Comparative Study of English and Russian Language Tweets in Ukraine-Russian Special Military Operation Dataset

Rahul Kevadia[1], Amriteshwar Singh[2], Kaining Huang[3]

***Abstract*—Sentiment analysis is a kind of analysis that helps to monitor emotions on Social Media, Brand Monitoring, Product Analysis, Market Research, and many more. Sentiment analysis helps to understand how people feel. Natural Language Processing is used to train the machine to better understand how the people feel about the special military operation. Social media has become a significant tool for disseminating information and influencing opinions and decisions. Furthermore, social media is having an increasing impact on political discourse. It allows institutions and citizens to connect directly with one another, allowing for more direct and active participation in political decision-making processes. As a result, sentiment and emotion analysis are at the heart of social media research, and it may be used as a significant content framing tool to boost virality. It is important to analyze the sentiment of the population on this matter as it can help to track and impact political opinions. In this comparative study, we reviewed the multiple paper about the sentiment analysis of Russian-language content and identified current challenges. In this paper, we presented methods to improve the quality of the applied sentiment analysis studies. We are trying to compare the sentiment of the Russian Language Tweets with the same tweets translated into English. Moreover, we have also performed sentiment analysis of Russian language tweets and tried multiple techniques to improve the quality of the sentiments lexicon in the Russian Language.**

***Index Terms*—Sentiment Analysis, Text Mining, Twitter Analysis Machine Learning, Deep Learning, Text Classification, Natural Language Processing.**

I. Introduction

Predicting, Inspecting, and Analysis of tweets on special military operations using computational power to gather information about a certain subject. With the use of Natural Language Processing and the concept of Machine Learning, we can train the computer to classify tweets from the dataset.

Our Survey differs from existing literature survey in that we focused on the application of sentiment analysis rather than existing sentiment analysis approaches and their classification quality. This paper is about the Sentiment analysis of tweets posted on one of the social media platforms called Twitter. BERT (Bidirectional Encoder Representations from Transformers) is the NLP (Natural Language Processing) Model, to predict the masked word from the tweets of the user for a special military operation. This Algorithm uses many NLP algorithms and architectures such as semi-supervised training, ULMFit, ELMo Embeddings, and OpenAI transformers. It is specially designed to use a masked language model, or MLM, to pre-train deep bidirectional representations from the unlabeled text. The processing of the data takes place and generally consists of two steps, pre-training and fine-tuning for creating models and for a wide range of tasks.

On 24 February 2022. The military conflict escalates between these two countries Russia and Ukraine. As soon as the operation starts, it attracts the attention of the public immediately and opinion towards this Special Military Operation keeps bubbling up, especially on social media. Having millions of users in multiple countries, Twitter is a social networking space where users from different places post and share their ideas. The opinion of Twitter users partly represents the public opinion towards this event. Therefore, the outcome of the research may reflect which country gains more support from the public. Considering the massive power of public opinion and its influence on culture, economy, and politics, we are interested in the attitude toward the war between Ukraine and Russia on Twitter users. Moreover, among the population who supports either side, the languages they use imply one's nationality, and thus we can understand the attitude of Twitter users of a specific country to some extent. By researching and analyzing the opinion of the public, we can understand which country the majority supports and what they think about this Special Military Operation.

II. Related Work

This section is to discuss the different Methodologies Review/ Literature Review and Motivation Outcomes from the given references. First of all, in the paper ”Multitask Learning for Fine-Grained Twitter Sentiment Analysis”, Georgios Balikas, Simon Moura, Massih-Reza Amini (Issue 2017) [1] proposed the approach of traditional sentiment analysis which solves problems such as 3-category called ternary and fined grained called 5-category classification by learning task individually. The author discusses that this classification is correlated with each other and finally they proposed a multitask learning approach. This study describes the potential of multitask model for a similar kind of problem and improved fine-grained Twitter sentiment classification problem. The challenges faced by the author to solve this issue are the common use of abbreviations, and creative languages used in tweets by the Modern generation. The author found data from two different Twitter sentiment classifications such as fine-grained and ternary, and they consider fine-grained to be the primary task as this task is challenging. According to the author, the solution to the given problem is to compare results with several models such as SVM, Logistic Regression, MaxEnt, and biLSTM to evaluate the multitask learning approach and they found that Support Vector Machine has the maximum margin classification algorithm that shows the competitive performance to solve several text classification problems.

