IEEE SEM

Boolean-Rule Based Sentiment Analysis Model for Election Prediction*

Maria Bangash, Dr Abdus Salam, Dr Javed Iqbal Bangash, Anum Waheed

[1](1st Affiliation) Department Name of Organization, Name of Organization, City, Country; [2](2nd Affiliation) Department Name of Organization, Name of Organization, City, Country.
Email: address desired (without hyperlink in E-mail)

ABSTRACT

For over a decade, data is extracted from social media platforms as Twitter to forecast the results of elections based on the positive, negative or neutral sentiments of people towards a political party. Sentimental analysis of informal texts such as tweets still remain a challenging task owing to their complex nature. Through this analysis technique, sentimental analysis or opinion mining will be used to explore the tweets of people published in Roman Urdu to analyze the sentiments and subjectivity. To deal with the challenges of sentimental analysis, a lexicon-based approach is used that includes the application of Boolean-based data analysis. The results generated through the extraction and filtering of tweets in Roman Urdu shows a positive correlation between the political party garnering highest positive tweets and its winning in the elections.

Keywords : Sentiments, Roman Urdu, Sentimental Analysis, Opinion Mining, Lexicon-based approach, Boolean

1 Introduction

1.1 Background

Recently, social medial platforms as Facebook, Twitter, and Instagram have emerged as one of the popular social-networking sites that people utilize to express their emotions, opinions, beliefs or sentiments. A close relationship is witnessed between the social media data and election results due to the extensive use of social media. These technologies are computer-mediated that enables millions of people to share their feelings and sentiments over the common topics. Through these actions that are constantly accumulating on social media can garner large-capacity, wift-acceleration, great-value, wide-variability, wide-range and huge-complexity data term as big social data [1]. The kinds of data used in various domains are medical, political and marketing [4]. Many researchers use this domain in various fields like fraud detection, market basket analysis and sentiment analysis.

Globally, many companies and consumers can utilize the ever-growing opiniated data to make decisions; however, millions of opinions are published each day that makes seeking manually and identifying it as negative or positive sentiment an impractical approach [3]. It increases the need and demand of sentiment analysis, which is an emerging and trending fields in social emdia for its opinions, sentiments, emotions and attitude analysis. Sentiment analysis is great value not just for research or computational purposes in critical circumstances. Significant catastrophic disasters or political changes are also included [2]. Nowadays several researcher works in sentiment analysis on different social media platform like Twitter, Facebook, Myspace, Digg, JISC listservs on the academic side etc. On these platforms, people share their opinion in different language.

Pakistan is home to dozens of languages that used to communicate on social media. Over the years, sentiment analysis is also classified as opinion mining that determines and extracts opinion (positive, negative, and neutral) and information about a related topic. Moreover, researches have the opportunity to select between three sentiment analysis approaches that are lexicon-based approach and machine learning approach [5].

In lexicon-based approach, the measure of polarization is the given content from the sentiment orientation words or phrase in documents. The aim of this approach is to identify the sentiment word or opinion expressed by user whether the words present in positive, negative or neutral. In contrast to the lexicon-based approach, machine learning can be unsupervised, semi-supervised, and supervised that demands training prior to data mining [6]. In sentiment analysis, its can be performed through classification algorithms that are further segmented into linear classifiers, decision tree, and probabilistic classifier. There are many types of probabilistic classifiers in supervised machine learning approach [7]. The approaches of supervised-machine learning give a good accuracy and experiential classification Naïve Bayes, type of method are less effective, as the machine learning approach method is limited struggle human labeled documents and quality and quantity of datasets [8].

1.2 Problem Statement

Social media platform as Twitter can emerge as critical tools to predict or forecast the outcomes of general elections in the country. Twitter users show their sentiments and emotions towards the influential topics in elections. In this study, the election prediction will be presented through the total number of positive, negative, and neutral tweets in Roman Urdu regarding a polictical party. Further, sentiment analysis was performed to analyze the tweets and predict their labels of three major parties in Pakistan: PTI, PPPP, and PML-N. In the past years, the sentiment analysis was performed on Tweets in English language; however, since in Pakistan most of the people use Roman Urdu on social media, the Roman Urdu Tweets were filtered and analyzed.

1.3 Objectives

The aim of the proposed research is to classify sentiment-orientated words from the Twitter Data present in Roman Urdu. The following objectives are set to achieve the aforementioned goal.

I. To organize the data in lexicon based approach of sentiment analysis for a given context in Urdu roman language.
II. To develop a Boolean rule based mechanism for the classification of opinion words in social media data.
III. To assess the proposed mechanism for accuracy using recall, F-measure, and precision.

