Exploring Twitter News Biases Using Urdu-based Sentiment Lexicon

Kamran Amjad[1], Maria Ishtiaq[1], Samar Firdous[2], Muhammad Amir Mehmood[1]
[1]{KICS, UET Lahore, Pakistan} , [2]{KEMU Lahore, Pakistan}
{kamran.amjad, maria.ishtiaq, amir.mehmood}@kics.edu.pk , dr_samarfirdous@yahoo.com

Abstract— **Social media has become a tremendous success in recent times. It has enabled people to keep in touch with each other anytime and anywhere around the globe. People share their opinions and experiences publically through such platforms. Twitter is one of the significant social networking platform that is used by many news media to disseminate breaking news instantaneously. Sentiment analysis of social media is successfully used to gain insights regarding collective behavior of the society. In addition, sentiment analysis is used to detect positive or negative content posted on the social media for various purposes such as riots detection.**

In this paper, we focus on the sentiment analysis of news tweets in Urdu language by major news sources in Pakistan. By gathering tweets data over the period of 10 months, we built a sentiment lexicon in Urdu language. Moreover, we devise an algorithm that classifies Urdu text into positive, negative, or neutral classes based on the cumulative sentiment score of the text. Our sentiment analysis algorithm achieves 77% accuracy. Furthermore, we have done perspective analysis in which we estimated the bias in the news reporting through tweets with respect to the government with 77.45% accuracy.

Keywords— Urdu News, Sentiment Analysis, Twitter, News Bias

I. INTRODUCTION

Today online social media such as Facebook, Twitter, Instagram, Snapchat etc., have gained tremendous popularity among people of all ages. It has been reported that Facebook alone had around 1.86 billion users at the end of 2016 [1]. Similarly, Twitter has 328 million monthly active users. These social networking platforms are generating petabytes of data regularly, directly from human users across the globe. As a result of the global popularity, these social networking sites have also become essential part of our daily life. With the rise of mobile broadband Internet, it has become very convenient to express one's views regarding social, political, and economic matters on social media anytime and anywhere. As a result, people often use such platforms to spread rumors, sectarian hatred, defaming political rivals, malicious activities, and even spreading blasphemous content.

Twitter is a micro-blogging site by nature that allows its users to express themselves with a restriction that each text message can only be of maximum 140-280 characters [2]. Such short text nature forces the user to convey a to-the-point opinion or sentiment in a tweet. The latest statistics show that more than 500 million tweets are generated each day with volumes going higher in case of some real world events. Thus, we have access to huge amounts of the data related to human interactions that can be used for the better understanding of the social issues.

Sentiment analysis is a task in which we infer polarity of a sentiment expressed in a given text. Several techniques have been proposed by researchers for sentiment analysis [3] [4] [5]. Mainly, there are two major techniques used for sentiment analysis. The first approach is based on supervised machine learning techniques. In this case, appropriate features for classification based on different sentiments are extracted from text and a labeled dataset is used to train the model. The final model is then applied to a real world data. Second approach for sentiment analysis is based on sentiment lexicon. In general, sentiment lexicon is a dictionary of words used in a language with their sentiment polarity assigned to them. In this paper, we rely on sentiment lexicon approach to perform sentiment analysis on Twitter news data. In literature, numerous resources are available for sentiment lexicon in English language that provide highly accurate results [6]. However, not much effort has been spent to prepare sentiment lexicon resources for regional languages especially for the Urdu language. The lack of such resources has resulted in practically no sentiment analysis or opinion mining studies on the text that is available in Urdu language on the social media.

In this paper, we focus on the sentiment analysis of Twitter feed generated by popular news sources in Pakistan and we build a sentiment lexicon based on the vocabulary extracted from tweets in Urdu language. We aim to classify the tweeting patterns of tweets from these news agencies. First, we created our own sentiment lexicon based on Urdu language. Second, we manually labeled our test data into positive, negative, or neutral tweets. Finally, we devise an algorithm to classify tweets written in Urdu language into positive, negative, or neutral sentiments. Our main contributions are as follows:

- **Urdu News Lexicon:** We have built Urdu lexicon of 20,171 unique words from 26,614 news tweets. 12,808 effective words (nouns and adjectives) were obtained after parts of speech tagging.
- **Sentiment Analysis:** Sentiment lexicon was built by assigning polarity labels to effective words. The overall accuracy of our algorithm for sentiment analysis is 77%.
- **News Media bias:** We performed perspective analysis to find media biases with respect to government with an accuracy of 77.45%.