Another paper ”Sentiment Analysis of Twitter Data for Predicting Stock Market Movements”, Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, Babita Majhi (Issue 2016) [2]proposed that how stock prices change of a company, such as high and low prices, how prices are correlated with the public opinions posted in the form of tweets about the company. This paper has 2 different textual representations Word2vec and Ngram, to analyze the public sentiments in tweets. The challenge faced by the author is to develop their sentiment analyzer because analyzers are trained with a different corpus. So, movie corpus and stock corpus analyzers are not the same. Therefore, they developed their sentiment analyzer for stock. In this paper, the authors provide the solution by applying sentiment analysis and the supervised machine learning principle. This principle was applied to the tweets and analyzed the correlation between stock market ups and down and sentiments in tweets. The results show a strong correlation between stock market movement and the public sentiments expressed in tweets.

Another publication like ”VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text”, C.J. Hutto Eric Gilbert (Issue 2015) [3] is about an underlying sentiment (or opinion) lexicon is used in a large variety of sentiment analysis algorithms. A sentiment lexicon is a collection of lexical features (such as words) that are classified as positive or negative (For example Linguistic Inquiry Word Count (LIWC), General Inquirer (GI), and Hu-Liu04) or along with valence scores for sentiment intensity in lexicons like ANEW, SentiWordNet, and SenticNet depending on their semantic orientation. Manually constructing and validating such lists of opinion-bearing attributes is one of the most time-consuming approaches for generating trustworthy sentiment lexicons, despite being one of the most robust. As a result, much of the applied sentiment analysis research depends largely on pre-existing human-built lexicons or uses machine learning algorithms to ”learn” the sentiment-relevant aspects of the text. Despite their widespread use for assessing sentiment in social media environments, lexicon-based sentiment analysis systems have two major drawbacks: 1) they have coverage issues, frequently neglecting essential lexical elements that are particularly relevant to social writing, such as acronyms, initialisms, emoticons, or slang in microblogs, and 2) certain lexicons disregard general sentiment intensity differentials for features within the lexicon. On the other hand, machine learning algorithms have difficulties when it comes to finding sentiment-relevant elements in the text. For instance, they necessitate large training data, which might be difficult to get, as with verified sentiment lexicons. Second, they expect the training set to have as many variables as feasible. Third, they are typically more computationally costly in terms of CPU processing, memory needs, and training/classification duration. VADER (Valence Aware Dictionary for sEntiment Reasoning) is a parsimonious rule-based model for social media text sentiment analysis. It overcomes these obstacles by combining qualitative and quantitative methodologies to create and then experimentally validate a gold-standard sentiment lexicon that is especially suited to microblog-like situations. It performs well with social media text while also applying to other domains. It is quick enough to utilize with streaming data online and does not suffer from a significant speed-performance trade-off.

Another paper ”Transformer Based Multi-Grained Attention Network for Aspect-Based Sentiment Analysis”, JIAHUI SUN 1, PING HAN 2, ZHENG CHENG 1, ENMING WU 1, AND WENQING WANG (Issue 2020) [4] is about the sentiment polarity for each aspect of a sentence is predicted using aspect-based sentiment analysis (ABSA). The majority of current approaches are based on sequence models, superimpose the emotional semantics of various tendencies, and lack syntactic structure information. Most models use a coarse-grained attention mechanism, which still has issues with weak aspect-context interactions. The most common issue with models performing ABSA tasks is when the aspect is a phrase rather than a single word. Although the average vector of several words is commonly used as a representation of a particular aspect, it cannot accurately reflect the characteristics of each word in a phrase. Further, it leads to the loss of information. Additionally, they perform single average pooling to learn the attention weight of each word in the context-specific aspect, which leads to the loss of useful information. In this paper, these issues are overcome with the help of Transformer based multi-grained attention network (T-MGAN) model proposed by the authors of this research paper. It uses the transformer module to learn word-level representations of aspects and contexts, and the tree transformer module to learn phrase-level representations of the context. To choose the key aspects and context characteristics, this model uses the dual-pooling method, which can limit the loss of learned features. The attention mechanism is then used several times to characterize the word-level and phrase-level interactions between aspect and context words.

