Web-sentiment Analysis Of Public Comments (Public Reviews) For Languages With Limited Resources Such As The Kazakh Language

Dinara Gimadi, Richard Evans, Kiril Simov
University of Wolverhampton, New Bulgarian University

{gimadi.dinara}@mail.ru

Abstract

In the pandemic period, the stay-at-home trend forced businesses to switch their activities to digital mode, for example, app-based payment methods, social distancing via social media platforms, and other digital means have become an integral part of our lives. Sentiment analysis of textual information in user comments is a topical task in emotion AI because user comments or reviews are not homogeneous, they contain sparse context behind, and are misleading both for human and computer. Barriers arise from the emotional language enriched with slang, peculiar spelling, transliteration, use of emoji and their symbolic counterparts, and code-switching.

1 Introduction

Top free Google Play apps in Kazakhstan are Ashyq, Telegram Messenger, Kaspi, Instagram, e-Gov mobile, OLX, Kolesa, and others. Kazakh-language application reviews are of high interest for companies developing and marketing mobile apps. They are valued by businesses to influence the decision-making of a potential customer and evaluate overall satisfaction or dissatisfaction with products, their new features, and services. The given problem predetermines the following obstacle: in the context of the absence of a standard dataset, comparison of techniques applied, and results obtained is a challenging task. The problem of sentiment classification of Kazakh comments includes code-switching between Kazakh and Russian languages by using Cyrillic script, ambiguous words, and emoji, especially the problem of combining negative emoji with a positive comment or vice versa, which have an impact on the overall tonality of the review. The relevance of the topic is justified by the scarcity of developed models and methods of sentiment analysis of the opinions of social networks users in the Kazakh language. Collection of 3000 labeled reviews data set on web apps and development of a method for achieving more correct results is the main goal, which consists of data compilation, pre-processing part, and classification by testing an aspect-based method.

2 Previous work

Researchers in the area of sentiment analysis and opinion mining have achieved significant progress. According to (Liu et al., 2018), sentiment is a subjective feeling expressed in text. Pang and Lee et al. (2008) stated that analysis of subjective feelings in texts refers to the assemble of techniques from fields such as natural language processing (NLP), information retrieval, and machine learning, to identify and extract subjective information from text reviews

Proceedings of the Student Research Workshop associated with RANLP-2021, pages 65–68,
held online, Sep 1–3, 2021.
https://doi.org/10.26615/issn.2603-2821.2021_010

and classify the documents, sentences or phrases into sentiment categories such as positive, negative and neutral. Sentiment analysis of public comments in Google Play Market has performed automatically with Google Cloud Natural Language API. Google Natural Language processing API is a pre-trained machine learning API for Google sentiment analysis, entity recognition, and syntax analysis. Google Cloud Natural Language sentiment analysis calculates a predicted floating-point value between -1 and 1 to evaluate the entire text string's emotional coloring. There are various directions, in which researchers paved their ways, to gain significant developments. As a rule, sentiment analysis is carried out by applying a linguistic method or a machine learning method, and neural networks. Machine learning techniques include Support Vector Machine (SVM), Logistic Regression, Naive Bayes, Random Forest, Linear Regression, k nearest neighbor (k-NN), and others. Machine learning techniques are supervised if they use pre-existing/ collected opinion datasets, which means 'they have a data set of input observations, each associated with some correct output (a 'supervision signal') according to (Jurafsky and Martin, 2020). Then it automatically learns all types of linguistic attributes and computes a model for each corpus. The trained model is later can be used for the classification of the test dataset. Petter Mæhlum et al. (2019) annotated evaluative reviews at three levels in the Norwegian language concerning being on-topic or not, and expressing the first opinion of the author or not, and a binary classification task about positive and negative sentiments. In this dataset sentences with positive, neutral, and negative categories were hand-labeled by reviewing lexical features and context. Also in Google PlayMarket there are existing labeled sentences. Labeled data includes the following categories: binary – positive and negative sentiment, ternary – positive, neutral and negative sentiment, and quinary – strongly positive, positive, neutral, negative, and strongly negative. Jurafsky and Martin (2020) indicated Naïve Bayes is a generative classifier, which is based on conditional probability, according to which the document belongs to the class, and uses the Bayes formula. The linguistic method by Yergesh et al. (2017) supposed designing sentiment dictionaries of Kazakh words and phrases as well as a set of linguistic rules, in which nouns and noun groups were considered as aspect categories. For example, the app 'Kaspi. Kz bank' has aspects such as technical issues, language, interface, and service to address the opinions of users. Users mention the aspects of the object in a comment with sentiment words. Aspect-based sentiment analysis correlates phrases to the aspects (categories that justify the reason for emotional evaluation). Aspect is a specific topic or feature to which sentiment will be assigned. In user comments from Google Play Market aspects for mobile applications are interface, language, technical performance, design, individual issues, and others. Sitaram et al. (2015), when extracting sentiments from the mixed language of Hindi and English, combined the sentiment of these sentence extracts incrementally, and trained data. Kozhirbayev et al. (2018) approached code-switching in a single two-step framework, consisting of unsupervised normalization and Naïve Bayes classification. Linguistic code-switching (LCS) is a linguistic phenomenon, when two or more languages are used in one sentence shifting traditional grammatical structure of the source language. Moreover, they applied model based on LSTM networks to classify reviews of users. Researchers Zharmagambetov et al. (2016) applied Word2Vec in the sentiment classification task for Kazakh and Russian languages and used these word vectors as an input to deep recurrent neural networks to deal with long term dependency of the textual document.