1.4 Scope

Primarily, the scope of the research entails classification of data as Tweets in Roman Urdu that present positive, negative, or neutal opinion towards the major political parties in Pakistan. These tweets have a correlation with the political results of Pakistan, making it imperative to discuss and analyse the underlying impact of showing certain emotion towards a political party. The data collection was performed through tools as Weka to extract Roman Urdu tweets prior to 2018 general elections in Pakistan.

1.5 Significance

The focus of this research is on sentiment analysis where we will find the polarity based sentiment orientation, using targeting general elections as a domain data. Sentiment analysis is important due to its scalability, real-time analysis platform, and consistent criteria. This technique is considered imperative in natural language processing (NLP) that develops an automatic tool to categorize the positive, neutral or negative opinions to ensure effective and timely decision-making process. Through effective technique, as sentiment analysis of the 2018 Election tweets in Roman Urdu will assist in assessing the underlying correlation of people sentiment on social media and the election results.

1.6 Structure

Firstly, the research paper expands on the previous research conducted in analysis of social media sentiments of people and its impact on election results. However, the research gaps will be identified to analyze the aspects not incorporated in the previous studies. Next, developed methodology to extract and analyze the data will be presented, followed by the discussion on the results generated. Finally, the conclusion based on data analysis results will be presented.

2 Literature Review

Alaoul et al., (2018) [27] presented a simple method of polarity based classification of the textual data. A method called novel adaptable approach is presented which is used to calculate the result of general election using the twitter tweets. The collecting of tweets to construct the dictionary and further classify them into positive and negative classes. The collection of words is classified and preprocessing method tokenization, stemming and filtering step work in tree tagger tool. The proposed method implemented for the micro-blogging site of election predication. The previous work of the election predication was collected from micro-blogging site and extracted lexicon-pattern most of researcher express their predication about upcoming election. The proposed technique is based Natural Language Processing (NLP) to explore the whole concept of sentiment analysis to generate opinion, where precision, recall, and F-score is used for calculating performance. Their approach was compared with the two classifiers: Naive Bayes and Google cloud predication API. The proposed method achieves good accuracy of 90.21, and 89.98% as compared to Naïve Bayes and Google predication API.

Asghar et al., (2015) [16] developed a simple information theory concept the proposed procedure enhanced feature weight scheme they demonstrated the problem lexicon dependent on domain and their reviews modified domain independent which was generated and labeled. The three benchmark datasets car, drug-reviews and hotel, used in the proposed method achieved high accuracy based on classification polarity. Further, the polarity of terms changes using domain-dependent lexicon in the proposed method. The proposed method performed well and produced higher performance experimental results as compared with proposed approach. The whole concept changes the polarity words and enhances the accuracy of classification. The comparison between the experimental results of remaining lexicon based and proposed method in term of classification polarity and their achieved high accuracies.

Charlton et al., (2015) [17] recommended the sentiment analysis concept to association sentiment points of twitter user and that user @-

describing each other to the concerned network configuration. They collected a huge amount of tweet dataset and categorized into three algorithms. Firstly, they identified that user have huge communication approach to sentiment use in a different formed and comparison to positive sentiment frequently to the negative sentiment frequently average users. In the second they get them after every months finding the stable twitter communities, they develop their stable sentiment level, and immediate variation in sentiment day by day can be affecting most of the cases can be traced outer event. Thirdly, they recommended and generate standardized one of the most used simple agent based model, which can calculate the feedback as the result as good as by other empirical dataset.

Amjad et al., (2017) [18] focused on Urdu language for the sentiment analysis. They build the polarity-based lexicon extracted from dataset, the dataset is collected by core Urdu news of Pakistan over the period of 10 months. They demonstrated the classification algorithm to classify the reviews of Urdu sentiment analysis polarity based in positive, negative and neutral classes based on textual data on the sentiment-score of the Urdu language. They applied their algorithm to classify sentiment-score using three steps i.e. filtration, segmentation, and calculation. The proposed system in the first phase breaks the given sentence into tokens. The second is segmentation phase; the white space area in each sentence is changed into uni-gram and used as delimiter area. In the next stage, they filter the sentiment lexicon tokens and uni-gram to extract the opinion word of the sentence. In the final stage, they compute their sentiment score of sentence and calculate each opinion word in polarity based. The experimental result of the proposed system proved that the algorithm achieved 71% accuracy.

Korovkinas (2017) [19] proposed the hybrid-method to recover SVM classify accuracy using hyperactive parameter tuning and trained dataset. The hybrid method is implemented in clustering to select parameter and training data to improve the classifier efficiency. The main goal and comparative study of this method, and advantage with a fore mentioned method, is the training data. In this paper, the selection of dataset randomly from Subset30K, in multiple runs, this might negatively affect the accuracy, result. In proposed method, they used clustering-based instance method to highlighting the data points with MAX, MIN and AVG. In these points shows the distances to each cluster center. In this paper, they adopted sentiment analysis machine based technique, which is called SVM (Support Vector Machine) increase the accuracy result.

