Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

DOI: 10.37943/17RHPH9724

Mukhtar Amirkumar
BSc, Assistant Instructor, Computer Sciences Department
mukhtar.amirkumar@sdu.edu.kz, orcid.org/0009-0005-7714-0292
SDU University, Kazakhstan

Kamila Orynbekova
MSc, Senior Lecturer, Computer Sciences Department
kamila.orynbekova@sdu.edu.kz, orcid.org/0000-0002-2182-2914
SDU University, Kazakhstan

Assem Talasbek
PhD, Assistant Professor, Computer Sciences Department
assem.talasbek@sdu.edu.kz, orcid.org/0000-0002-0944-1772
SDU University, Kazakhstan

Dauren Ayazbayev
MSc, Lecturer, Information Systems Department
dauren.ayazbayev@sdu.edu.kz, orcid.org/0000-0001-9973-2145
SDU University, Kazakhstan

Selcuk Cankurt
PhD, Assistant Professor, Department of Computer Engineering
s.cankurt@vistula.edu.pl, orcid.org/0000-0003-0581-1913
Vistula University, Poland

COMPARATIVE EFFECTIVENESS OF RULE-BASED AND MACHINE LEARNING METHODS IN SENTIMENT ANALYSIS OF KAZAKH LANGUAGE TEXTS

Abstract: Sentiment analysis is increasingly pivotal in natural language processing (NLP), crucial for deciphering public opinions across diverse sectors. This research conducts a comparative examination of rule-based and machine learning (ML) methods in sentiment analysis, specifically targeting the Kazakh language. Given the Kazakh language's limited exposure in computational linguistics, the study meticulously evaluates datasets from news articles, literature, and Amazon product reviews, aiming to compare the efficiency, adaptability, and overall performance of these distinct approaches.

Employing a detailed set of evaluation metrics such as accuracy, precision, recall, and computational efficiency, the study provides a comprehensive analysis of the strengths and limitations of rule-based techniques versus ML models like Logistic Regression, Multinomial Naive Bayes, Decision Trees, Random Forest, and XGBoost. The findings suggest rule-based methods excel in identifying nuanced emotional expressions within literary texts, while ML models demonstrate superior adaptability and robustness, particularly effective in handling the linguistic variations found in news and reviews.

Despite the strengths identified, the study also reveals significant limitations of the rule-based approach, especially in broader contexts beyond literary analysis. This highlights an imperative for future research to integrate sentiment dictionaries or domain-specific lexicons that cater to a wider array of linguistic styles, potentially enhancing sentiment analysis tools' applicability in Kazakh and similar less-studied languages.

DOI: 10.37943/17RHPH9724

This investigation contributes significantly to the sentiment analysis discourse, offering invaluable insights for both researchers and practitioners by elucidating the complexities of applying NLP technologies across diverse linguistic landscapes, thus advancing the understanding and methodologies of sentiment analysis in the Kazakh language context.

Keywords: sentiment analysis; machine learning; rule-based approach; Logistic Regression; Multinomial Naive Bayes.

Introduction

Sentiment analysis is the technique of determining a text's emotional tone and classifying it as positive, negative, or neutral [1-2]. Numerous studies have been carried out for a number of years on sentiment analysis across a variety of languages, including languages that are agglutinative [3-5]. On the other hand, academic studies concentrating on sentiment analysis in the Kazakh language are limited. In the field of natural language processing, sentiment analysis is extremely important because of how common social media and online platforms are becoming. The potential of rule-based methodology for sentiment analysis, particularly within the Kazakh language context, is still not well understood, despite the fact that several recent research have used traditional machine learning techniques in this field. Therefore, it becomes necessary to investigate the feasibility and effectiveness of rule-based approaches in order to further the field of sentiment analysis study in Kazakh.

Yergesh B. et al. [6] presents an ontological model and morphological rules-based approach that makes use of the recently created dictionary. The study of sentiment subtleties in Kazakh literature is improved by this organized framework. The development of an ontological model that offers a conceptual and methodical framework for sentiment rule extraction is a significant advance. The rule-based approach is effective, as evidenced by empirical assessment, which achieves 83% accuracy for short phrases. This supports the method's feasibility and possible applicability to sentiment analysis in Kazakh. The work not only closes a significant gap but also opens the door for customized sentiment analysis techniques that take into account Kazakh's particular linguistic traits.

