: Machine learning and natural language processing

Tweets of twitter user was necessary for ML phase. But additional cleaning is also required because of URLs, user mention and hashtag words, nulls, blanks, special characters, numbers, currency symbols and stop words present in twitter data. “Neat text” library provides neat text. functions module for handling such things. So, neat text.

functions were used to remove URLs, user mention and hashtag words, extra white spaces. Special characters, stop words, punctuation and emoticons were not removed from the dataset because VADER handle them well as shown in figure 3. VADER were used for finding sentiment polarity of tweets. VADER work well with uppercase text data. So, all tweets were converted to uppercase as shown below. Final dataset consists of 1599999 tweets.

```
analyzer.polarity_scores("- Annu, that's a bummer. You shoulda got David Carr of Third Day to do it. Wink ")
{'neg': 0.14, 'neu': 0.86, 'pos': 0.0, 'compound': -0.3818}

analyzer.polarity_scores("- ANNU, THAT'S A BUMMER. YOU SHOULD GOT DAVID CARR OF THIRD DAY TO DO IT. WINK ")
{'neg': 0.14, 'neu': 0.86, 'pos': 0.0, 'compound': -0.3818}

analyzer.polarity_scores("- ANNU, THAT'S A BUMMER. YOU SHOULD GOT DAVID CARR OF THIRD DAY TO DO IT. WINK ")
{'neg': 0.172, 'neu': 0.828, 'pos': 0.0, 'compound': -0.516}
```

Figure 3: VADER work well with stop words, punctuations and uppercase tweets

Phase 6: Machine Learning Model

Cluster module from “sklearn” library was import to make the clusters for sentiment analysis purpose. “K-means” algorithm is selected for clustering tweets as described earlier. Initially ten cluster were made. Then elbow method was used to select the optimal value of clusters and it was three as Positive, Negative, Neutral. Elbow method shown in figure 4. While, the researcher is interested in fine grain sentiment clustering as strong positive, positive, neutral, negative, strong negative. So, five number of clusters was chosen. Moreover, it can be seen from the elbow method that graph become smoother at value five.

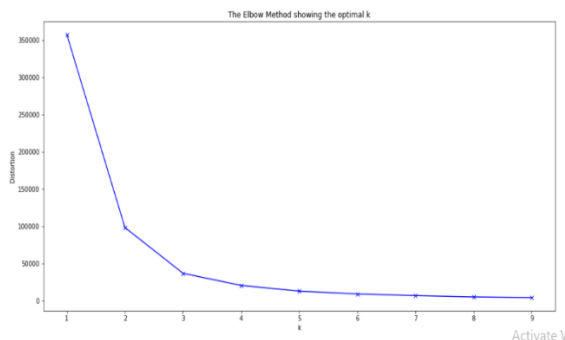


Figure 4: Elbow method for optimal value of cluster

V. DISCUSSION

Every one of the five clusters with centers depicted in the figure is characterized by different sentiment types. The first cluster, which is at index 0, has as its center, 0. It is equal to 76647436 and characterizes those tweets containing words with a very positive correlation. The second cluster, located at the second index, was -0. It is equal to

37926272 and includes tweets with negative attitude.

The third cluster which is closest to the origin has an index number of 2. Information derived from the database associated with PAN number 00192567 shows no bias, and therefore the text is completely neutral. The fourth cluster is also the last one and, in terms of the index, it is the 3-nd one with the center 0. The set 4382477 is a set of tweets that contained positive sentiment. Regarding the last cluster, at index 4, has -0.69052762 and refers to tweets with highly negative polarity. The distribution of tweets among these clusters is as follows: the first cluster have 351965 tweets while the second cluster have 208367, the third cluster have most significant number of 538057, the fourth cluster have 342536, and the last cluster 159075.

```
c.cluster_centers_
array([[ 0.76647436],
       [-0.37926272],
       [ 0.00192567],
       [ 0.4382477 ],
       [-0.69052762]])
```

Figure 5: Tweets fall into strong negative categories

```
df['result'] = wcss
df.result.value_counts().head()
2    538057
0    351965
3    342536
1    208367
4    159075
Name: result, dtype: int64
```

Figure 6: Tweets fall into strong negative

It can be seen that more than three lac and fifty thousand tweets are Strong Positive and their polarities are between 0.6 and 1.0. Scatter plot of Negative tweets shows that more than two lac tweets are Negative and their sentiment polarities lies between greater than -0.55 and less than -1.0. Results show that more than five lac tweets are neutral and their polarities are greater than 0.2 and less than 0.23., it can also be seen that more than three lac and forty thousand tweets are Positive and their sentiment polarities are greater than 0.2201 and less than 0.6023. More than 1, 50,000 tweets are Strong Negative and their polarities are range between -0.9991 and -0.5349 all can be seen from scatters diagram Table 2 and 3.