The remainder of the paper is structured as follows: In Section II, we discuss the related work present in the literature. Different aspects of the data acquisition process are explained in Section III. In Section IV, we describe our methodology used for sentiment analysis. We discuss our results in Section V and Section VI concludes the paper along future work.

978-1-5386-1658-1/17/$31.00 ©2017 IEEE

TABLE I. DETAILS OF TWITTER ACCOUNTS USED IN OUR STUDY

Twitter ID	User Name	Number of Followers	Number of Tweets
Daily Jang	@jang_akhbar	0.23M	7,000
BBC Urdu	@BBCUrdu__	1.14M	10,000
Dawn News	@Dawn_News	0.91M	9,417

TABLE II. STATISTICS OF THE DATASET

Characteristics of Data	Total
Number of Tweets	26,614
Number of Unique Tweets	21,991
Duration of Data Collection	10 Months(Sep, 2016- July, 2017)
Number of Unique Words	20,171

II. RELATED WORK

In the past, researchers have used Twitter data extensively for sentiment analysis in various different domains. For lexicon based sentiment analysis, specific domain knowledge is used for building a lexicon. In [5], Twitter data is used to study the propagation of a disease in a certain geographical area by fetching tweets of a particular region. Authors have analyzed words related to the condition of persons related to the disease in the text. e.g., "I am having flu" or "My flu is getting worse". SentiWordNet [6] was used as a baseline for scoring the sentiment value of words after modifying the relevant words according to the security informatics domain. In [7] authors have gathered Twitter data prior to some national security related events and performed sentiment analysis to study negative tweets being generated that can contribute to any law and order situation.. In [8], the authors used tweets by the consumers and rely on lexicon related to a particular domain to gain consumer insights about their product.

However, most of the sentiment lexicons are available in English language. Just like [6], SentiStrength is a publicly available sentiment lexicon in English that is also widely used [9]. The sentiment analysis in different regional languages has been an active area of research in recent times. In [10], authors present a lexicon in Spanish Language. Similarly, sentiment lexicon for Arabic and Persian Languages are discussed in [11] [12]. In case of Urdu language, only a few researchers have tried to build a sentiment lexicon [13] [14]. The sentiment lexicons used in these studies are built on a limited data and is not domain specific.

III. DATASET

In this section, we provide details of data gathered in our study. For this purpose, we rely on publicly available Twitter data through Twitter APIs [15] which is commonly used for research in social opinion mining and sentiment analysis. First, we selected Twitter accounts of three renowned news websites in Pakistan. These Twitter accounts have collectively 2.28 million followers and they post their content in Urdu language frequently. We acquired the data for a period of 10 months from Sep, 2016 to July, 2017 that constituted of more than 26,000 tweets written in Urdu language. The long duration of data capturing enables us to achieve the desired richness in our corpus. The details of our dataset are explained in Table I and Table II.

TABLE III. DIFFERENT SAMPLES OF URDU TWEETS

S.No	Urdu Sample Tweets
1	کالم پڑھنے کے لئے لنک پر کلک کیجیئے
2	چیف جسٹس نے خلیفہ دوم حضرت عمر کی مثال دے کر کیا کہا؟
3	پاکستان میں سائیکیبیماری میں اضافہ
4	ہیٹی میں سیلاب سے زندگی متاثر
5	خواتین کے لیے چند ایسے آئیڈیاز جنہیں پہن کر وہ نہایت خوبصورت نظر آئیں گی
6	وزیراعظم نے تحقیقاتی ٹیم سے تعاون نہیں کیا

There are numerous issues with Twitter data due to the fact that people use short hands, misspell words, and often join two words by mistake. Social media users do not usually post content by following proper language rules. For our study, we chose news sources that post content only in standard text. Moreover, we have removed emoticons and other non-standard text characters. Another problem was auto generated tweets. These tweets were pointers to some link without any sentiment notion and were tweeted regularly by these news websites. These duplicate and spam tweets were filtered from the data. Table III shows different types of tweets that we have in our corpus. The first line shows a spam tweet. The second line shows a tweet that is a question statement in nature. Sentiment classification of such statements is quite difficult. In our work, these tweets are dealt as general tweets and are labeled according to the sentiment polarity score. The third line explains the word joining problem. This was dealt in two different ways. Wherever possible, such joining were corrected and in other cases, such words were filtered out from the corpus. The fourth and fifth line obey proper language rules. If we look closely, we can see that the tweet in fourth line presents a negative sentiment. The fifth line shows a positive sentiment. In the sixth line, presence of negation reverses the overall notion of the sentiment.