Low-resource agglutinative languages have two fundamental obstacles that make text categorization difficult: the lack of labeled data in target domains and the morphological variety of derivations within language structures. To overcome these obstacles, a workable approach that makes use of pre-trained language models and refines them to provide useful feature extractors for subsequent text classification tasks is required. Li Z. et al. [7] presents AgglutiFiT, a low-resource agglutinative language model fine-tuning method, in answer to this demand. The process involves stem extraction and morphological analysis to provide a low-noise fine-tuning dataset. Then, using this dataset, the previously trained language model is refined. In addition, the paper proposes an attention-based fine-tuning strategy to enhance the pre-trained language model's recognition of relevant syntactic and semantic information. These improved attributes are then used for next text classification tasks.

Social media has revolutionized communication, providing a dynamic platform for interactions between people and brands. This unstructured dialogue space offers valuable insights into customer views, crucial for informed decision-making and brand management. The rise of big data has propelled sentiment analysis, particularly due to the abundance of sentiment-rich social media content. Kurian D. et al. [8] have developed methods tailored for effective sentiment analysis on large-scale datasets. They processed a Twitter dataset using Hadoop and evaluated performance based on accuracy and speed metrics, demonstrating the effectiveness of their approach. This highlights the utility of scalable frameworks like Hadoop and advances sentiment analysis research, showcasing the potential of computational frameworks in ex-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

tracting insights from vast and dynamic social media datasets, thereby shaping the evolving landscape of sentiment analysis in the era of big data.

Effective preprocessing is essential for accurate sentiment analysis of user comments, which often feature diverse languages and spelling errors. Niyazmetova K. et al. [9] focus on sentiment analysis of Tashkent restaurant reviews from Google Maps, emphasizing dataset preprocessing as pivotal for algorithm optimization. They employ logistic regression models for their robust statistical foundation in binary results, well-suited for sentiment categorization tasks. Specifically, they integrate preprocessing techniques like stemming, tailored for agglutinative languages characterized by complex word formation. Evaluation results demonstrate the system's effectiveness, underscoring the benefits of preprocessing—especially stemming—for agglutinative languages, as it standardizes text and enhances emotion discernment. The study highlights the critical role of preprocessing in sentiment analysis, particularly in diverse linguistic contexts like Tashkent restaurant reviews. It advances sentiment analysis techniques by leveraging logistic regression models and language-specific preprocessing, showcasing the reliability of their method in obtaining sentiment analysis insights.

Tussupov J. et al. [10] investigate word normalization algorithms and morphological models tailored to the unique linguistic features of Kazakh. They explore synthesizing normalized forms and identifying word bases in Kazakh, offering guidelines for handling non-dictionary concepts and nonexistent terms. This inclusive approach accommodates the dynamic nature of language, suitable for languages with evolving vocabularies. Their development of a Kazakh thesaurus for scientific and technical terms in information technology demonstrates the algorithm's flexibility and reliability, particularly in specialized fields. The study enhances textual data processing, especially in morphologically complex languages like Kazakh, through the formulation of normalization rules. By addressing Kazakh's linguistic complexities, the research advances linguistic tools and underscores the importance of specialized methods for unique morphological structures. The creation of a customized thesaurus exemplifies the algorithm's utility in specific language contexts.

Zhumabekova A.K. et al. [11] explore the intricacies of this translation process. They highlight the significant role of Russian as a mediator between English and Kazakh and discuss how linguistic and cultural differences between the languages influence the translation outcome. Through a comparative analysis of translated texts and examination of semantic shifts and stylistic adaptations, the authors elucidate the challenges faced by translators in preserving meaning and cultural nuances. They also propose strategies for mitigating these challenges, emphasizing the importance of linguistic competence and cross-cultural sensitivity. This study contributes to the understanding of indirect translation practices and provides insights for translators and researchers working in multilingual contexts.