Hailong et al., (2014) presented comparative study about one of the most important common topic in sentiment analysis. In this survey paper, comparison between lexicons based and machine learning methods were discussed. The two approaches are used in this paper discovered cross-domain and lingual. The performance of proposed method proved that the supervised-machine learning approach gives a good accuracy and experiential classification. The proposed method is used studded deep learning approach to solving the problem.

Ruz et al., (2020) Concept of sentiment analysis is one of the common and famous topic in now a day. In this paper they presented machine language for twitter data. The twitter dataset divided into two natural disasters one is Chilean earthquake in Spanish 2010 and second Catalan independence referendum in 2017. In this paper they apply sentiment analysis common classifiers Bayesian to check the result effectiveness. In this performance to check the all-competitive result to Training dataset. The experimental result of the proposed system proved that the algorithm achieved 80% accuracy.

Asghar et al., (2014) [21] focus one of the common and famous research topic now a day is sentiment analysis. The researcher adopts a lot of research topic in many different languages but one of important language they adopt is roman language i.e. Pashto roman and Urdu roman. About 30 million people across the world share their feelings, reviews and experiences over the internet used in roman Urdu language. This growing field has major opportunities to determine the people feelings or behavior, emotions, and attitude. The User analyzed to generated content can be useful to product acceptance, popularity measuring, summarization of reviews, and predictive analysis etc. In this paper, researcher focuses on the proposed system on lexicon-based approach. Using Sentiment analysis approach, i.e. Lexicon based approach has been proposed that is performed of fluent data (English and Roman Urdu). The proposed system analysis and collect the tweets data related to election 2018 held in Pakistan to sentiments expressed in them.

Ahmad et al., (2017) [22] proposed a treebank manufacture of a multilayered phrase structure design. There are three layers of treebank in this paper. The grammatical function, semantic roles and phrases. In the primary level phrase consist of 12 tags. Each phrase label following some grammatically function which is inspired by lexicon functional grammar. These phrases monitor same semantic role. Semantic role label already using prop bank roles. The treebank guideline1 using the CLE Urdu Digest Corpus 1,300 sentence.

Ruz et al., (2020) [24] further researchs in current era topics as cyber-crime is an unethical behavior fraud or illegal gambling exposed. In modern society reflex detection, a language on social platforms is a major challenging field. The researcher deal with difficulty of natural language While, now they focused on common languages i.e. English. Roman Urdu and Urdu language resource-rich and two script of writing language on social networking sides. The researcher use Urdu writing in Urdu characters whereas, the Roman writing uses the English language characters. The two spoken natural languages are extremely similar to one another i.e. Urdu and Hindi language but the writing scripts are completely different with each other. The focus of researcher to detection the user’s comments accessible in an Urdu language. The researcher proposes to collect user-generated comments dataset from social media of Urdu offensive language.

In the research gap, it is presented that in sentiment analysis extensive studies are conducted previously. In this field motivating method and many developed scheme, are handle by lot of researcher to grip the research problem. In sentiment analysis, the most of methods are proposed and development for studied where been taken in other language such as English language etc. and existing method cannot work accurately in other language like Arabic, Punjabi, Pashto, and Roman Urdu. Therefore, it is very useful for researcher to propose a method and develop the resource for these languages. We study a lot of recent and up to date work in different language on sentiment analysis we point out the useful method for the roman Urdu language sentiment analysis polarity based.

3 Methodology

A sentiment analysis approach is proposed, where we apply adoptable lexicon-based approach. Lexicon based-approach is un-supervised method. The proposed method consists of followings steps, which are shown in Figure 1.

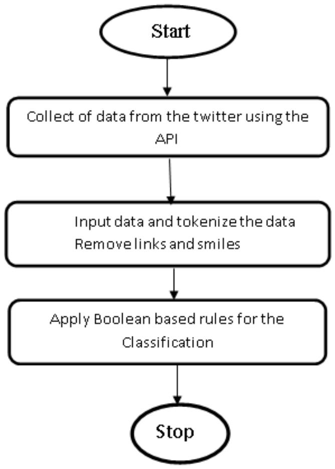
Start
Collect of data from the twitter using the API
Input data and tokenize the data
Remove links and smiles
Apply Boolean based rules for the Classification
Stop

Fig. 1 Proposed methodology for the sentiment analysis conducted on Roman Urdu Tweets.