”Distributed Real-Time Sentiment Analysis for Big Data Social Stream.”, Amir Hossein Akhavan Rahnama. (Issue 2016) [4] is as the data sources increase and the speed of data grows, more challenges arise in querying and data mining. Currently, real-time sentiment analysis is still not applicable (Bifet, Albert, 2010). Thus, solving the real-time sentiment analysis regarding the growing data is the main challenge. Problems need to be solved in data processing, architecture, classification, etc. to achieve real-time sentiment analysis. In the Processing of the data stream, the synopsis data structure can minimize the reaction time. The data sources and streams can be read by ADWIN, which is used to solve the growing data sources and streams. To solve the challenges in the process of data mining, Rahnama suggests the algorithm: Vertical Hoeffding Tree, a parallel decision tree classifier, which has significantly better accuracy and shorter measuring time than using Multinomial naïve Bayes and Hoeffding tree.

Another paper "Integrated Real-Time Big Data Stream Sentiment Analysis Service.", Sun Sunnie Chung, Danielle Aring, (Issue 2018) [4] is about high speed of data processing, massive datasets, and complex structures bring challenges to clean, transform, and process data. (Chung, S. S., Aring, D., 2018). The arrival of high-speed and massive data also challenges the time and space processing of the data stream. Regarding real-time sentiment analysis, the accuracies of predictions and stability of the model become the problems that need to be solved. To store the data stream, the Sentiment analysis service uses a multi-layer structure: Data Storage and Extraction Layer, Data Stream Layer, Data Preprocessing and Transformation Layer, Feature Extraction Layer, Prediction Layer, and Presentation Layer. (Chung, S. S., Aring, D., 2018) to process data for real-time sentiment analysis. To solve the challenges of stabilizing the model and improving the accuracies of the sentiment models, the Sentiment analysis service applies the Deterministic model and Probabilistic model. By using cross-validation with multinomial Naïve Bayes Classifier for the models, accuracies and stability can be ensured. Thus, the Sentiment analysis service can perform efficient real-time data processing. [4].

III. Problem Definition

Using the dataset from Kaggle, Ukraine Conflict Twitter Dataset[7]. This data contains 36.28 million tweets and around 11 Giga Bytes datasets in the form of Binary data about the conflict and is updated every day. The dataset consists of multiples columns such as userid, username, access (user bio), location, following (count), followers (count), totaltweets, usercreatedts (time), tweetid, tweetcreatedts (time), retweetcount, text, hashtags, language, coordinates, favorite count, extractedts (time).

Various people and groups tend to speak different languages in Russia because it is a multilingual nation. More than 150 languages were identified by linguistic experts, starting with Russian, which is spoken by 96.25 percent of the world's population. Numerous studies were analysed and addressed the simultaneous analysis of multiple languages, which gives them the ability to cover a wider range of informational compare the emotion represented in several sources and languages discussing the same subjects. We translated the sentiment scales from Russian into English languages in order to classify simple monolingual sentiment analysis by translating text into a single language. We observed that the sentiment polarity of Russian language and English language are different for the same tweets. In this paper, we are comparing the Russian tweets and English translated tweets and finding the solution for the difference we observed while analysing same tweets in two different language.