Table 2: Sentiment Clusters and Polarity

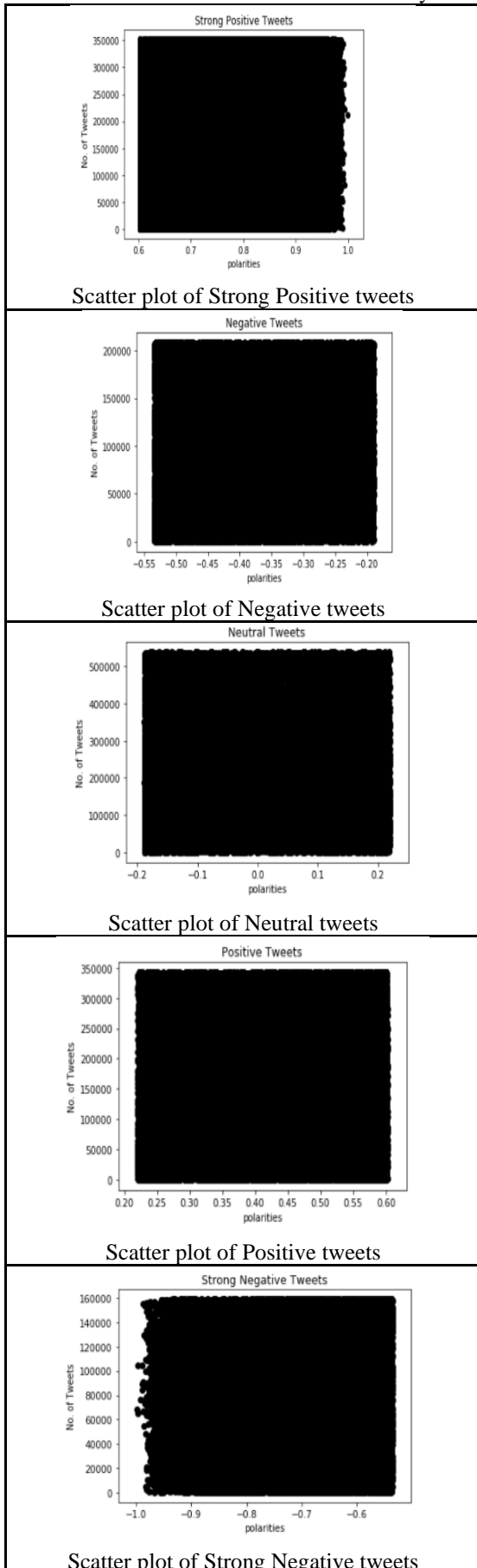
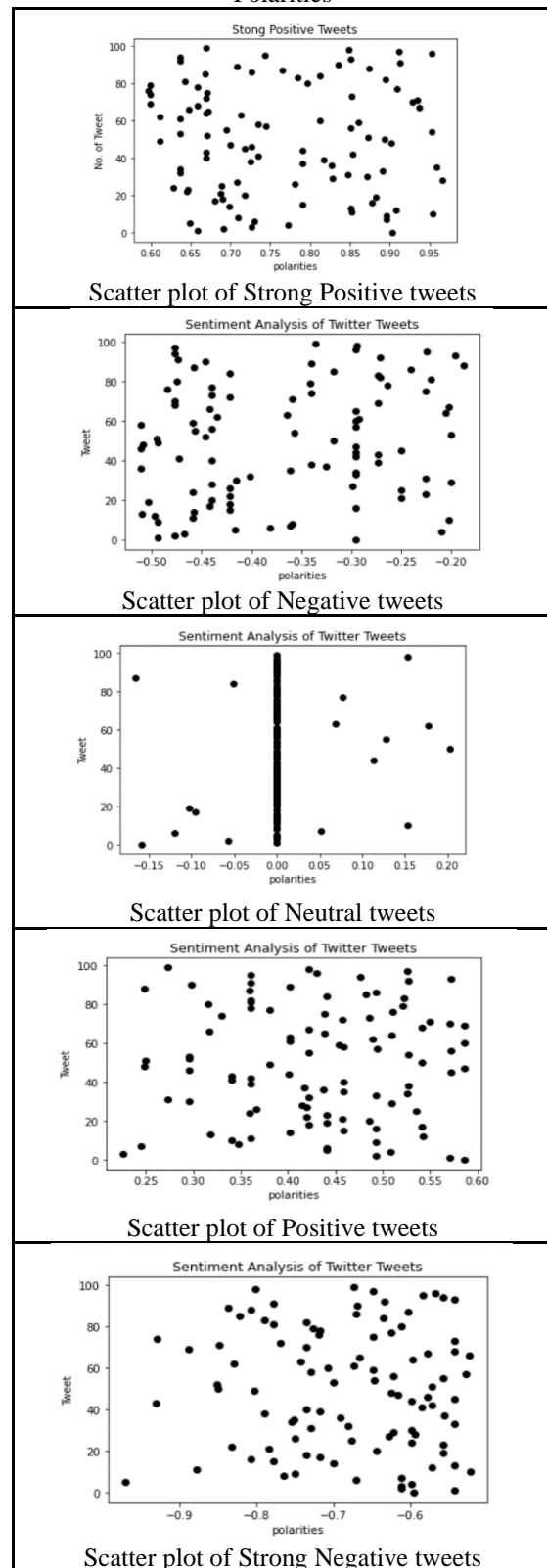