The data is collected from twitter. Then we will apply pre-processing technique and rule based methodology to select words, based on sentiment orientation categories that strongly reflect code-mixing behavior and will construct the dictionary [7]. In interpretation phase polarity, score of the tweeted word or sentence, will be checked to get the required positive, negative and neutral words. In the proposed method we apply the Boolean based rule methodology which is suitable for best accuracy. The methodology called SRule mostly used for text mining and is a standard choice for sentiment analysis. We will build a prototype by analyzing the three political parties related twitter tweets to show that which party is the favorite in order to identify strong candidates.

The proposed system is summarized into three stages. Figure 3showsthe first stage. We will construct the dictionary for Urdu roman sentiment words which are divided into positive, negative and neutral words. In the next step, we will classify the dataset and identify the opinion words using the Lexicon-based approach and predicate data.

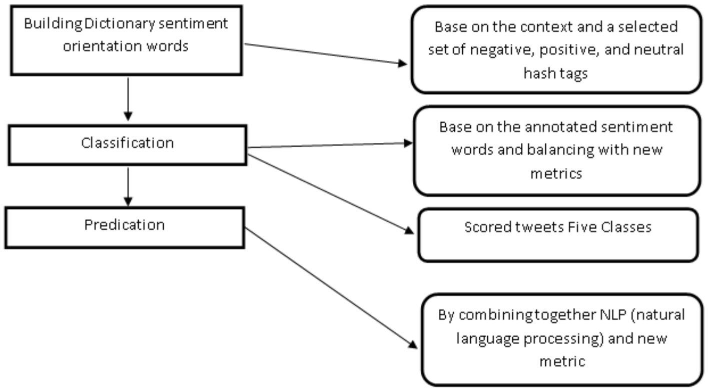
Building Dictionary sentiment orientation words
Base on the context and a selected set of negative, positive, and neutral hash tags
Classification
Base on the annotated sentiment words and balancing with new metrics
Predication
Scored tweets Five Classes
By combining together NLP (natural language processing) and new metric

Fig. 2 Application of Lexicon-based approach on the dataset.

The research design is completed through several stages.

3.1 Dataset Creation

There are four different ways of getting twitter data., Firstly, to retrieve data from twitter public API, secondly to find an existing twitter dataset, third to purchase data from twitter and fourthly access or purchase from a Twitter service provider. We will collect data from google API of Urdu roman words or tweets from micro-blogging side twitter for further phases. Using these datasets in our proposed method the first phase will be implemented which is called pre-processing phase.

3.2 Pre-processing Phase

In the first phase, we will build a dictionary that remove duplicates and rearrange the textual data using text mining techniques i.e. conversion, tokenization, filtering, morphology, and removing duplicated data. The concept of conversion is to eliminate or remove the repetition of words. These extended words repeated some letters more than two successive letters for example pakistanaaan by Pakistan GOOOOD by good and removing the Uppercase letters into lowercase letter e.g. HAPPY by happy. Tokenization is a process that segments or fragments the words. These segments mean phrases and whole sentence called tokens. Therefore, the tokenization is used for discarding some line breaks or punctuation marks such as full stop, comma, blank space or question mark. This is the best indicator for sentiment analysis. It filters empty words, plurals, conjugation and for representing adjective and verbs. By filters we easily get proper sentiment word i.e. positive, negative and neutral words. Morphology is the study of words or structure of words. It stems the plural gender, root words, prefixes, suffixes and conjugation. In this phase, we will be eliminating frequent or duplicate word data. It removes duplicate tweets to avoid misleading results e.g. #Good#Good by #Good.

3.3 Emotion-classification: Lexicon-based

After the pre-processing steps we will get a lot of data in the form of casual words, emoticon/emoji, short form, abbreviation, spelling mistake and single English words [16]. According to our prospective, we will only collect data in roman Urdu. For this type of data, we will use Google translator to convert English words into roman Urdu words. We can clean this type of noise data to perform and analyze the effect of emoticons.

3.4 Interpretation Phase

After creating the lexicon-based dictionary in our proposed method, we will assign that polarity score to all lexicon words e.g. +1 for positive, -1 for negative and 0 for neutral words. In the next step, we will apply algorithm to get enhance and find positive sentiment words dictionary, negative sentiment words dictionary and neutral sentiment words dictionary. After these steps, we will apply two actions for polarity degree of scoring of lexicon-based word.

Balancing is a simple concept of an extended word to reduce the polarity degree of the post. In our proposed method, this kind of words needs full attention. In our proposed method, for balancing we will use scoring metric in which words of polarity will be 0 (zero) if there doesn’t exist any word, +1 for positive words and -1 for negative words. For checking the overall concept of polarity, we will use the polarity formula, which collects polarity score and polarity degree:

Polarity (t) = $\sum_{k=1}^{n} wt$ (1)

Given:

t = tweet

n = length of tweet (or number of words in a tweet or sentence)

w = word

After creating the lexicon-based approach we can further classify sentiment orientation words (positive, negative and neutral). We use word cloud technique to highlight most common positive, negative and neutral word used in this dictionary.