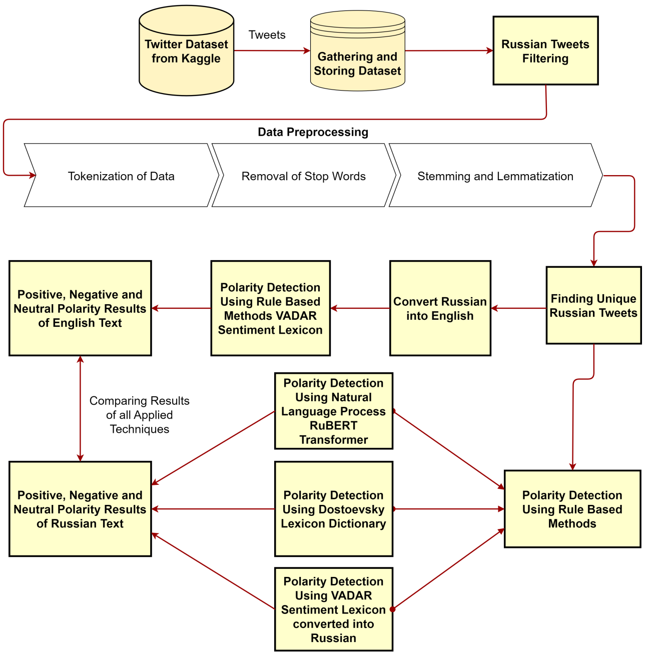
Twitter Dataset from Kaggle
Tweets
Gathering and Storing Dataset
Russian Tweets Filtering
Data Preprocessing
Tokenization of Data
Removal of Stop Words
Stemming and Lemmatization
Positive, Negative and Neutral Polarity Results of English Text
Polarity Detection Using Rule Based Methods VADAR Sentiment Lexicon
Convert Russian into English
Finding Unique Russian Tweets
Comparing Results of all Applied Techniques
Polarity Detection Using Natural Language Process RuBERT Transformer
Positive, Negative and Neutral Polarity Results of Russian Text
Polarity Detection Using Dostoevsky Lexicon Dictionary
Polarity Detection Using Rule Based Methods
Polarity Detection Using VADAR Sentiment Lexicon converted into Russian

Fig. 1. System Architecture / System Flow Diagram

IV. System Model Overview

To start with, we collected data from the Ukraine conflict twitter dataset and stored them. Then we filtered Russian tweets from the dataset. Next, the data preprocessing consists of three steps: 1. Tokenization, which tokenize each single word in the sentences; 2. Removal of stop words, which removes unnecessary words in the sentences; 3. Stemming and lemmatization, which settle the word in the root form and present tense; After preprocessing the data, we analyzed sentiments of Russian texts by two paths. 1. Converting Russian into English and using VADER Lexicon sentiment to detect the polarity shift in the English text. 2. Our method of using VADER Sentiment Lexicon is English Language sentiment lexicon which we have converted into Russian, Dostoevsky Lexicon, and RuBERT Transformer to detect positive, negative, and neutral polarity results of Russian text. Finally, we compared results of all techniques we applied and concluded the outcome.

V. Problem Formulation

In order to identify the attitude of Russian tweets, we have done sentiment analysis of Russian tweets using four approaches. In each of them, the texts were categorised into 3 classes such as Negative, Neutral, and Positive. Before implementing all the methods, data has been through the process of data pre-processing.

A machine cannot correctly interpret language in its natural state, therefore we process the language to make it simpler for the computer to understand better. Tokenization, or the division of strings into smaller units called tokens. This is the first data preprocessing step in making meaning of the data. In