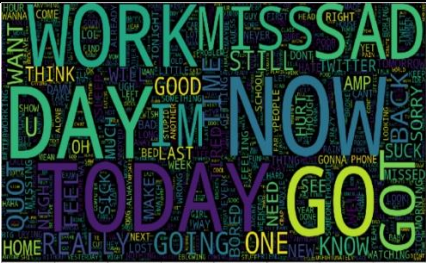
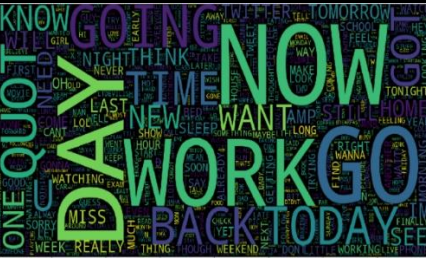
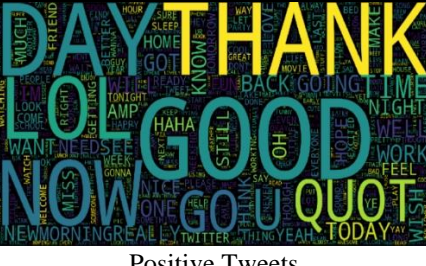
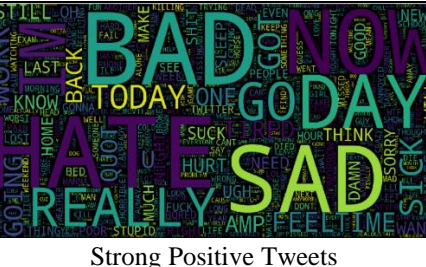
Table 3: Scatter plot of Sample Tweets and their Polarities



Words that were used in Strong Positive tweets are Good, Thank, Great, well, Awesome, Haha, Happy, Cool etc., words that used in Negative tweets are Missed, Suck, Sad etc., words that were used in Neutral tweets are Day, Work, Go, Time etc., words

that used in Positive tweets are Good, Thank, lol, Thinking, Work etc. and words in Strong Negative tweets are Bad, Hate, Suck, Sad, Hurt etc. Word clouds of all types of tweets can be seen from the scattered plot Table 4.

Table 4: Words in Category of Cluster

 <p>Strong Positive Tweets</p>
 <p>Negative Tweets</p>
 <p>Neutral Tweets</p>
 <p>Positive Tweets</p>
 <p>Strong Positive Tweets</p>

Percentage of positive, strong positive, neutral, negative and strong negative tweets can be seen from Figure 1. Pie chart shows that 33.6% tweets are

neutral, 13.0% tweets belong to negative, 22.0% tweets belong to strong positive, and 9.9% percent tweets are strong negative, while 21.4% tweets are positive.

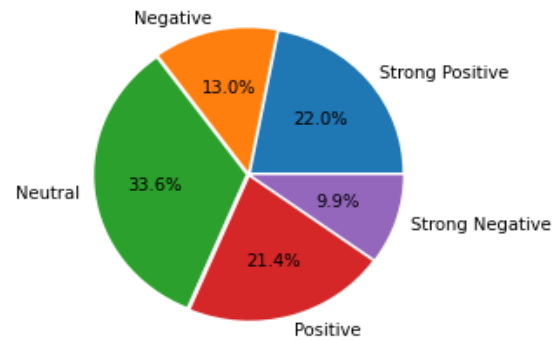


Figure 7: Pie Chart of all Clusters

VI. CONCLUSIONS, LIMITATION AND FUTURE WORK

The framework that was used in this research is big data analytic framework which works both for big data environment and non-big data environment. In this research, researcher analyze the sentiments of Twitter data and determine its nature as strong positive, positive, neutral, negative, strong negative. Cluster k1 0.76647436 representing Strongly Positive Sentiments and contain 351965 tweets which is 22% of tweets. And their polarities are between 0.6 and 1.0, Cluster k2 which is -0.37926272 representing Negative Tweets and containing 208367 which is 13% of tweets and polarities range are between -0.55 to -1.0, Cluster k3 0.00192567 representing neutral sentiments and containing 538057 tweets which is 33.6% of the dataset and polarity range is -0.2 to 0.23. Cluster k4 0.4382477 representing positive and containing 342536 tweets which is 21.4% of tweets and polarity range is 0.2201 to 0.6023. Cluster k5 -0.69052762 representing strongly negative sentiment and containing 159075 tweets which is 9.9% of the total data and polarity range is -0.9991 and -0.5349. Research provided interface which can work on text, emoticons and emoji's as well.