Sentiment analysis, an essential part of natural language processing, is concerned with polarity-based text classification. Opinion mining is essential in this field, especially when it comes to figuring out what people think about movies or goods. It is impossible to overestimate the importance of user views on purchase decisions; for example, movie star ratings have a big influence on prospective audiences and shape their tastes. Similar to this, consumers' perceptions and decisions are greatly influenced by product reviews. The Naive Bayes classifier technique, a probabilistic method often employed in sentiment research, is utilized for classification in Surya, P. P. et al. [12] research. The Amazon product review dataset, which has about 600 entries, is the dataset that was taken from the UCI repository. Every record is analyzed using the Naïve Bayes technique, which ultimately yields a probabilistic matrix. Next, an accuracy matrix is used to evaluate the suggested strategy. This study explores the use of sentiment analysis and offers important new information on how well the Naive Bayes classifier distinguishes between different textual datasets' feelings.

DOI: 10.37943/17RHPH9724

The importance of customer reviews is examined in this review of the literature with reference to Shopee, an online store. Shopee's ongoing purchasing and selling activity generates a growing number of user evaluations, which are essential product references. Customers are allowed to express their opinions in these comments on the Shopee website, both favorable and bad. In order to deal with this situation, Hariguna, T. et al. [13] suggest using sentiment analysis, which combines a naïve Bayes classifier with the K-means clustering method. K-means is used to help classify comments into different groups, and the naive Bayes classifier is used to evaluate how well these groups are classified. According to the study, K-means clustering achieves an accuracy of 77.12% by identifying 116 negative and 37 positive comments in product evaluations. This accuracy is noteworthy since it is higher than the 56.86% accuracy that was attained using K-means, the Naive Bayes classifier, and manual data. The results highlight the frequency of unfavorable remarks, which are especially noticeable when it comes to the product «High Heels Women Knot Ribbon Ikat FX18» by Spatuafa. As a result, the study emphasizes how crucial sentiment analysis is to comprehending customer feedback and the possible influence of unfavorable remarks on how products are seen and assessed in the e-commerce space.

The literature review raises a pertinent question regarding the choice of methods for sentiment analysis, language processing, and translation tasks: Which methods are better to use? This question emerges from the diverse approaches adopted in the studies reviewed, ranging from rule-based methods to machine learning models. While some studies demonstrate the effectiveness of rule-based approaches, others showcase the potential of machine learning techniques. The comparison between these methodologies prompts further investigation into their respective strengths and weaknesses, considering factors such as accuracy, scalability, computational efficiency, and adaptability to different linguistic contexts. Ultimately, the choice of method may depend on the specific requirements of the task at hand, the availability of labeled data, computational resources, and the desired level of interpretability. Therefore, exploring the comparative advantages of rule-based and machine learning approaches becomes essential in determining the most suitable method for achieving optimal results in sentiment analysis, language processing, and translation tasks.

The aim of the study is to determine the comparative effectiveness of rule-based methods against machine learning algorithms for sentiment analysis.

Many studies have been conducted on sentiment analysis in natural languages, and notable progress has been made in well-known languages such as English, Turkish, and Russian. However, due to a lack of necessary tools and resources, the Kazakh language is still largely underdeveloped in this field. Proposed methodology incorporates three datasets: D. Chapaev's Sentiment Analysis Dataset, Serek's Agglutinative Language Sentiment Dataset, and Kaggle Amazon Sentiment labeled Dataset. Two methods were compared: Rule-based sentiment analysis, focusing on Kazakh language adjectives and Machine Learning models included Logistic Regression, Multinomial Naive Bayes, Decision Trees, Random Forest, and XGBoost.

This study underscores the importance of dataset-specific considerations in sentiment analysis tasks in Kazakh, highlighting the complementary strengths of both rule-based and ML approaches.

Methods and Materials

A. Dataset

In this research, three distinct datasets were utilized to conduct a comprehensive analysis. The details of each dataset are outlined below:

1. *D. Chapaev's Sentiment Analysis Dataset:*
   Source: Dauren Chapaev's sentiment analysis dataset on github [14].

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

Composition: The dataset comprises 20,014 sentences extracted from various news websites. Within this dataset, 5,993 sentences are classified as negative, 4,422 as positive, and the remaining 9,599 are labeled as neutral.