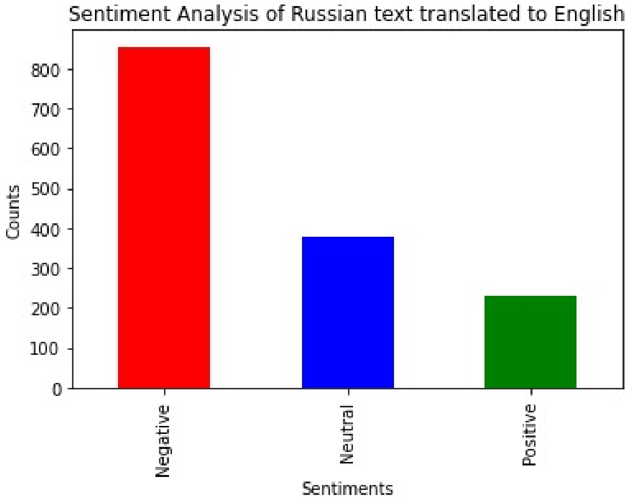

Fig. 2. Sentiment Analysis of Russian Text Translated to English

writing, a token is a group of letters that functions as a whole. The tokens might be made up of words, emojis, hashtags, URLs, or even single characters, depending on how we make them. By dividing the text into tokens based on white space and punctuation, language may be broken down into fundamental units. Furthermore, the Stemming Technique involves eliminating affixes from words. Stemming is a technique that trims the endings of words while only using simple verb forms. The word is transformed into a normalised form via the lemmatization algorithm, which analyse the word's structure and context. Then, we extracted 1455 Russian texts from the data set and translated the unique ones into English language. They are all analyzed by the VADER lexicon sentiment analysis. VADER is a English language lexicon and rule-based sentiment analysis tool used to expressed sentiments in social media. To analyse the sentiments in VADER we calculate the compound score by adding the valence score of every word in the lexicon followed by the rule and then normalised between -1 to +1. We use below normalisation,

$$x=\frac{x}{\sqrt{x^2+c}}$$

where, x is sum of valence score and c is normalisation constant

In the Fig. 2 we observe that all the tweets divided into 3 classes: Negative, Neutral, and Positive. We found the majority of tweets were classified as negative and almost equal proportions were classified as Neutral and Positive. Each of them did not reach half of negative tweets.

In the next approach we applied another sentiment analysis library: Dostoevsky package on Russian texts and categorized texts into the same classes as in the first approach. As shown in the Fig. 3 that majority of tweets were classified as Neutral. Positive and negative tweets did not reach half of neutral ones and they almost took equal proportion. The result of this approach is not similar with which translated Russian text to English.