In these second phase the feature identification phase, we identify the orientation sentiment or opinion and compare these words for checking positive, negative and neutral words.

The following Boolean rules are given below to identify orientation sentiment of words:

Rule 1: The concept of logical AND, when there are two positive and one negative occur then the polarity of orientation sentiment word is negative.

Rule 2: The concept of logical OR, when there are two negative and one positive occur then the result will be positive.

Rule 3: The concept of logical NAND, when there are three positive occur then the result will be positive.

Rule 4: The concept of logical XNOR, when there is one negative and two are positive the result will be negative.

The researchers have mostly used following performance evaluations to find the classification performance i.e. Precision, Recall and F-measure [14].

Firstly, precision can be calculated as:

Precision is the ratio of correctly called Positive predicated value. High precision relates to the low false positive rate. Find mathematically we can use precision formula [15].

Precision = TP/TP+FP (2)

Next, Recall is also known as, sensitivity. Recall is the ratio of correctly predicted positive observations in actual class. Mathematically, it can be written as, the following, [15].

Recall = TP/TP+FN (3)

The F-Measure is also known as, F-Score. F-Measure is the partisan average of precision and recall. For calculating the F-Measure, we use the following equation [15].

F-Measure = 2*(Recall * Precision) / (Recall + Precision) (4)

The methodology applied in the research can be summarized as shown in figure 3.

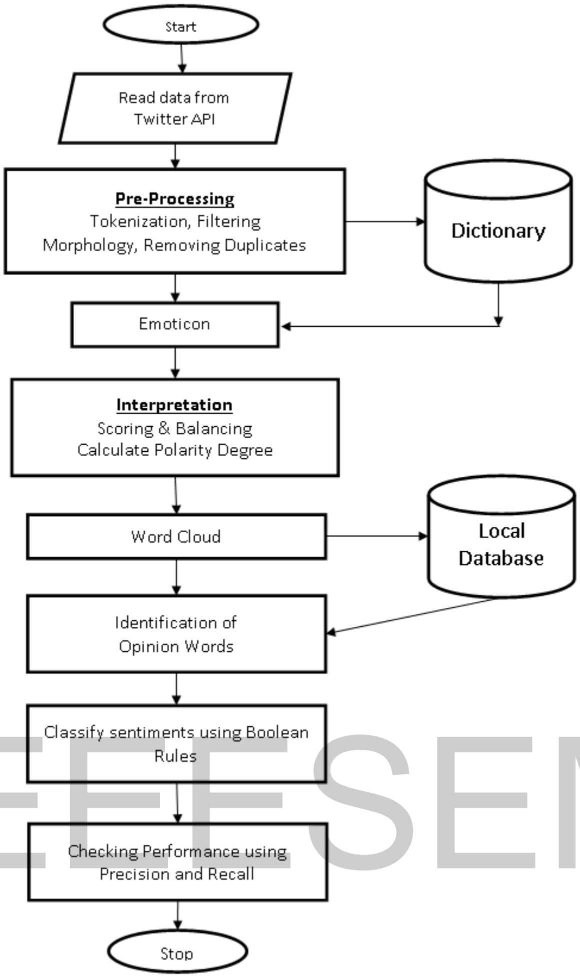
Start
Read data from Twitter API
Pre-Processing
Tokenization, Filtering
Morphology, Removing Duplicates
Dictionary
Emoticon
Interpretation
Scoring & Balancing
Calculate Polarity Degree
Word Cloud
Local Database
Identification of Opinion Words
Classify sentiments using Boolean Rules
Checking Performance using Precision and Recall
Stop

Fig 3. The research methodology used in the sentimental analysis.

4 Results and Discussion

In this setion, the results and discussion of the sentimental analysis conducted is presented to study the objectives and the results generated. The Python language will be used for system implementation. Python is a general-purpose language, which is used mostly now a day in the text mining and machine learning. It is open source, easy to learn and it provided efficient machine learning requirements like processing big data, mathematical calculation, NLTK etc. NLTK (Natural Language Toolkit) is the popular library and one of the most famous packages in python. It works with human languages and provides user-friendly interfaces to lexical resources such as tokenization, classification, stemming, filtering etc. Other than that Weka tool will be used as a framework for classification, which is popular now a day for data or text mining. The requirement of experimental classification our proposed system Boolean-rule based mechanism is easily handles specific type of dataset.

