IRSTI 11.15.23

[1]Moldir Tlepova
[1]Suleyman Demirel University, Kaskelen, Kazakhstan

SENTIMENT ANALYSIS ON TWEETS ABOUT THE ELECTION CANDIDATES ISING TEXTBLOB

Abstract. Sentiment analysis is the categorization of the speaker's, writer's, or other subject's perspective on a certain issue. Since Kazakhstan's presidential election in 2022 is one of the most discussed events we distinguish people's perspectives on the election and each candidate. The possible leader of the country will be influenced by the public's perception of a candidate. A diverse data set illustrating the current public perceptions of the candidates are gathered from the Twitter platform. The collected tweets are examined using a lexicon- based methodology TextBlob to ascertain the public's sentiments. In this study, we analyze the collected tweets to identify the polarity and subjectivity measures that provide light on the user perception of a certain candidate.

Keywords: Sensual analysis, presidential election, TextBlob.

***

Аңдатпа. Мәтіндердің сезімдік талдауы – белгілі бір мәселе бойынша сөйлеушінің, жазушының немесе басқа субъектінің көзқарастарын санаттау. 2022 жылғы Қазақстандағы президенттік сайлау ең көп талқыланатын оқиғалардың бірі болғандықтан, халықтың сайлауға және әрбір кандидатқа деген көзқарасын анықтағымыз келді. Твиттер әлеуметтік платформасынан үміткерлер туралы қазіргі қоғамдық пікірді бейнелейтін әртүрлі деректер жинағы жиналды. Жиналған твиттер қоғамдық көңіл-күйді анықтау үшін лексикалық негізделген TextBlob әдістемесі арқылы тексерілді. Бұл мақалада біз твиттерді пайдаланушының белгілі бір кандидатты қалай қабылдайтынын және полярлық, субъективтілік өлшемдерін анықтау үшін талдаймыз.

Түйін сөздер. Сезімдік талдау, президенттік сайлау, TextBlob.

***

Аннотация. Анализ тональности текстов - это категоризация взглядов говорящего, пишущего автора или другого субъекта на определенный вопрос. Поскольку президентские вы- боры в Казахстане в 2022 году являются одним из самых обсуждаемых событий, мы хотим определить взгляды людей на выборы и на каждого кандидата.

Разнообразный набор данных, иллюстрирующий текущее общественное мнение о кандидатах, соби- рается с социальной платформы Twitter. Собранные твиты проверяются с помощью основанной на лексической методологии TextBlob для определения настроения публи- ки.В этом исследовании мы анализируем собранные твиты, чтобы определить меры полярности и субъективности, которые проливают свет на восприятие пользователем определенного кандидата.

Ключевые слова: Чувственный анализ, президентские выборы, TextBlob.

***

Introduction

In the modern world, dominated by the power of social networks and the opinion of the majority, determining the mood and sentiment of the individual on a specific topic and, therefore, the different attitude of the whole group of individuals plays a significant role. With the advent of the Internet, every person can be an author, a writer, or speaker in the media space. Each tweet on Twitter or review left in online stores has its impact. We are used to sharing our experiences or opinion on some product, a movie, or political events. In addition, we form our opinion based on the information on the Internet. Thus, people's beliefs, sentiments, and emotions on the Web influence business, entertainment, politics, and society. Every text on the Web is a digital data, and organizations, and scientists use and analyze these data for their purposes.

Sentiment Analysis(SA) is one of the studies of Natural Language Processing that classifies text based on sentiment orientation. It derives information from social blogs, and discussion forums and then distinguishes them by their opinion polarity, which is positive, negative, and neutral. Sentiment Analysis has many difficulties and challenges in determin- ing the accurate polarity of a text because of some NLP problems. Human language has many intricacies and underlying meanings that the computer cannot handle.

Literature Review

The language of communication of people on Internet platforms is rapidly changing, consequently, Sentiment Analysis needs updates to carry out novelties and additions. SA has different methods and approaches to analyzing huge data. And the accuracy of these methods varies.