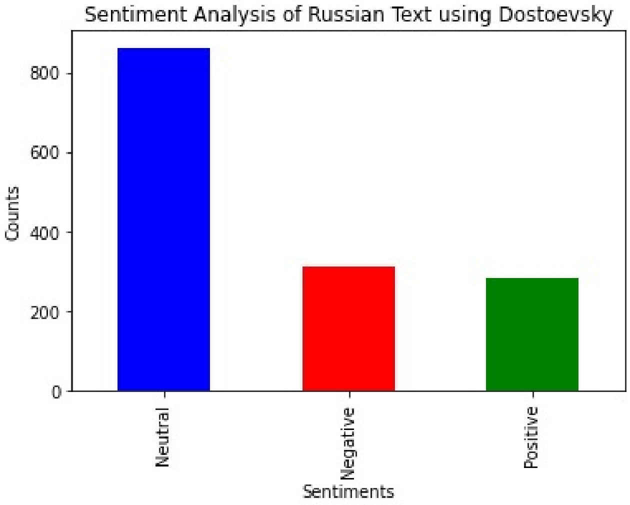

Fig. 3. Sentiment Analysis of Russian Text

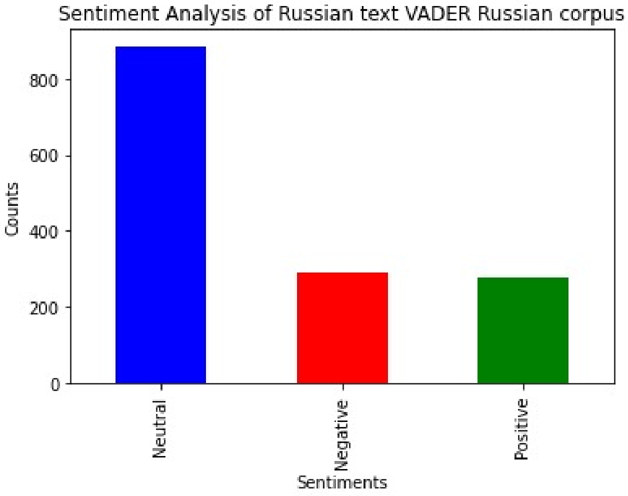

Fig. 4. Sentiment Analysis of Russian Text VADER Russian Corpus

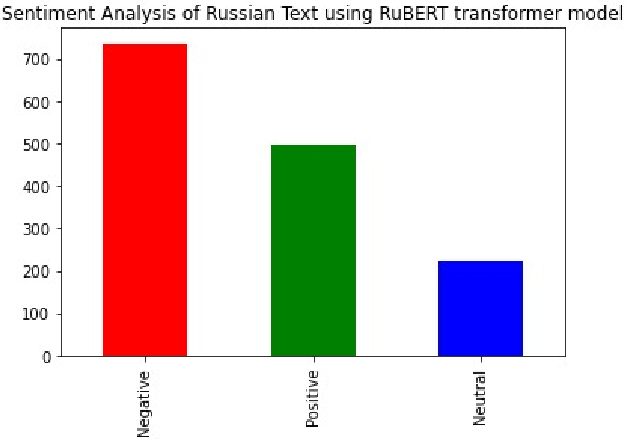

Fig. 5. Sentiment Analysis of Russian Text using RuBERT transformer model

Next, refactoring the source code for VADER sentiment lexicon was performed for Russian language. In this approach we extracted the words list from the VADER sentiment lexicon and translated them to Russian assuming that the intensity rating will be consistent after translation. After updating the intensity rating for all the English words to corresponding Russian ones, we run the same code to calculate the sentiment intensity for Russian text and the result can be observe in Fig. 4. We didn't get the expected outcome, but our results were similar to the results obtained from the second approach using Dostoevsky library.

Finally, RuBERT model was performed in the sentiment analysis. In this approach, we applied Fine-tuned Multilingual Encoder, RuBERT, and Multilingual Universal Sentence Encoder for the sentiment classification in Russian tweets. The categorization is the same as the previous ones: Neutral, positive, and negative. As presented in Fig. 5, we found negative tweets take the majority, which was similar with the result of translating Russian to English. In addition, such a result is consistent with English Tweets sentiment analysis. Thus, the outcome produced by performing RuBERT transformer model was the most ideal for our sentiment analysis.

According to the result, we observed that the sentiment analysis performed using the RuBERT Transformer Model shows the best outcome to problem of dissimilarity of positive, negative and neutral polarity of same tweets but in two different languages, English and Russian. Polarity in the RuBERT transformer model shows almost similar kind as it could classify majority of Russian tweets as negative as expected as they are all based on Ukraine Russia Special Military Operation.

VI. Challenges

1) Implementing the Dostoevsky Package:
The main challenge we faced while implementing the Dostoevsky package was to install its dependencies. It requires the C++ build tools to install all dependencies of this package. Therefore, we installed Microsoft Visual Studio C++ for additional 25 Gigabytes tools in our system. After installing Microsoft Visual Studio C++, we installed all the dependencies and can able to get the results of Russian Sentiment Tweets.

2) Converting VADER Lexicon word into Russian:
Another challenge we faced while converting VADER Lexicon words into Russian was that we had translated all the English words manually, available in VADER Lexicon. The result obtained from this technique is similar to the result obtained from technique in which Dostoevsky package used.

3) Computational Cost:
The computational cost for this project starts with downloading the 11 Gigabytes Ukraine Russia Conflict dataset from Kaggle, which requires a good internet connection and storage space on the disk. Additionally, the dataset contains a collection of gzip files, accessed using a for loop and decompressed simultaneously.