The process of building a lexicon for sentiment analysis is quite extensive and requires a large amount of linguistic data and effort. A predefined set of rules for a language is employed while listing words into a lexicon. We note that for news data, the vocabulary size is small compared to the complete vocabulary in the language. The reduced vocabulary factor was our prime reason to prefer Twitter data compared to web news sources. We manually generated word-list from our 26,000 tweets by removing duplicate words. We got 20,171 unique words from the original corpus of tweets as shown in Table II. Unique word-list was generated by tokenizing the tweets based on white space delimiter and then duplicates were removed. This word-list was further refined for building lexicon.

TABLE IV. TWITTER ACCOUNTS DETAILS FOR PERSPECTIVE ANALYSIS

Twitter ID	User Name	Number of Tweets
ARY News Urdu	@ARYNewsAsiaUrdu	500
GEO News Urdu	@geonews_urdu	500
Express News	@ExpressNewsPK	500

After building the lexicon, we require test data for our work. We built test data for sentiment and perspective analysis separately. For sentiment analysis, we took a dataset comprising of 500 tweets from the same three sources as mentioned in Table I. We obtained ground truth by labeling each tweet by at least three different subjects and the final label was selected by majority vote. In cases, where we could not get a majority vote, we further asked a language expert to break the tie in order to label the sentiment. As for the perspective analysis, we obtain 1500 tweets from three other news sources as mentioned in Table IV and labeled them in the same manner as described above. For each news item, we gather same tweets from these news sources in order to ascertain their reporting bias. We gathered 500 tweets from each source regarding the same issue and their bias was determined using the bias perspective in the manner of reporting the same news by different news agencies.

IV. METHODOLOGY

In this section, we present our methodology of sentiment analysis of our Twitter data. First, we discuss the Parts of Speech (POS) tagging of the word-list obtained as described in the above section. This is done in order to extract effective words from our corpus which are words that carry some notion of positive and negative sense. Next, we describe the process of labeling the words according to the polarity. Finally, we describe our algorithm to classify the text based on the cumulative score of the sentiment.

A. POS Tagging

The major challenge of building a lexicon is the identification of the part of speech of a specific word. This process is referred to as POS Tagging. In general, the sentiment of a text is expressed using adjectives and nouns where adjectives either enhance or suppress the sentiment of a given text. In our work, we call the collection of adjectives/adverbs and nouns as "*effective words*". All words that are not tagged as adjectives and nouns are marked as "*ineffective words*". There are lots of POS tagging tools available for the English language, however, this is not the case for the Urdu Language. For POS tagging, we rely on freely available tool from Urdu Summary Corpus (USC) that claims an accuracy of 88.7% [16]. In order to enhance the POS tagging of our own data, we manually checked ineffective words generated by the POS tagger and corrected the misclassified words. Table V shows the distribution of words after POS tagging.

B. Labeling

The POS tagged word-list was further examined and labeled into positive, negative, and neutral based on the sentiment that these words express in their most generalized use. In general, nouns and adjectives are the main contributors to the notion of sentiment expressed in a text document. After the completion of the process, we obtained more than 12,808 unique nouns and adjectives/adverbs as shown in Table V. For our purpose, we ask human subjects with the Urdu language expertise to label each word as positive, negative, or neutral to generate a labeled unique word-list. We call this list as "Sentiment-Lexicon". It includes the semantic information and morphological aspects of words. Unigram model of the words was used while labeling the words.

TABLE V. STATISTICS OF EFFECTIVE WORDS

POS Tag	Total Number
Nouns	10,356
Adjective/Adverbs	2,452
Total	12,808

C. Algorithm

Next, we need an algorithm that classifies the sentiment of the overall text based on the sentiment value present in each effective word. Our algorithm consists of three stages i.e., segmentation, filtration, and sentiment score calculation. For the sentiment analysis, first we need to break a given text into tokens. In segmentation stage, the text is converted into unigram tokens where white space is used as a delimiter. Next, we compare these tokens in our sentiment lexicon list to filter effective words of the text that has some sentiment value. In addition, we assign the polarity score of (+1, -1, 0) of each effective word of the text according to our labeled data where +1 means a positive word, -1 means a negative word, and 0 means a neutral word. We obtained the overall sentiment score by adding the polarity score of the individual effective words. The formula for calculating the sentiment score is as follows

$$\text{Text Sentiment Score} = \sum (\text{Sentiment score of Effective words})$$

After sentiment score is calculated, the final step in the sentiment analysis is to check for the presence of any negation and its position of occurrence in the text. For this purpose, we have created a separate list of negation words used in the Urdu language. The algorithm checks for the presence of words in this list against the text. If negation is detected, then overall sentiment value is reversed. Line 6 in Table III explains this process.