2. *Serek's Agglutinative Language Sentiment Dataset:*

Azamat Serek's and his coauthors' work on distributed sentiment analysis [15], utilizing sentences from the artistically instructive work «Bir atanyn balasy» (1973) by Mukhtar Makhauin.

This dataset consists of 732 sentences, categorized into 231 positive, 228 negative, and 273 neutral sentiments. The selection of sentences is based on the emotionally charged content of the book, focusing on emotions such as anger, fear, resentment, sadness, hopelessness (considered negative), and inspiration, anticipation, joy, euphoria, delight, interest, admiration, satisfaction (considered positive), with all other cases designated as neutral.

3. *Kaggle Amazon Sentiment labeled Dataset:*

Acquired from Kaggle, this dataset was originally in English [16] and was subsequently translated to Kazakh. Created for the research 'From Group to Individual Labels using Deep Features' [17].

The dataset contains sentences labeled with positive (score 1) or negative (score -1) sentiment. Initially sourced from three different websites/fields—imdb.com, amazon.com, and yelp.com—each website provides 500 positive and 500 negative sentences. For this research, sentences specifically selected from the amazon.com domain were utilized.

For the rule-based sentiment analysis component of this study, a set of Kazakh language adjectives was compiled from Sozdik Qor [18]. Sozdik Qor is a comprehensive platform that facilitates access to words and stable phrases from diverse industry dictionaries and encyclopedias. It encompasses ancient words in the Kazakh language, input words, and the meanings of newly emerging technological terms in the realm of regional and information technologies. The portal's search engine allows users to explore word definitions, synonyms, antonyms, homonyms, and their occurrences in phraseological phrases or within sentences, all conveniently presented on a single page.

Specifically, the focus was on gathering adjectives in the Kazakh language from this platform. However, these adjectives were initially unlabeled. Sentiment labels in the range of -1 to 1 were manually assigned to address this issue. The final dataset contains 5,539 adjectives, with 1,902 tagged as near to positive sentiment, 1,657 as close to negative sentiment, and 1,980 as close to neutral sentiment.

B. Rule-Based Sentiment analysis

The following principles have been developed in order to formalize the rules guiding the evaluation of sentiment in Kazakh language phrases:

1. *Adjective Tonality*: Let T_a stand for an adjective's tone. The tone of the adjective determines the tone of the phrase (T_p) directly: $T_a = T_p$

Example: jaqsy adam –> Positive (tonality is 1)

Explanation:

→ jaqsy – positive adjective with a sentiment label of 1.

→ adam – a noun

2. *Adverb Influence*: Let T_a represent the adjective's initial tonality. Let T_p to be the tonality of the phrase. If an adverb (A) precedes the adjective: $T_p = 2 \cdot T_a$

Example: ote jaqsy adam –> Positive (tonality is 2)

Explanation:

→ ote – an adverb

DOI: 10.37943/17RHPH9724

→ jaqsy – positive adjective with a label 1
→ adam – a noun

3. *Negation Impact:* Let T_a be the original tonality of the adjective. Let T_p be the tonality of the phrase. If a negation (N) appears after the adjective: $T_p = -T_a$

Example: jaqsy adam emes –> Negative (tonality is -1)
Explanation:
→ jaqsy – a positive adjective with a label 1
→ adam – a noun
→ emes – a negation

Cumulative Adjective Tonality: When a sentence has more than one adjective, the overall tone of the phrase is derived from the tonal aggregate of all the adjectives. For instance, the combined tonality of 'ademi +0.7' and 'ote jaman -2' in a phrase is determined to be 0.7 - 2 = -1.3, indicating a Negative (-) attitude.

These principles provide the fundamental components for a deliberate approach to analyzing sentiments in Kazakh. A richer, more complex understanding of the emotions conveyed in the text can be gained by exploring the complex relationship between adjectives, adverbs, and negations.

C. Data Processing and Machine Learning Approaches

- *Data Preprocessing:*

To ensure that the data flowing into the models was in the best possible state, thorough cleaning of the text was conducted before the utilization of the machine learning models. Several processes were involved in this cleanup:

→ *Lowercasing:* To maintain consistency and get rid of case-related differences, convert all text to lowercase.