4.1 Sentimental-Analysis Dataset

To conduct the analysis, in the first stage, the tweets were extracted prior to the 2018 Election day in Pakistan. However, a filter was used to only select the tweets in Roman Urdu Language, which reduced the dataset to only keep the Roman Urdu Tweets. Next, a dicitionary is formed that consists of phrases as “fouj, drama, siyasi, and laanat” to filter the Roman Urdu tweets. Basically, the purpose is to only analyze the tweets that show either a positive, negative or neutral reaction or sentiments towards one of the three parties in the elections. Moreover, these sentiments can be focused on external sources as other countries or establishment in Pakistan, which people think might facilitate the winning of any party.

Once the dataset was filtered for lexicon-based or boolean approach, the results were presented in a confusion matrix.

4.2 Boolean-based Dataset

In the research, three main political parties were targeted: PTI, PPPP, and PMLN that formed the largest three parties in 2018 elections. After the application of filters to only analyze the sentimental tweets regarding the political parties, the dataset consisted of 312 Tweets in Roman Urdu. The next stage was to analyze the tweets through the sentimental analysis that predicted if the tweet was focused towards PTI, PPPP, and PMLN.

The tweets were analyzed through boolean-based approach that utilizes algorithms to analyse, which party was most focused in the elections. In the confusion matrix, the results generated after the Boolean-based approach are summarized. The political parties were presented through the following approach: A represented PTI, B was an anocrym for PMLN and C was abbreviated for PPP.

Through the confusion matrix, it can be analyzed that the highest number of tweets were redirected towards A (PTI). In total, these tweets were calculated through Boolean-based results as 121. It was followed by the tweets for B (PMLN) that were the most strong competitors of A and they had the total number of tweets as 106. Lastly, the tweets were for C (PPPP) that didn't have high or strong competition from either of the parties.

Moreover, it was analyzed through the confusion matrix that highest number of tweets were singularly for the parties as only A, B, and C. There were no tweets against A that was in favor of B or C. Similarly, no tweets were presented against B that supported C; however, 5 tweets supported the political party A. In contrast, there were tweets against C tha favored both A (1 tweet) and B (3 tweets).

Through the confusion matrix, it was generated through the prediced labels that F1 was calculated as 97.13 (macro) and 97.12 (micro). Next, P was calculated as 97.51 (macro) and 97.12 micro. Lastly, we calculated R that had a value of 96.83 (macro) and 97.12 (micro). The summation of all the factors led to a total accuracy of 97.12 in the selection and analysis of tweets based on the boolean-based data.

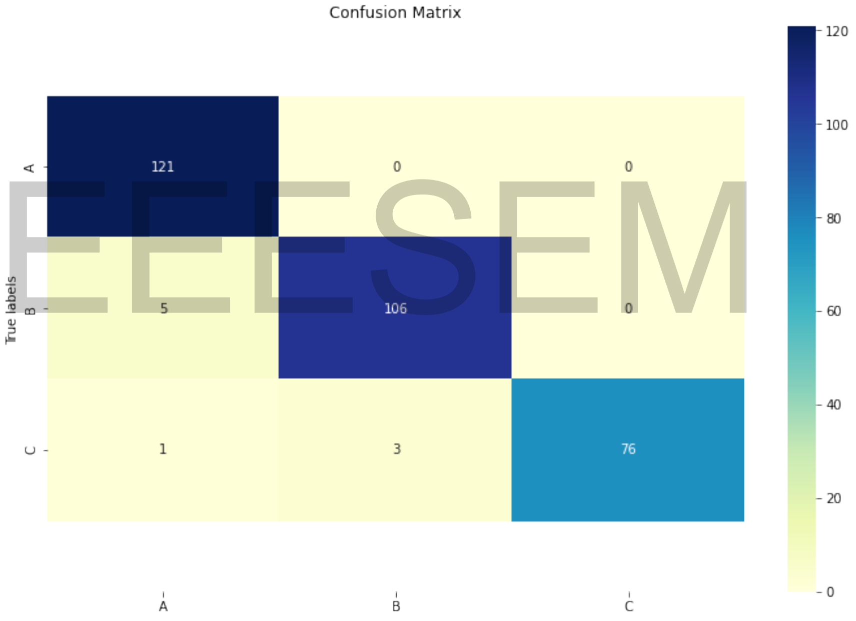

Fig. 4 Confusion Matrix for Boolean-based Dataset

4.3 Results

Based on the results, it was predicted that highest number of tweets were for the Part A or PTI and based on the objectives and research study, it should win the elections. Based on the results of 2018 elections in Pakistan, the winning party was PTI that showed a positive correlation between the highest number of tweets and sentiment of people and the winning of the political party.

5 Conclusion

This paper focuses on using social media tweets for the forecast of election results. Over the years, sentiment analysis is also classified as opinion mining that determines and extracts opinion (positive, negative, and neutral) and information about a related topic. Moreover, researches have the opportunity to select between three sentiment analysis approaches that are lexicon-based approach and machine learning approach.