TABLE VI. BREAKDOWN OF A SAMPLE TWEET

Sample News Tweet	اپوزیشن کا رویہ درست نہیں
Segmentation	اپوزیشن | کا | رویہ | درست | نہیں
Filtering out the Effective words	اپوزیشن | رویہ | درست | نہیں
Adjective	"درست"
Negation	"نہیں"
Perspective Word	"اپوزیشن"

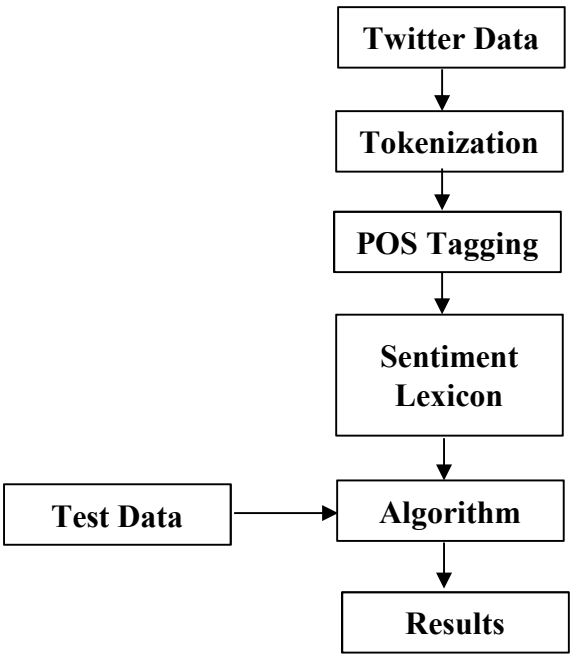

Fig. 1. Schematic diagram of methodology

Figure 1 presents a schematic diagram of the whole process. Table VI shows the process step by step and explains it by implementing the process on a sample tweet. It is evident that the presence of negation changes the sentiment of the sentence altogether.

V. RESULTS

In this section, we discuss our findings of different experiments. First, we present our results of sentiment analysis of our data. Next we describe a case study of a perspective analysis with respect to a particular domain.

A. Sentiment Analysis

In order to test the accuracy of our algorithm, we randomly selected 500 tweets from our original corpus of tweets. We compared manually labeled data as described in section II which is our ground truth data with the labels given by our algorithm. The results of sentiment analysis of our algorithm are shown in Table VII. In general, news tweets are mostly negative or neutral in nature as news agencies report news in sensational manner such as news of accidents, terrorism, crimes, and political instability. Our randomly selected data represents this trend as well. We found that our positive, negative, and neutral tweets were classified with precision of 56%, 75%, and 90% respectively by our algorithm. The precision of predicting neutral tweets is high due to the fact that in neutral tweets adjectives are mostly absent and neutral nouns are mostly present. Furthermore, the precision of negative tweets is better than positive tweets due to the reason that negative news contain explicit negative words that are easily distinguishable. On the other hand, positive news are mostly not composed of expressive words. The overall accuracy achieved by our algorithm using our sentiment lexicon is 77%.

TABLE VII. STATISTICS OF SENTIMENTS CLASSIFIED

Class Label	Ground Truth	Total Predicted	Correct Classified	Wrong Classified
Positive	67	90	50	40
Negative	194	238	180	58
Neutral	239	172	155	17

TABLE VIII. SAMPLE WORDS FOR BIAS AGAINST/TOWARDS GOVT.