→ *HTML Tag Removal:* Remove HTML tags in order to eliminate any unnecessary content.

→ *Special Character, Number, and Punctuation Removal:* For a writing that is clearer and more concentrated, remove punctuation, special characters, and number values.

→ *Stopword Removal:* To cut down on noise and highlight words that provide meaning, omit frequently used stop words.

→ *Stemming*: Utilize stemming to break down words into their most basic form, which will enable more efficient feature extraction.

- *Vectorization Techniques:*

The processed text data was vectorized using two widely-used methods to provide numerical characteristics for machine learning models:

1. *TF-IDF Vectorizer:* The Frequency-Inverse Document Frequency (TF-IDF) method is employed to measure a term's significance in relation to a group of documents. It takes into account the inverted document frequency over the whole dataset in addition to the term's frequency in a document. The selection of TF-IDF stems from its capacity to draw attention to important terms in a document while minimizing common terms. This strengthens the sentiment analysis features' ability to discriminate [19].

Mathematical Formulation:

For a term t in a document d, TF-IDF is calculated as in (1):

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

where IDF is the inverse document frequency and TF is the term frequency. This procedure minimizes common terms while highlighting the significance of phrases in a text.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

2. *Count Vectorizer:* The text is transformed into a sparse matrix that displays the number of each phrase in the document using count vectorization. Count vectorization is selected due to its simplicity of usage and efficacy in figuring out the word frequency in a document. It provides a straightforward representation of word occurrences, which is helpful for some types of sentiment analysis tasks [20].

Mathematical Formulation:

The count of each term in a document is represented as a matrix element (2):

$$Count(t,d) = number\ of\ occurrences\ of\ term\ t\ in\ document\ d \tag{2}$$

that produces a sparse matrix that highlights word occurrences.

- *Machine Learning Models:*

The vectorization algorithms were used in combination with two commonly used classifiers:

1. *Logistic Regression:* A linear model that works well for binary classification applications is Logistic Regression. It is useful for sentiment analysis tasks and predicts the likelihood that a sample belongs to a specific class. The simplicity, effectiveness, and interpretability of logistic regression make it a preferred option. In tasks involving text categorization, it frequently does well [21].

Mathematical Formulation:

Logistic regression predicts the probability that a sample belongs to a specific class using the sigmoid function (3):

$$P(y=1) = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+\ +b_nx_n)}} \tag{3}$$

where $b_0, b_1,..., b_n$ are coefficients and $x_0, x_1,..., x_n$ are features.

2. *Multinomial Naive Bayes:* The probabilistic classifier Multinomial Naive Bayes is based on the Bayes theorem. It works especially well for text classification jobs since it makes the assumption that the characteristics are conditionally independent given the class. The selection of Naive Bayes is based on its efficacy and efficiency in managing sparse and high-dimensional data. Tasks involving word frequencies in documents are a good fit for it [22].

Mathematical Formulation:

Naive Bayes calculates the probability of a document belonging to a class given its features (4):

$$P(class \mid features) = \frac{P(features \mid class) \times P(class)}{P(features)} \tag{4}$$

The assumption of conditional independence simplifies the computation.

3. *Decision Trees:* A versatile model for classification and regression that segments the dataset into branches, making it highly interpretable. Decision Trees can be applied effectively in categorizing textual data based on feature thresholds.

Mathematical Formulation:

Decision trees use criteria such as Gini impurity or information gain to split data, aiming to create subsets that are as pure as possible at each node.

4. *Random Forest:* An ensemble of decision trees that improves prediction accuracy and controls overfitting. It combines the predictions from multiple decision trees to produce a more accurate and stable prediction.

DOI: 10.37943/17RHPH9724

Mathematical Formulation:

For classification tasks, the Random Forest model takes the majority vote from its decision trees. In regression tasks, it averages the outputs. The randomness injected into the model building process helps in reducing overfitting.

5. *XGBoost:* Stands for eXtreme Gradient Boosting, a scalable and efficient implementation of gradient boosting. It is known for its performance and speed in data competitions. XGBoost is particularly useful in handling structured data for both classification and regression tasks.