In lexicon-based approach, the measure of polarization is the given content from the sentiment orientation words or phrase in documents. The aim of this approach is to identify the sentiment word or opinion expressed by user whether the words present in positive, negative or neutral. In contrast to the lexicon-based approach, machine learning can be unsupervised, semi-supervised, and supervised that demands training prior to data mining. In sentiment analysis, its can be performed through classification algorithms that are further segmented into linear classifiers, decision tree, and probabilistic classifier. There are many types of probabilistic classifiers in supervised machine learning approach. The approaches of supervised-machine learning give a good accuracy and experiential classification Naïve Bayes, type of method are less effective, as the machine learning approach method is limited struggle human labeled documents and quality and quantity of datasets. Through this study, a positive correlation between the highest number of positive tweets garnered by the political party and its winning in the election is found.

6 APPENDIX

6.1 Acknowledgements

I would like to thank our professor Dr. Abdus Salam who throughout the research paper assisted and supervised me.

6.2 References

[1] P. Keung, J. Salazar, Y. Lu and N. Smith, "Unsupervised Bitext Mining and Translation via Self-Trained Contextual Embeddings", *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 828-841, 2020. Available: 10.1162/tacl_a_00348.

[2] A. Devitt and K. Ahmad, "Is there a language of sentiment? An analysis of lexical resources for sentiment analysis", *Language Resources and Evaluation*, vol. 47, no. 2, pp. 475-511, 2013. Available: 10.1007/s10579-013-9223-6.

[3] M. Araújo, A. Pereira and F. Benevenuto, "A comparative study of machine translation for multilingual sentence-level sentiment analysis", *Information Sciences*, vol. 512, pp. 1078-1102, 2020. Available: 10.1016/j.ins.2019.10.031.

[4] D. Yuret and M. Yatbaz, "The Noisy Channel Model for Unsupervised Word Sense Disambiguation", *Computational Linguistics*, vol. 36, no. 1, pp. 111-127, 2010. Available: 10.1162/coli.2010.36.1.36103.

[5] J. Zhao, K. Liu and L. Xu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions", *Computational Linguistics*, vol. 42, no. 3, pp. 595-598, 2016. Available: 10.1162/coli_r_00259.

[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-Based Methods for Sentiment Analysis", *Computational Linguistics*, vol. 37, no. 2, pp. 267-307, 2011. Available: 10.1162/coli_a_00049.

[7] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis", *Computational Linguistics*, vol. 35, no. 3, pp. 399-433, 2009. Available: 10.1162/coli.08-012-r1-06-90.

[8] B. Klebanov, N. Madnani and J. Burstein, "Using Pivot-Based Paraphrasing and Sentiment Profiles to Improve a Subjectivity Lexicon for Essay Data", *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 99-110, 2013. Available: 10.1162/tacl_a_00213.

[9] F. Bravo-Marquez, M. Mendoza and B. Poblete, "Combining strengths, emotions and polarities for boosting Twitter sentiment analysis", *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '13*, 2013. Available: 10.1145/2502069.2502071 [Accessed 30 April 2021].

[10] A.Bermingham, A.Smeaton" On using Twitter to monitor political sentiment and predict election results",CLARITY: Centre for Sensor Web Technologies,pp.1-10, Nov.23. 2011.

[11] R.Bhatt, V.Chaoji , R. Parekh ,"Predicting product adoption in large-scale social networks". In: Proceedings of the 19th ACM international conference on Information and knowledge management. New York: ACM; 2010. p. 1039–48.

[12] S. Jain, N. K.Sharma, and S. Gupta, "Business strategy predication system for market basket analysis," Journal of Big data, pp. 93–106, Oct. 2017.

[13] M. Herland, T. M. Khoshgoftaar, and R. A. Bauder, "Big Data fraud detection using multiple medicare data sources," Journal of Big Data, pp. 1–21, 2018.

[14] S.Mukund, R.K. Srihari "Analyzing Urdu Social Media for Sentiments using Transfer Learning with Controlled Translations",Association for Computational Linguistics (LSM 2012), pp. 1–8, Montreal, Canada, June 7, 2012.

[15] W. Medhat, A.hassan, H. korashy, "Sentiment analysis algorithm and application: A survey" Ain Shams engineering journal, Vol. 5, pp. 1093-1113, 2014.

[16] Asghar, M., A., Khan, S. Ahmad, I. A. Khan and F. M. Kundi,"A unified framework for creating domain dependent polarity lexicon from user generated reviews". PloS one, 10(10), e0140204, 2015

[17] Charlton, N., C., Singleton and D. V. Greetham,"In the mood: the dynamics of collective sentiments on twitter". Royal society open science, 3(6), 2016.