This research is limited to fine grain sentiment analysis (five categories as Strong positive, positive, neutral, negative and strong negative). Also, this research has been limited to twitter dataset. Future research can be on other resources like Facebook, WhatsApp, Messenger, and other social media etc. Research may be conduct by Introducing performance and accuracy parameters. Also, Future research will be on other algorithms for unsupervised learning. Future research using this framework can be on reinforcement learning. Future work may be on other languages such as Arabic, Spanish and Hindi.

REFERENCES

- [1] M. Saad, R. Khosla, N. Zaki, N. Al-Masri, and I. Aljarah, "Arabic sentiment analysis: A systematic review," *Inf Process Manag*, vol. 60, no. 2, p. 103093, 2023.
- [2] A. Qazi, T. Ali, M. Khushi, I. Siddique, R. Khalid, and H. Ahmad, "Sentiment analysis of hybrid network-based social media big data using a context-aware convolutional neural network." *Sustain Cities Soc*, vol. 82, p. 103915, 2023.
- [3] Y. Zhang, Y. Zhang, and J. Zhang, "Towards interpretable sentiment analysis: Combining aspect extraction and sentiment classification with deep neural networks," *Neurocomputing*, vol. 485, pp. 215–227, 2022.
- [4] S. Pawar, P. Nikam, and K. Kotecha, "Multi-aspect sentiment analysis using reinforcement learning-based capsule network model," *Knowl Based Syst*, vol. 234, p. 107571, 2022.
- [5] M. Al-Ayyoub *et al.*, "Arabic sentiment analysis using bidirectional recurrent neural network and convolutional neural network," *J Ambient Intell Humaniz Comput*, vol. 13, no. 1, pp. 663–675, 2022.
- [6] I. Arpacı, O. Turetken, and S. Ozkan, "Analysis of Twitter data using evolutionary clustering during the COVID-19 pandemic," *Computers, Materials & Continua*, vol. 65, no. 1, pp. 193–204, 2020.
- [7] M. Bibi, W. Aziz, M. Almarashi, I. H. Khan, M. S. A. Nadeem, and N. Habib, "A cooperative binary-clustering framework based on majority voting for Twitter sentiment analysis," *IEEE Access*, vol. 8, pp. 68580–68592, 2020.
- [8] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. Lehmann, "An 'infodemic': Leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak," *Open Forum Infect Dis*, vol. 7, no. 7, 2020.
- [9] F. Z. Kermani, F. Sadeghi, and E. Eslami, "Solving the Twitter sentiment analysis problem based on a machine learning-based approach," *Evol Intell*, vol. 13, no. 3, pp. 381–398, 2020.
- [10] M. Umar, A. M. Qamar, S. M. Anwar, and A. Hassan, "Multimodal sentiment analysis using deep learning techniques: A review," *IEEE Access*, vol. 9, pp. 29523–29537, 2021.
- [11] A. Hassan, A. M. Qamar, and S. M. Anwar, "Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression," *Artif Intell Rev*, vol. 54, no. 1, pp. 573–602, 2021.
- [12] M. Khurram Iqbal, K. Abid, M. fuzail, S. din Ayubi, and N. Aslam, "Omicron Tweet Sentiment Analysis Using Ensemble Learning", doi: 10.56979/402/2023.
- [13] F. Malik *et al.*, "A Hybrid Machine Learning Model to Predict Sentiment Analysis on í µíµ• ", doi: 10.56979/602/2024.
- [14] S. Al-Saqqa, A. Awajan, and N. Al-Rajebah, "Unsupervised sentiment analysis approach based on clustering for Arabic text," *Procedia Comput Sci*, vol. 170, pp. 4243–4254, 2020.
- [15] M. I. Dar and G. Naymat, "The impact of applying different preprocessing steps on review spam detection," *Procedia Comput Sci*, vol. 113, pp. 273–279, 2020.
- [16] X. Liang, Z. Li, C. Wang, Y. Zhang, J. Tang, and M. Xu, "Learning sentiment distribution with hybrid deep sentiment network for microblog sentiment analysis," *Applied Intelligence*, vol. 52, no. 1, pp. 960–975, 2022