Mathematical Formulation:

XGBoost optimizes a loss function by iteratively adding trees that predict the residuals or errors of prior trees, with an objective to minimize these errors across all predictions.

The inclusion of Decision Trees, Random Forest, and XGBoost alongside Logistic Regression and Multinomial Naive Bayes enriches the analysis by covering a broad spectrum of machine learning approaches, ranging from simple to complex models. This diverse set of models was selected for their complementary strengths in handling different aspects of sentiment analysis, with considerations for factors such as interpretability, efficiency, and effectiveness in managing high-dimensional data.

- *Data splitting:*

The datasets were each separated into training and testing sets to facilitate the evaluation of the proposed models' performance across multiple data sources. Specifically, for each dataset, 80% of the data was allocated for training purposes, while the remaining 20% was reserved for testing. This consistent partitioning approach ensures a uniform evaluation framework for all three datasets. Prior to model training and testing, extensive data preprocessing procedures were implemented to ensure that the data entering the models was in the best possible condition. This preparation involved several cleaning processes tailored to each of the datasets.

- *Computational Environment:*

Google Colab, a cloud-based platform for machine learning and Python programming, was used for the analysis. An easy and scalable environment for running code is offered by Google Colab, which is especially helpful for resource-intensive activities like machine learning. This platform ensures that the method described may be executed with ease and that collaboration is made easier.

Results and Discussion

The F1 score for the applied model on each individual dataset is used to illustrate the results, giving a thorough picture of how different vectorizers, classifiers, and the rule-based method performed (Tables 1, 2, and 3).

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

Table 1. Dauren Chapaev's Sentiment Analysis Dataset results

#	Method	Accuracy	Precision	Recall	F1 Score
1	TF-IDF + Logistic Regression	0.78	0.79	0.78	0.79
2	TF-IDF + Multinomial Naive Bayes	0.80	0.80	0.80	0.80
3	Count + Logistic Regression	0.78	0.78	0.78	0.78
4	Count + Multinomial Naive Bayes	0.81	0.81	0.81	0.81
5	Decision Tree	0.66	0.65	0.66	0.65
6	Random Forest	0.75	0.75	0.75	0.75
7	XGBoost	0.72	0.73	0.72	0.72
8	Rule-Based	0.39	0.40	0.40	0.40

Table 2. Azamat Serek's Agglutinative Language Sentiment Dataset results

#	Method	Accuracy	Precision	Recall	F1 Score
1	TF-IDF + Logistic Regression	0.59	0.61	0.59	0.60
2	TF-IDF + Multinomial Naive Bayes	0.58	0.58	0.57	0.57
3	Count + Logistic Regression	0.58	0.60	0.58	0.59
4	Count + Multinomial Naive Bayes	0.59	0.60	0.59	0.59
5	Decision Tree	0.61	0.63	0.61	0.62
6	Random Forest	0.64	0.68	0.64	0.66
7	XGBoost	0.61	0.62	0.60	0.61
8	Rule-Based	0.79	0.77	0.77	0.77

Table 3. Amazon Sentiment Labeled Sentences Dataset

#	Method	Accuracy	Precision	Recall	F1 Score
1	TF-IDF + Logistic Regression	0.75	0.75	0.75	0.75
2	TF-IDF + Multinomial Naive Bayes	0.74	0.74	0.73	0.73
3	Count + Logistic Regression	0.74	0.74	0.74	0.74
4	Count + Multinomial Naive Bayes	0.75	0.76	0.75	0.75
5	Decision Tree	0.72	0.72	0.72	0.72
6	Random Forest	0.76	0.76	0.76	0.76
7	XGBoost	0.72	0.72	0.72	0.72
8	Rule-Based	0.31	0.33	0.33	0.33

Dauren Chapaev's Sentiment Analysis Dataset Results:

Analysis of Dauren Chapaev's dataset reveals that Logistic Regression and Multinomial Naive Bayes, particularly when combined with TF-IDF vectorization, perform commendably with F1 scores reaching up to 0.80. Decision Trees, Random Forest, and XGBoost show a range of effectiveness, with Random Forest presenting a notable performance. The Rule-Based approach, while not as effective in this dataset, highlights the diverse linguistic challenges present.