[18] Amjad, K., M. Ishtiaq, S. Firdous and M.A. Mehmood,"Exploring Twitter news basis using urdu-based sentiment lexicon". In open source systems & Technologies (ICOSST), pp. 48-53, 2017

[19] K.korovkinas, "SVM and k-Means Hybrid Method for Textual Data Sentiment Analysis" Baltic J. Modern Computing, Vol. 7 , No. 1, pp. 47–60, January 2019 (pp 47-60).

[20] Hailong, Z., G. Wenyan and J. Bo. 2014 "Machine learning and lexicon based methods for sentiment classification": A survey. In Web Information system and application conference (WISA), 2014 11th (pp.262-265), IEEE.

[21] Asghar, M. Z., A. Khan, and S. Ahmad."Lexicon-Based Sentiment Analysis in the social web" Journal of Basic and Applied Scientific Research 2014 (pp.238-248).

[22] M. Ahmad, S. Aftab and I. Ali, "Sentiment analysis of tweets using SVM" International Journal of Computer Applications (0975 – 8887) Volume 177 – No.5, November 2017 (pp 25-29).

[23] A. Tupsoundarya, S. Dandannavar"Sentiment Expression via Emoticons on Social Media: Twitter"Volume 6 Issue VI, June 2018, International Journal for Research in Applied Science & Engineering Technology (IJRASET)

[24] G. A.Ruz, P. A. Henriquez, A.Mascareno “Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers” Journal Vol:106, 13 January 2020.

[25] P. Chen, Z. Sun, L. Bing and W. Yang, "Recurrent Attention Network on Memory for Aspect Sentiment Analysis", *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017. Available: 10.18653/v1/d17-1047 [Accessed 30 April 2021].

[26] D. Ruck, N. Rice, J. Borycz and R. Bentley, "Internet Research Agency Twitter activity predicted 2016 U.S. election polls", *First Monday*, 2019. Available: 10.5210/fm.v24i7.10107.

[27] I. E. Alaoul,Y. Gahi, and R. Messoussl, "A novel adoptable approach for sentiment analysis on big data," Journal of Big data , pp. 1–18, 2018.

[28] M. De. Choudhury, M.Gamon, “Predicting depression via social media”, Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, pp.128-137, 2013.

[29] G. Brown, J. Wyatt, R. Harris and X. Yao, "Diversity creation methods: a survey and categorisation", *Information Fusion*, vol. 6, no. 1, pp. 5-20, 2005. Available: 10.1016/j.inffus.2004.04.004.

[30] J. Devlin, M. Chang, K. Lee and K. Toutanova, *Proceedings of the 2019 Conference of the North*, 2019. Available: 10.18653/v1/n19-1423 [Accessed 30 April 2021].

[31] D. Farias and P. Rosso, "Irony, Sarcasm, and Sentiment Analysis", *Sentiment Analysis in Social Networks*, pp. 113-128, 2017. Available: 10.1016/b978-0-12-804412-4.00007-3 [Accessed 30 April 2021].

[32] A. Ghosh et al., "SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter", *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015. Available: 10.18653/v1/s15-2080 [Accessed 30 April 2021].

[33] "Preprint repository arXiv achieves milestone million uploads", *Physics Today*, 2014. Available: 10.1063/pt.5.028530.

[34] S. Rani, "Hybrid Model using Stack-Based Ensemble Classifier and Dictionary Classifier to Improve Classification Accuracy of Twitter Sentiment Analysis", *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 7, pp. 2893-2900, 2020. Available: 10.30534/ijeter/2020/02872020.

[35] P. Nakov et al., "Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts", *Language Resources and Evaluation*, vol. 50, no. 1, pp. 35-65, 2016. Available: 10.1007/s10579-015-9328-1.

[36] K. Garcia and L. Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA", *Applied Soft Computing*, vol. 101, p. 107057, 2021. Available: 10.1016/j.asoc.2020.107057.

[37] J. Lochter, R. Zanetti, D. Reller and T. Almeida, "Short text opinion detection using ensemble of classifiers and semantic indexing", *Expert Systems with Applications*, vol. 62, pp. 243-249, 2016. Available: 10.1016/j.eswa.2016.06.025.

[38] R. Silva, T. Alberto, T. Almeida and A. Yamakami, "Towards filtering undesired short text messages using an online learning approach with semantic indexing", *Expert Systems with Applications*, vol. 83, pp. 314-325, 2017. Available: 10.1016/j.eswa.2017.04.055.

[39] "Trump and Muslims During US Presidential Elections 2016: A Sentiment Analysis of Muslim Community on Twitter", *Media Education (Mediaobrazovanie)*, vol. 60, no. 2, 2020. Available: 10.13187/me.2020.2.30.