3 Methodology

This research focuses on the development of a new method for aspect-based sentiment analysis for Kazakh-language reviews of Android apps posted on the Google Play Market. The research questions set during the research planning are:

1) Is the rule-based approach for identifying different aspects of apps mentioned in the reviews, in the context of scarce data, more accurate than the data-driven approach?

2) What techniques of machine learning are relevant to overcome challenges related to the scarcity of language resources and tools?

3) To what extent the output obtained from rule-based and data-driven methods for aspect-based sentiment analysis will correlate with each other?

4) How do emoji contribute to proper sentiment extraction in line with a textual message?

5) Is code-switching in Kazakh-Russian combined sentences can be resolved for sentiment classification by applying machine learning techniques?

The relevance of the research topic is justified by the scarcity of developed models and methods of sentiment analysis of the opinions of social networks users in the Kazakh language. Collection of 3000 labeled reviews data set on web apps and development of a method for achieving more correct results is a contribution and can be used for future research. The object of the research is the user reviews from Android Google Play Market in the Kazakh language. The subject of the research is sentiment analysis methods and algorithms, as a special case of semantic analysis of texts in natural language.

The main steps of sentiment analysis are data compilation and preprocessing, as well as classifying it with the use of machine learning algorithms and neural networks. The first stage contains an accumulation of Kazakh-language reviews of Android apps posted on the Google Play Market. Classification consisted of labeled data such as positive, negative, neutral labels based on my personal standpoint, reviews containing only emoji (without words), emoji and words, reviews with code-switching, Google Play Market ratings (1-5 stars), binary classification by Google Play Market. The compiled data consists of 3000 annotated tokens, which can be used for training and testing.

The second stage includes data preprocessing that implies clearing data from unneeded characters and punctuation marks, converting all words to lower case, extracting stop words that do not possess any semantic load, and removing affixes and endings (stemming), or modifying words into an indefinite form (lemmatization). Up to far tokenization was performed with the use of Python, reduced all tokens to lower case, added notes containing corrected tokens for misspelled tokens, which refers to any type of data created by users, and far from standards of literary language. National Laboratory of Kazakhstan contributed to developing the KazNLP tool for text preprocessing, and NLA-NU laboratories developed Kazakh Dependency Treebank for parsing.

In the third stage, it will be tested whether a rule-based approach for identifying different aspects of apps mentioned in the reviews, in the context of scarce data, is more accurate than a data-driven approach.

4 Current progress

At the current stage, 3000 Kazakh comments are compiled. They were labeled manually according to emotional polarity and compared with pre-defined sentiment classification on Google Play Market, The second stage was followed by defining code-switching, peculiar misspelling, and emoji identification. At the third stage, there is an ongoing process of text pre-processing. User comments were tokenized, lemmatized, and keywords were extracted in Sketch Engine to define a list of aspects. The next step will be processed in order to define the effective way of applying machine learning tasks for sentiment extraction.

Acknowledgements

I would like to express my gratitude to my supervisors Dr Richard Evans (University of Wolverhampton) and Prof Kiril Simov (New Bulgarian University) for their support, recommendations and valuable feedback for my research. I would also like to appreciate Prof Ruslan Mitkov, the EM TTI coordinator, he encouraged me to be brave enough to share my experience with this research.

References

Bing Liu, Lei Zhang and Shuai Wang. 2018. Deep learning for sentiment analysis. *Wiley*

Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4):e1253.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1-135. https://doi.org/10.1561/15000000114

Daniel Jurafsky and James H. Martin. 2020. Naive Bayes and Sentiment Classification. *Speech and Language Processings*. Stanford University Press, USA. https://web.stanford.edu/~jurafsky/slp3/

Petter Mæhlum, Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian. SANT project (Sentiment Analysis for Norwegian Text. In *the Proceedings of SANT project (Sentiment Analysis for Norwegian Text), Research Council of Norway, 10*

Banu Yergesh, Gulmira Bekmanova, and Altynbek Sharipbay. 2017. Sentiment analysis on the hotel reviews in the Kazakh language. In *Proceedings of the International Conference on Computer Science and Engineering (UBMK).* Pages 790-794. https://doi.org/10.1109/UBMK.2017.8093531

D. Sitaram, Savitha Murthy, Debraj Ray, Devansh Sharma, and Kashyap Dhar. 2015. Sentiment analysis of mixed language employing Hindi-English code switching. In *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC)*. Pages 271-276.

Zhanibek Kozhirbayev, Zhandos Yessenbayev, and Aibek Makazhanov. 2018. Document and Word-level Language Identification for Noisy User Generated Text. In *Proceedings of IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*. Pages 1-4.

Arman Zharmagambetov and N.S. Sakenovich. 2016. On One Approach of Solving Sentiment Analysis Task for Kazakh and Russian Languages. In *Proceedings of ICCCI 8th International Conference on Computational Collective Intelligence.* Halkidiki, Greece, pages 1-9.
https://doi.org/10.1007/978-3-319-45246-3_51