Azamat Serek's Agglutinative Language Sentiment Dataset Results:

Within Azamat Serek's dataset, the ensemble models, especially Random Forest, and Decision Trees display an enhanced ability to handle the dataset's linguistic complexities. This per-

DOI: 10.37943/17RHPH9724

formance underscores the potential of these models in analyzing texts with nuanced language structures. The Rule-Based method's continued success emphasizes its capacity to grasp the subtleties of literary expressions effectively.

Amazon Sentiment Labeled Sentences Dataset Results:

The Amazon Sentiment Labeled Sentences Dataset shows the strengths of ensemble methods, with Random Forest leading in performance. This suggests their robustness in adapting to the varied linguistic patterns typical of online reviews. Decision Trees and XGBoost also offer strong alternatives, highlighting the diversity of effective approaches available for sentiment analysis.

Comparative Analysis:

- *Vectorization Techniques:* The consistent performance of TF-IDF and Count Vectorizer across all datasets underscores their reliability in capturing textual features, regardless of the linguistic context.
- *Classifier Performance:* Logistic Regression and Multinomial Naive Bayes demonstrate versatility across the datasets, suggesting their general applicability to sentiment analysis. The decision on which model to use may depend on specific project needs related to interpretability and computational demands.
- *Dataset-Specific Observations:* The models exhibit a capacity to adapt to the varied language patterns encountered, from the broad scope of Dauren Chapaev's dataset to the specific challenges posed by Azamat Serek's literary work. The Rule-Based method's performance in the latter case points to its effectiveness in contexts where capturing emotional depth is crucial.
- Model Selection: The study illustrates the efficacy of both machine learning models and Rule-Based approaches in their respective domains. Ensemble methods, in particular, show promise for their adaptability and robust performance across different text types.

Considerations for Sentiment Analysis in Kazakh:

The findings emphasize the critical role of model and preprocessing technique selection in sentiment analysis. While ensemble methods like Random Forest prove to be highly effective across various text types, the value of Rule-Based approaches in capturing nuanced emotional content should not be overlooked. The diversity of the Kazakh language, with its range of literary and journalistic expressions, necessitates flexible and nuanced analysis strategies.

This revised discussion integrates Decision Trees, Random Forest, and XGBoost as foundational elements of the study, providing a nuanced comparison of their performance against traditional models and a Rule-Based approach across diverse datasets.

Conclusion

The study's findings provide factual support for the continuing debates in the area and mark a substantial advancement in our knowledge of sentiment analysis techniques. Additionally, professionals looking to utilize sentiment analysis methodologies might benefit from the practical insights gained from this comparison investigation.

The findings conclusively show that rule-based methods excel in identifying the nuanced emotional content typical of literary works. Conversely, machine learning (ML) models exhibit remarkable flexibility in addressing the language variations often encountered in news articles and reviews. This distinction underscores the effectiveness of rule-based strategies in specific emotional contexts within literature, and the adaptability of ML models to a wide range of linguistic patterns prevalent in news and review datasets.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

Nevertheless, the analysis revealed notable limitations in the performance of the rule-based model, particularly evident with the Sentiment Labeled Sentences Data Set. This highlights the challenges in accurately capturing a wide spectrum of sentiment expressions and language nuances, particularly in contexts beyond literary texts. Such findings underscore the urgent need for further research to enhance and develop current sentiment analysis techniques.

Prospective future paths for study might involve integrating sentiment dictionaries or domain-specific lexicons that are designed to fit the variety of language styles that are common outside of literary fields. It is possible that these efforts will improve sentiment analysis models' ability to function in the complex linguistic environment of Kazakh, making them more flexible to a wider range of sentiment expressions and linguistic subtleties in non-literary instances.