From the beginning of the 21st-century numerous research papers and results have been made and new approaches have been introduced.

Bo Pang and Lillian Lee in 2002 collected movie reviews and using machine learn- ing methods classified them into positive and negative

categories. The authors used three different methods: Naive Bayes, maximum entropy classification, and support vector ma- chines. Also, they concluded that sentiment-based analysis is more complex and more challenging[1]. Later they published overall research about opinion mining and sentiment analysis and discussed general trends and interests in this area[2].

Besides machine learning methods there are unsupervised approaches that take ex- actly the sentiment of words and lexicon. Peter D. Turney presents a classification of reviews by the semantic orientation of phrases, use of adjectives and adverbs. He builds the relation- ship between the given phase and the words "excellent" and "poor" and gives them a score of the particular sentiment[3]. Hu and Liu use almost the same method as Peter Turney on analyzing customers' reviews about some product, but presents some novelties[4]. Maite Taboada et al. use different approach Lexicon-Based Methods. They create dictionaries of words with semantic orientation and with their use assign polarities to texts. In addition, they show how to establish proper and reliable dictionaries[5].

Liu and Zhang presented an article that gives a detailed view of every aspect of se- mantic analysis and discusses some problems of distinguishing accurate orientation because of some nuances of human language and spotting spams[6].

There are some articles presenting the application of the SA in different areas on Twitter. Farha Nausheen et al. in their research with the help of a twython library col- lected tweets during the election in the USA and they filtered these tweets using NTLK (Natural Language Toolkit). By application of opinion mining the authors differed tweets of each candidate[7]. During impactful world events, it is very important to take into ac- count the opinion of the public. And one of those events was the spread of the coronavirus. K.Manguri et al. analyzed in [8] the latest people's opinions regarding COVID-19 on Twit- ter. They use a similar method as in the previous article. B. Gupta, M.Negi tested all machine learning algorithms and compare their accuracy using Twitter data. They applied different training models by using Natural Language Processing (NLTK), SCIKIT-LEARN libraries: Classification, Regression, Clustering. The authors presented some difficulties in classification and how to deal with them[9].

The papers above are based on English's grammar and semantic rules. So basically, there are many techniques, methodologies, and libraries of Semantic analysis for English texts. But for other languages, the libraries don't work as they do for English. Kotelnikov E. and Klekovkina M. discuss machine learning methods of Semantic analysis for Russian texts[10]. Osokin V.and Shegay M. use the Naive Bayes classifier for defining semantics of Russian text. They showcased an excellent accuracy in distinguishing polarity[11].

Methods and Materials

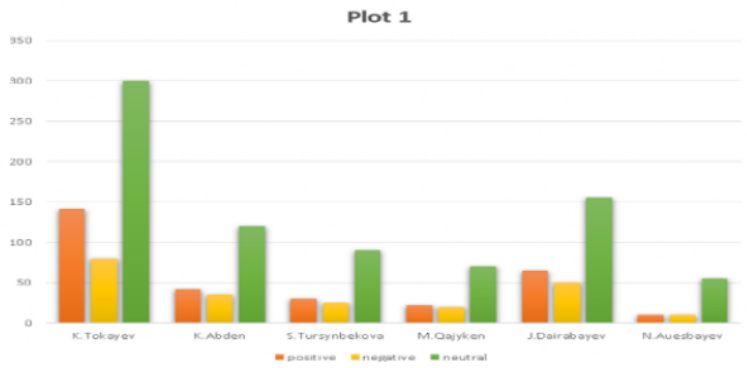
The first task in this project is to collect the necessary information from Twitter. In our case, sentiment analysis analyzes tweets about candidates for elections in Kazakhstan, which took place in November 2022. We use a social networking service SNScrape to collect data from Twitter. This service scrapes data like usernames, hashtags, or searches and returns the discovered items by some keywords. The period was chosen from the announcement of the elections on 1 September to the end of the election on 30 November. Moreover, we apply for the Twitter filter such keywords as "выборы Казахстан" and names of candidates. Thus, we collected more than 1500 tweets in Russian.