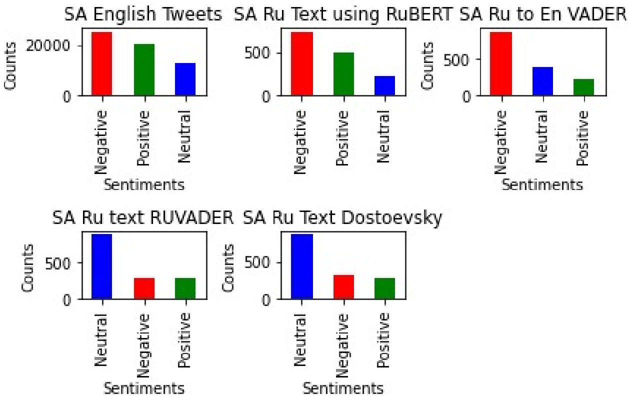

Fig. 6. Comparing applied methods and finding best approach

VII. Simulation Results

In the result, we have observed that the sentiment analysis performed in Russian Language using RuBERT transformer is identically showing the same pattern as it is observed in the VADER sentiment in English language as shown in Fig.6 comparison of methods. We can say that RuBERT Transformer gives the best solution because the pattern of classification in RuBERT Transformer matches with the VADER classification which performs best among the other methods.

VIII. Discussion

The results of the RuBERT model and VADER sentiment analysis after translating Russian to English imply that dissimilarity still exists in the counts of neutral and positive tweets. Opposition claims might argue that the result of RuBERT is not consistent with that of VADER analysis after translation. However, it is noticeable that negative tweets took the majority part as the conspicuous result in the analysis, which is the focus of the paper. Additionally, the result of the RuBERT model is consistent with that of English sentiment analysis which consists of the biggest part of the dataset. Thus, we still consider that the RuBERT model gives the best solution.

IX. Conclusion

As military conflict between Russia and Ukraine erupts, it becomes the centroid of social media attracting people to post and comment. In this paper, multilingual tweets of twitter were extracted and gathered into the dataset. By breaking languages into fundamental units, eliminating affixes and transforming them to normalized forms, the data was processed to be in the state which a machine could interpret. In order to understand the emotion of Russian texts and investigate which is the most ideal sentiment analysis method, two paths were performed. On the one hand, Russian was converted into English and

VADER lexicon was conducted afterwards. On the other hand, Dostoevsky package, VADER lexicon for Russian, and RuBERT model were performed to conduct sentiment analysis. The results illustrated that RuBERT model obtained similar outcome with VADER sentiment analysis after translating Russian into English. Furthermore, the results of sentiment analysis using RuBERT are consistent with that of English tweets, which are the majority of the dataset. The results of these analysis highlighted that negative tweets took the dominant place in twitters of Russian languages. In the future, further research can be done to decrease the dissimilarity of negative, neutral, and positive tweets by performing VADER lexicon after translation and RuBERT model.

References

[1] ”Multitask Learning for Fine-Grained Twitter Sentiment Analysis”, Georgios Balikas, Simon Moura, Massih-Reza Amini (Issue 2017)

[2] Sentiment Analysis of Twitter Data for Predicting Stock Market Movements”, Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, Babita Majhi (Issue 2016)

[3] ”VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text”, C.J. Hutto Eric Gilbert (Issue 2015)

[4] ”Transformer Based Multi-Grained Attention Network for Aspect-Based Sentiment Analysis”, JIAHUI SUN 1, PING HAN 2, ZHENG CHENG 1, ENMING WU 1, AND WENQING WANG (Issue 2020)

[5] ”Distributed Real-Time Sentiment Analysis for Big Data Social Stream.”, Amir Hossein Akhavan Rahnama. (Issue 2016)

[6] ”Integrated Real-Time Big Data Stream Sentiment Analysis Service.”, Sun Sunnie Chung, Danielle Aring, (Issue 2018)

[7] https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows

[8] “Identifying top sellers in underground economy using deep learning-based sentiment analysis,” in Proc. IEEE Joint Intell. Secur. Informat. Conf., W. Li and H. Chen (Issue 2014)

[9] “Sentiment-analysis-in-russian: Fine-tuned multilingual Bert and multilingual use for sentiment analysis in Russian. Rureviews, Rusentiment, Kaggle Russian news dataset, LINIS crowd, and RuTweetCorp were utilized as training data.” Sismetanin. (Issue 2021).