Positive Words	حکومت؛مسلم لیگ ن؛وزیر اعظم؛نواز شریف؛پاکستان؛ن لیگ؛ حکومت پاکستان؛وفاق؛وفاقی حکومت؛حکومت پنجاب
Negative Words	اپوزیشن؛اپوزیشن لیڈر؛پی ٹی آئی؛پیپلز پارٹی؛عوامی تحریک؛تحریک انصاف؛لانگ مارچ؛سول نافرمانی؛متحدہ اپوزیشن؛احتساب

Figure 2 shows the performance metrics of our classification algorithm. Precision, Recall, and F1-Score are calculated individually for each of three classes present in our test data. For calculating the cumulative overall accuracy, we used the following equation.

$$\text{Overall Accuracy} = \frac{\text{True Positive} + \text{True Negative} + \text{True Neutral}}{\text{Total Number of Predictions}}$$

B. Perspective Analysis- A Case Study

The general sentiment analysis can be further enhanced by using domain specific knowledge. Perspective analysis is a hot research area in recommendation systems where one can ascertain the bias of a user towards a particular domain. In perspective analysis a domain is fixed and the text is further classified according to the domain. For instance, one may want to check reviews on any movie or a product. Another challenging field is to classify news tweets according to any organization, event, or any personality.

Next, we present a case study of perspective analysis from our data to find a media bias towards or against the government. We focused on the bias of news tweets generated by TV news channels in the favor of the government. A percentage formula was devised to detect the biasness. Positive percentage showed that a news channel is biased toward government, the negative percentage showed that the channel is biased against government and zero showed that the channel is neither biased towards government nor unbiased. Biasness was calculated by subtracting total negative sentiments from total positive sentiments and the result was then divided by the total number of sentiments.

A mathematical measure was designed to gauge the level of bias expressed by a news source. The percentage formula used to calculate the biasness is given below:

$$\text{Bias} = \left(\frac{\text{TotalPositiveSentiments} - \text{TotalNegativeSentiments}}{\text{TotalSentiments}}\right) \times 100$$

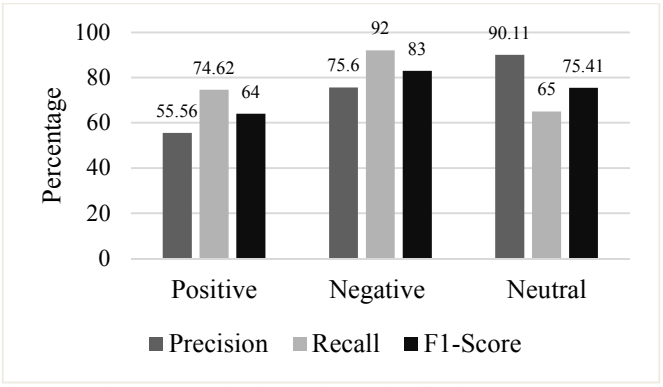

Fig. 2. Performance metrics for sentiment classification

TABLE IX. PERSPECTIVE BIAS

News Channel	Percentage Measure
Geo News	+35.47
Express News	-10.51
ARY News	-33.21

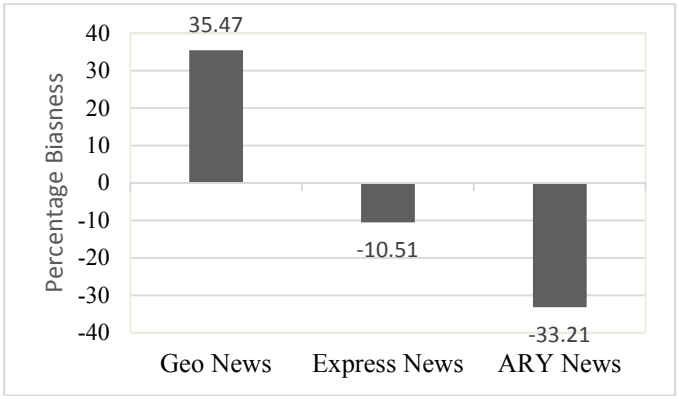

Fig. 3. Level of percentage biasness

We applied our Bias formula on 500 tweets that were collected from three media channels i.e., "Geo News", "ARY News", and "Express News" each. We prepared a list of perspective words with respect to government as shown in Table VIII. First of all, sentiment of a text was calculated and then the text was checked against the list of perspective words. If a noun has a negative perspective with respect to government then the polarity of general sentiment was reversed. Table IX shows the bias results. Our findings show that "Geo News" was biased in favor of the government and "Express News" is slightly biased against the government. However, "ARY News" was highly biased against the government. Overall, the accuracy of our perspective analysis is 77.45%.