The next step is preparing the tweets in a suitable format. Empty rows and duplicate texts have been removed, as well as tweets containing only photos or videos. All texts in Kazakh and Russian were translated into English. The translation was made because in the next step when using the analysis, there are no good and free libraries for determining the polarity of the text. English has advanced much more than other languages.

For the third part, we use the TextBlob library to determine the sentiment of each tweet. This library uses NLTK algorithms and take the color of each word and gives a summed result for each text. The sentiment of the text ranges from -1 to 1.

Data and Results

As we said above the algorithm of TextBlob is used to the filtered tweets in order to calculate the total number of negative, positive and neutral tweets for Kassym-Jomart Tokayev, Meiram Qajyken, Jiguli Dairabaev, Qaraqat Äbden, Saltanat Tursynbekova and Nurlan Auesbaev. Depending on the polarity of the specific words used in the tweet, the sentiment of the text is categorized. A positive term (such as good, great, etc.) receives a score of 1, a negative word (such as bad, worse, pathetic etc.), and a neutral word (such as quite, average, etc.) receives a score of 0. Polarity values vary from $[-1, +1]$. The ratings of all the individual words are then added to determine the overall polarity of a tweet. The results of the analysis is presented in Plot 1.



Discussion and Conclusion

According to the graph, we can see that Kassym-Jomart Tokayev has more positive tweets than the other candidates, but the number of negative tweets is also higher than the rest. Because he has more attention on the social platform. The difference between negative and positive tweets is quite the same for other candidates. Therefore, we can predict that Kassym-Jomart Tokayev has a higher chance of winning the president election.

Approximately the same results were obtained in Farha Nausheen's article about president elections in the USA [7]. In their study, they suggest a lexicon-based sentiment analyzer that categorizes tweets according to their sentiment value. The differences between this paper and their work lies in the way of collecting and processing data, and applying some other NLTK algorithm. Due to the wide range of English speakers and twitter users, their derived data is much bigger than the collection of tweets that we use in this project. Also the accuracy of their analysis is higher because of the direct use of language. in our case we need to translate the original tweet to English in order to apply the NLTK methods.

To obtain a more accurate result in the future, you can use NLTK methods directly for the texts without translation, as the SA for Russian language is still developing.

References

1 Pang, B., Lee, L. and Vaithyanathan, S., 2002. Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070.*

2 Pang, B. and Lee, L., 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2), pp.1-135.

3 Turney, P.D., 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032.*

4 Hu, M. and Liu, B., 2004, August. Mining and summarizing customer reviews. *In Pro- ceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

5 Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2), pp.267-307.

6 Zhang, L. and Liu, B., 2017. Sentiment Analysis and Opinion Mining.

7 Nausheen, F. and Begum, S.H., 2018, January. Sentiment analysis to predict election results using Python. *In 2018 2nd international conference on inventive systems and control (ICISC)* (pp. 1259-1262). IEEE.

8 Manguri, K.H., Ramadhan, R.N. and Amin, P.R.M., 2020. Twitter sentiment analysis on worldwide COVID-19 outbreaks. *Kurdistan Journal of Applied Research, pp.54-65.*

9 Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., Badhani, P. and Tech, B., 2017. Study of Twitter sentiment analysis using machine learning algorithms on Python. *International Journal of Computer Applications*, 165(9), pp.29-34.

10 Котельников, Е.В. and Клековкина, М.В., 2012. Автоматический анализ тонально- сти текстов на основе методов машинного обучения. *Компьютерная лингвистика и интеллектуальные технологии*, 2(11), p.27.

11 Осокин, В.В. and Шегай, М.В., 2014. Анализ тональности русскоязычного текста. Интеллектуальные системы. *Теория и приложения*, 18(3), pp.163-174.