Comparison of LR, NB and SVM Techniques on Sentiment Analysis of Azerbaijani Texts

Mehdi Rasul
Landau School, Baku, Azerbaijan
mehdi@rasul.az

Abstract—**The prediction of sentiment in text has been a challenging problem across domains. This paper analyzes movie reviews in English and Azerbaijani using machine learning techniques like Logistic Regression, Naive Bayes, and Support Vector Machine (SVM). The results indicate the importance of language corpus and compare performance between languages.**

Keywords — sentiment analysis, machine learning, Logistic Regression, Naïve Bayes, Support Vector Machine

I. INTRODUCTION

Sentiment analysis, also known as opinion mining, identifies and extracts subjective information such as opinions, emotions, and attitudes from text data using natural language processing (NLP) and computational linguistics. It has applications across business, politics, and social media monitoring. Companies use it to evaluate customer feedback, monitor reputation, forecast behavior, etc.

Recent research shows machine learning algorithms, particularly supervised learning and deep learning, produce good results in automating textual sentiment analysis. Studies show that combining computer science and math can improve accuracy in sentiment analysis automation.

Well-defined English language corpus aids in building powerful and accurate models for English text analysis. The abundant NLP libraries and tools in English, like word correction, grammar checking, tokenization, etc. make accurate sentiment models easier to create. However, research on under-resourced languages like Azerbaijani is limited despite its linguistic distinctness and approximately 30 million speakers worldwide. The lack of labeled data and

For the Azerbaijani language, Rustamov and Hasanli

classified 12,000 Azerbaijani tweets into positive and negative sentiment using LR, NB and SVM with accuracy up to 94% using bag-of-words features [5]. Suleymanov et al. (2019) did text classification on Azerbaijani news articles, finding SVM with TF-IDF features gave 93% accuracy, outperforming NBat 56.53% accuracy [6].

Overall, prior research indicates machine learning approaches can effectively classify sentiment for both Azerbaijani and English texts. Algorithms like SVM, NB and LR have been most widely used.

sophisticated corpus in Azerbaijani makes sentiment analysis uniquely challenging.

This paper compares several machine learning techniques on an English and Azerbaijani movie review dataset including Logistic Regression (LR), Naive Bayes (NB) and Support Vector Machine (SVM). The results address the importance of language corpus and differences in techniques for Azerbaijani versus English sentiment analysis.

II. REVIEW OF LITERATURE

There have been several studies that have applied machine learning techniques for sentiment analysis across different domains.

Neethu and Rajasree performed sentiment analysis on Twitter data using classifiers like SVM, NB, Maximum Entropy and Ensemble methods [1]. They achieved up to 90% accuracy. Chandra and Jana (2020) also used machine learning and deep learning models on Twitter data, finding that LSTM gave 97% accuracy while LR, NB and SVM gave around 82% accuracy [2].

For political sentiment analysis, Heredia et al. used deep convolutional neural networks to analyze tweets and predict the 2016 US elections with 84% accuracy [3]. Additionally, in Indonesia, tweets of @jokowi, account of current President of the Republic of Indonesia, have been scraped and fitted into different machine learning models to find the sentiment, specifically positive and negative labels. The results indicated that Sequential Minimal Optimization (SMO) gave 82.7% of accuracy, precision and recall [4].

III. EXPERIMENT

In the project, Python has been utilized in data cleaning and modeling stages. Specifically, Python built-in library sk-learn was useful by simplifying the use of various ML algorithms. The trained dataset is about movie reviews in English, and it was translated into Azerbaijani. There are 2000 reviews, half of which was in Azerbaijani, and the other half was in English language. The process for both Azerbaijani and English languages has been outlined below.

A. *Azerbaijani Language Processing Approach*

Data Cleaning

In the data cleaning phase, several essential steps were undertaken to prepare the Azerbaijani dataset for sentiment analysis. Words with fewer than 3 characters were removed, and non-alphabetical characters, such as digits and punctuations, were eliminated because they do not contribute to the sentiment of the text. To maintain dataset relevance, a comprehensive list of Azerbaijani stop-words was manually created and applied to the text. To ensure consistency, all text entries were converted to lowercase, and sentence tokenization was employed.

Modeling

To enhance result quality, experiments were conducted using various machine learning models, specifically LR, NB and SVM. Different feature extraction techniques, including TF-IDF and Countvectorizer, were explored. Given the relatively small dataset, the k-fold cross-validation method was employed to ensure robust model training and evaluation.

In the initial stage of the pipeline, text data was tokenized and countvectorized using a (1,2) n-gram with a word analyzer. The data was divided into an 80-20 train-test split. Among the models tested, LR, with BOW, achieved an impressive 84.25% accuracy and 86% precision (Table II). However, the models exhibited susceptibility to overfitting, with training accuracy values ranging from 95% to 99%. To mitigate this, regularization techniques were applied.

Table II. Accuracy Scores with Default Parameters

Algorithm	**TF-IDF**	**BOW**
LR	86%	84.25%
NB	74%	79%
SVM	85.25%	77.5%

Later, regularization was applied to minimize the possibility of overfitting. Overfitting is the phenomenon in machine learning when the model adapts to the training data too much, and, consequently, the model performs exceptionally well on training data, but poorly on testing data. Therefore, it is preferred to build a model that achieves similar performance with both training and testing data and over the predefined thresholds. After L1 regularization was applied, most models still showed either no or very little improvement in accuracy score. The only model which saw significant increase in accuracy score from L1 regularization was NB with BOW. The results show that NB algorithm with BOW produced 86% accuracy, which is a 7% increase from the result with default parameters.