Figure 3 presents a graphical representation of the findings that Table IX shows. We can see that there is a total reversal of biasness in cases of "Geo News" and "ARY News". Figure 4 shows the performance metrics for the Perspective analysis.

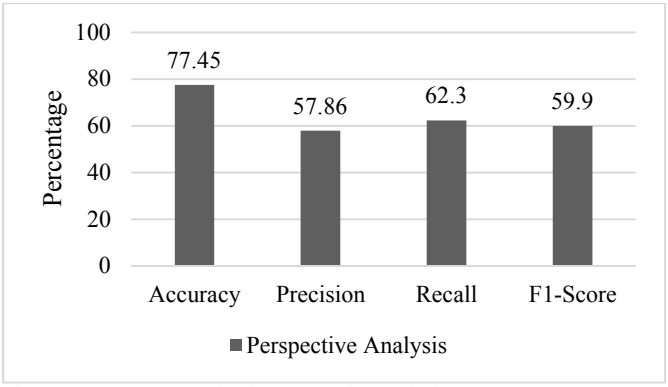

Fig. 4. Performance metrics for perspective analysis

VI. CONCLUSION

In this paper, we have studied news tweets from major news sources in Pakistan. Our core contribution is to devise a methodology to classify Urdu news tweets' text into positive, negative, or neutral sentiment. We relied on lexicon-based approach to build an extensive sentiment lexicon in Urdu language. Furthermore, we presented an algorithm to perform domain knowledge based perspective analysis. In our case we performed our analysis from the perspective of the government. We note that our methodology is general in nature and can be applied with respect to any domain. In future, we plan to implement machine learning approach to learn more features to obtain better results.

ACKNOWLEDGMENTS

This work is done under funded project "Urdu Search Engine" by Ignite National Technology Fund, Pakistan.

REFERENCES

[1] Facebook, "Facebook Newsroom," [Online]. Available: https://newsroom.fb.com/company-info/.

[2] Twitter, "Twitter Support," [Online]. Available: https://support.twitter.com/.

[3] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002.

[4] C. Musto, G. Semeraro and M. Polignano, "A comparison of lexicon-based approaches for sentiment analysis of microblog posts," *Information Filtering and Retrieval,* vol. 59, 2014.

[5] V. Carchiolo, A. Longheu and M. Malgeri, "Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics," in *International Conference on Information Technology in Bio-and Medical Informatics*, 2015.

[6] S. Baccianella, A. Esuli and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for

Sentiment Analysis and Opinion Mining," in *LREC*, 2010.

[7] A. Jurek, Y. Bi and M. Mulvenna, "Twitter sentiment analysis for security-related information gathering," in *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*, 2014.

[8] W. Chamlertwat, P. Bhattarakosol, T. Rungkasiri and C. Haruechaiyasak, "Discovering Consumer Insight from Twitter via Sentiment Analysis," *J. UCS,* vol. 18, no. 8, pp. 973--992, 2012.

[9] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the Association for Information Science and Technology,* vol. 61, no. 12, pp. 2544--2558, 2010.

[10] M. Amores, L. Arco and C. Borroto, "Unsupervised Opinion Polarity Detection based on New Lexical Resources," *Computacion y Sistemas,* vol. 20, no. 2, pp. 263-277, 2016.

[11] H. S. Ibrahim, S. M. Abdou and M. Gheith, "Sentiment analysis for modern standard Arabic and colloquial," *International Journal on Natural Language Computing (IJNLC),* vol. 04, no. 02, pp. 95-109, 2015.

[12] M. E. Basiri, A. R. Naghsh-Nilchi and N. Ghassem-Aghaee, "A framework for sentiment analysis in persian," *Open Transactions on Information Processing,* vol. 1, no. 3, pp. 1--14, 2014.

[13] Z. U. Rehman and I. S. Bajwa, "Lexicon-based sentiment analysis for Urdu language," in *Sixth International Conference on Innovative Computing Technology (INTECH)*, 2016.

[14] A. Z. Syed, M. Aslam and A. M. Martinez-Enriquez, "Lexicon based sentiment analysis of Urdu text using SentiUnits," in *Mexican International Conference on Artificial Intelligence*, 2010.

[15] "Twitter Streaming API," [Online]. Available: https://dev.twitter.com/streaming/public.

[16] B. Jawaid, A. Kamran and O. Bojar, "A Tagged Corpus and a Tagger for Urdu," in *LREC*, 2014.