References

[1] Mehta, P., & Pandya, S. (2020), A review on sentiment analysis methodologies, practices and applications. *Int. J. Sci. Technol. Res.*, 9 (2), 601–609.
[2] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.*, 55 (7), 5731–5780.
[3] Parlar, T., Ozel, S., & Song, F. (2019). Analysis of data pre-processing methods for sentiment analysis of reviews. *Comput. Sci.*, 20, 123-141.
[4] Özçift, A. (2022). FastText Word Embedding Model in Aspect-Level Sentiment Analysis of Airline Customer Reviews for Agglutinative Languages: A Case Study for Turkish. *International Conference on Artificial Intelligence and Applied Mathematics in Engineering,* Springer, 691–702.
[5] Matlatipov, S., Rahimboeva, H., Rajabov, J., & Kuriyozov, E. (2022). Uzbek sentiment analysis based on local restaurant reviews. *ArXiv Prepr,* 220515930.
[6] Yergesh, B., Bekmanova, G., Sharipbay, A. & Yergesh, M. (2017). Ontology-based sentiment analysis of kazakh sentences. *Computational Science and Its Applications–ICCSA 2017: 17th International Conference*, *Springer*, 669–677.
[7] Li Z., Li X., Sheng J., & Slamu, W. (2020), AgglutiFiT: efficient low-resource agglutinative language model fine-tuning. *IEEE Access*, 8, 148489–148499.
[8] Kurian, D. D. M. K., Vishnupriya, S., Ramesh, R., Divya, G., & Divya, D. (2015). Big data sentiment analysis using hadoop. *Int. J. Innov. Res. Sci. Technol.*, 1 (11), 92–96.
[9] Niyazmetova, K., Raximov, K., Anvarova, D., & Bekjanov, R. (2023). Formation of a Database For Sentiment Analysis of Texts in the Uzbek Language. *Sci. Innov.*, 2 (C11), 20–23.
[10] Tussupov, J., Sambetbayeva, M., Idrisova, I., & Yerimbetova, A. (2015). Development and implementation of a morphological model of kazakh language. *Eurasian J. Math. Comput. Appl.*, 3 (3), 69–79.
[11] Zhumabekova, A. K., & Mirzoyeva, L. Y. (2016), Peculiarities of indirect translation from English into Kazakh via Russian language. *TOJET*. pp. 189-194
[12] Surya, P. P., & Subbulakshmi, B. (2019). Sentimental analysis using Naive Bayes classifier. *2019 International conference on vision towards emerging trends in communication and networking (ViTECoN)*, 1-5.
[13] Hariguna, T., Baihaqi, W. M., & Nurwanti, A. (2019). Sentiment analysis of product reviews as a customer recommendation using the naive Bayes classifier algorithm. *International Journal of Informatics and Information Systems*, 2(2), 48-55.
[14] Open Access Kazakh News Sentiment Labeled Dataset. https://github.com/chapayevdauren/sentiment-analysis-for-kz/blob/master/data/sample.csv.
[15] Serek, A., Issabek, A., & Bogdanchikov, A. (2019). Distributed sentiment analysis of an agglutinative language via Spark by applying machine learning methods. *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, *IEEE*, 1–4.
[16] Sentiment Labeled Sentences Data Set of Product Reviews. https://www.kaggle.com/datasets/marklvl/sentiment-labelled-sentences-data-set?rvi=1.

DOI: 10.37943/17RHPH9724

[17] Kotzias, D., Denil, M., De Freitas, N., & Smyth, P. (2015). From group to individual labels using deep features. *21th ACM SIGKDD international conference on knowledge discovery and data mining*, 597–606.
[18] Sozdikqor.kz: Comprehensive Kazakh Language Portal for Diverse Word Meanings and Phrases. https://sozdikqor.kz/
[19] Abubakar, H. D., Umar, M., & Bakale, M. A. (2022). Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec. *SLU Journal of Science and Technology*, 4(1 & 2), 27-33.
[20] Goyal, R. (2021). Evaluation of rule-based, CountVectorizer, and Word2Vec machine learning models for tweet analysis to improve disaster relief. *2021 IEEE Global Humanitarian Technology Conference (GHTC),* 16-19.
[21] Saad, S. E., & Yang, J. (2019). Twitter sentiment analysis based on ordinal regression. *IEEE Access*, 7, 163677-163685.
[22] Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. *IJCSNS Int. J. Comput. Sci. Netw. Secur*, 19(3), 62.