In a separate experiment, the k-fold cross-validation approach was adopted, dividing the dataset into equally sized folds and iteratively using one-fold for testing while training on the remaining k-1 folds. This method allowed assessment of the models' performance to be more comprehensive. The accuracy scores for the three selected algorithms are as follows:

Table III. Accuracy Scores using K-fold Approach

Algorithm	**TF-IDF**	**BOW**
LR	81%	84.6%
NB	78%	79%
SVM	81%	82%

B. Language Processing on English Dataset

An English version of the same dataset was utilized in the modeling process for predicting the general sentiment of reviews. Python, enriched with built-in libraries for English language modeling, particularly through nltk, facilitated tasks like word tokenization, stemming, and feature extraction without requiring custom implementations. Text files were read in Python and structured using the Pandas library.

In subsequent stages, common stop-words like "and," "or," and "is" were removed from the texts using a list obtained from the nltk corpus library. Additionally, each text was tokenized using the word tokenizer function. The Porter Stemming method was applied to maintain the root of various word versions in the dataset, ensuring consistency across words like "computers," "computation," and "computed," which were all transformed to "comput" in the texts.

Further preprocessing involved retaining words with a minimum length of 4 characters while removing those with 3 or fewer characters.

In the modeling stage, various machine learning algorithms were employed. Initially, models were trained with default parameter values. Subsequently, regularization was applied to address issues such as overfitting. Data was divided into train and test datasets, with the k-fold approach also implemented to assess model performance and minimize overfitting risks.

Based on default parameter values, LR with TF-IDF feature extraction yielded an accuracy of 87%, while SVM and Multinomial NB with TF-IDF achieved 85% accuracy (Table IV).

Table IV. Accuracy Scores for
Sentiment Analysis for Dataset in English

Algorithm	**TF-IDF**	**BOW**
LR	87%	78%
NB	85%	81%
SVM	85%	82%

IV. CONCLUSION AND FUTURE WORK

In this research, various ML algorithms have been used to predict the sentiment of the movie reviews in both English and Azerbaijani. The study has indicated the achievements attained in building models with various techniques. TF-IDF and BOW (or Count Vectorizer) have been implemented for feature extraction methods from the texts and tested in different models, namely LR, NB, and SVM.

For Azerbaijani version of the dataset, combined with TF-IDF feature extraction, LR and SVM algorithms produced the best result (81% accuracy, measured as the mean of 10 folds' results). When BOW feature extraction was used, LR produced accuracy of 84%, while SVM reached 82% accuracy score.

The same dataset in English has also been modeled to compare the results with different preprocessing techniques that were not applied in Azerbaijani version, such as stemming, stopwords list, etc. Overfitting was much less of as issue for models based on English dataset. The highest score was attained with BOW feature extraction combined with LR method. Additionally, SVM and NB algorithms performed well with TF-IDF feature extraction method and achieved 85% accuracy score.

The research indicates that the results are similiar for both language models. However, it should be noted that the dataset was relatively small for sentiment analysis. Therefore, the amount of data in the dataset should be increased so that a better comparison can be made.

Python has enriched libraries for building language models in English. These include as list of stopwords, word tokenization, stemming and lemmatization techniques. Particularly, English language corpus is well developed, which helps achieve higher results. However, Azerbaijani language lacks the language corpus which makes it harder to build generalized models. Also, there are not any libraries like NLTK and BERT for Azerbaijani language, which makes creation of accurate sentiment analysis models a lot harder. Therefore, a lot of resources should be dedicated to research and creation of language corpus for Azerbaijani Language, which would help getting higher accuracy scores.

ACKNOWLEDGEMENT

I am extremely grateful to Mr. Mammadzada for his guidance and technical support throughout this research experiment. In addition, I am thankful to Dr. Rustamov for providing the essential dataset for the research project.

REFERENCES

[1] M.S. Neethu & R. Rajasree (2013). “Sentiment analysis in twitter using machine learning techniques”. Proceedings of the 2013 Fourth International Conference on Computing, Communications and Networking Technologies, 2013, pp.1-5. https://doi.org/10.1109/ICCCNT.2013.6726818

[2] Y. Chandra & A. Jana (2020) “Sentiment analysis using machine learning and deep learning”. Proceedings of the 2020 7th International Conference on Computing for Sustainable Global Development, New Delhi, India, pp.1-4. https://doi.org/10.23919/INDIACom49435.2020.9083703

[3] B. Heredia, J.D. Prusa, & T.M. Khoshgoftaar (2018). “Location-based twitter sentiment analysis for predicting the U.S. 2016 Presidential Election”. Proceedings of the 31st International Florida Artificial Intelligence Researc Society Conference, vol. 2009, pp.265–270.

[4] F.A. Wenando, R. Hayami, Bakaruddin & A.Y. Novermahakim (2020). “Tweet sentiment analysis for 2019 Indonesia presidential election results using various classification algorithms”. Proceedings of the 2020 1st International Conference on Information Technology, Advanced Mechanical and Electrical Engineering, Yogyakarta, Indonesia, pp.279-282. https://doi.org/10.1109/ICITAMEE50454.2020.9398513

[5] H. Hasanli & S. Rustamov (2019). “Sentiment analysis of Azerbaijani twits using logistic regression, Naïve Bayes and SVM”. Proceedings of the 2019 IEEE 13th International Conference on Application of Information and Communication Technologies, Baku, Azerbaijan, pp.1-7. https://doi.org/10.1109/AICT47866.2019.8981793

[6] U. Suleymanov, B.K. Kalejahi, E. Amrahov, & R. Badirkhanli (2019). “Text classification for Azerbaijani language using machine learning and embedding”. arXiv:1912.13362 https://doi.org/10.48550/arXiv.1912.133