Ministry of Science and Higher Education of the Republic of Kazakhstan
SDU University

Mukhtar Amirkumar

Sentiment Analysis of Texts in the Kazakh Language Using Machine Learning and Rule-based Methods

THESIS
Presented in Partial Fulfilment for the

Degree of Master of Technical Science in Computer Science
(degree code: 7M06102)
Department of Computer Science
Faculty of Engineering and Natural Sciences

Supervisor: **Kamila Orynbekova**
Kaskelen, June 2024

SDU University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty of Engineering and Natural Sciences

Assistant Professor, PhD Akhmedov Ramis

_____

« _____» _____ 2024

Topic of the thesis:

Sentiment Analysis of Texts in the Kazakh Language Using Machine Learning and Rule-based Methods

---

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science" , SDU University

Head of Department	Zhanar Mukash
Academic Supervisor	Kamila Orynbekova
Master Student	Mukhtar Amirkumar

Kaskelen, 2024

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Mukhtar Amirkumar

June 2024

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Kamila Orynbekova, for her invaluable guidance, unwavering support, and insightful feedback throughout the entire process of completing this thesis. Her expertise and encouragement have been instrumental in shaping the direction of my research and in helping me navigate through the challenges.

I am also immensely thankful to Dauren Ayazbayev, Merarslan Meraliyev for their assistance, encouragement, and valuable contributions at various stages of this research. Their input and constructive criticism have been crucial in refining the ideas presented in this thesis.

Finally, I would like to express my gratitude to all the participants who generously shared their time and insights, without whom this research would not have been possible.

Dedication

This thesis is dedicated to: Me, whose unwavering support and belief in me have been my guiding light throughout this endeavor.

Abstract

This thesis explores the effectiveness of rule-based and machine learning (ML) approaches in the sentiment analysis of Kazakh language texts across diverse domains in natural language processing (NLP). Recognizing the complexities and limited resources available for Kazakh, this research incorporates edit distance algorithms to refine rule-based methods, enhancing their accuracy and robustness. The study undertakes a comparative exploration between these methodologies, aiming to comprehensively assess and compare their performance, effectiveness, and adaptability.

Diverse datasets, including daily news, Kazakh literary texts, and Amazon product reviews, have been utilized to delineate the strengths and weaknesses of both rule-based and machine learning approaches. By employing metrics such as accuracy, precision, recall, and computational efficiency, a detailed evaluation has been conducted. It has been found that subtle emotional nuances in literary texts are effectively captured by rule-based methods, while the adaptability of ML models to the varied linguistic elements commonly found in news and consumer reviews is confirmed.

However, significant limitations in the rule-based approach have been exposed, revealing the superior scalability and adaptability of ML models when applied across different datasets. The extension of edit distance application to comprehensive sentence analyses and the integration with ML techniques to develop hybrid models are anticipated as future research directions. These advancements are expected to improve both the precision and the applicability of sentiment analysis tools, benefiting not only Kazakh but also other languages with complex linguistic structures. This work has advanced NLP capabilities for underrepresented languages and has promoted the development of more inclusive language processing tools.

Аңдатпа

Бұл дипломдық жұмыс табиғи тілді өңдеудің (NLP) әртүрлі домендері бойынша қазақ тіліндегі мәтіндерді сезімдік талдаудағы ережеге негізделген және машиналық оқыту (ML) тәсілдерінің тиімділігін зерттейді. Қазақ тілі үшін қолжетімді қиыншылықтар мен шектеулі ресурстарды ескере отырып, бұл зерттеу ережелерге негізделген әдістерді нақтылау, олардың дәлдігі мен сенімділігін арттыру үшін өңдеу қашықтағы алгоритмдерін біріктіреді. Зерттеу олардың өнімділігін, тиімділігін және бейімделгіштігін жан-жақты бағалауға және салыстыруға бағытталған осы әдістемелер арасында салыстырмалы зерттеу жүргізеді.

Ережеге негізделген және машиналық оқыту тәсілдерінің күшті және әлсіз жақтарын анықтау үшін күнделікті жаңалықтарды, қазақша әдеби мәтіндерді және Amazon өніміне шолуларды қоса алғанда, әртүрлі деректер жинақтары пайдаланылды. Дәлдік, еске түсіру және есептеу тиімділігі сияқты көрсеткіштерді қолдану арқылы егжей-тегжейлі бағалау жүргізілді. Көркем мәтіндердегі нәзік эмоционалды реңктер ережеге негізделген әдістермен тиімді түрде түсірілетіні анықталды, ал ML үлгілерінің жаңалықтар мен тұтынушылар шолуларында жиі кездесетін әртүрлі тілдік элементтерге бейімделуі расталады.

Дегенмен, ережеге негізделген тәсілдегі елеулі шектеулер әр түрлі деректер жиындарында қолданылған кезде ML үлгілерінің жоғары масштабталатындығы мен бейімделгіштігін ашып көрсетті. Болашақ зерттеу бағыттары ретінде гибридті модельдерді әзірлеу үшін ML әдістерімен интеграциялау және сөйлемдерді кешенді талдауға дейінгі қашықтықты өңдеу қолданбасын кеңейту күтілуде. Бұл жетістіктер тек қазақ тілінде ғана емес, сонымен қатар күрделі тілдік құрылымы бар басқа тілдерде де пайда әкелетін сезімді талдау құралдарының дәлдігін де, қолдану мүмкіндігін де жақсартады деп күтілуде. Бұл жұмыс жеткіліксіз ұсынылған тілдер үшін NLP мүмкіндіктерін жетілдірді және анағұрлым инклюзивті тілді өңдеу құралдарының дамуына ықпал етті.

Аннотация

В этой диссертации исследуется эффективность основанного на правилах и машинного обучения (ML). подходы к анализу тональности текстов на казахском языке в различных областях обработки естественного языка (НЛП). Признавая сложности и ограниченные ресурсы, доступные для казахского языка, это исследование включает в себя дистанционное редактирование горитмы для совершенствования методов, основанных на правилах, повышая их точность и надежность. В исследовании проводится сравнительное исследование этих методологий. стремясь всесторонне оценить и сравнить их результативность, эффективность, и адаптивность.

Разнообразные наборы данных, включая ежедневные новости, казахские литературные тексты и обзоры продуктов Amazon, были использованы для определения сильных и слабых сторон обеих компаний. подходы, основанные на правилах и машинном обучении. С помощью таких показателей, как точность, точность, полнота и вычислительная эффективность, была проведена детальная оценка. проведенный. Было обнаружено, что тонкие эмоциональные нюансы в литературных текстах эффективно улавливаются методами, основанными на правилах, а адаптируемость моделей МО к различным языковым элементам, обычно встречающимся в новостях и отзывах потребителей, подтверждена.

Однако были выявлены существенные ограничения подхода, основанного на правилах, что свидетельствует о превосходной масштабируемости и адаптируемости моделей МО при их применении в различных сферах. разные наборы данных. В качестве будущих направлений исследований ожидается расширение применения дистанционного редактирования до комплексного анализа предложений и интеграция с методами машинного обучения для разработки гибридных моделей. Ожидается, что эти достижения повысят как точность, так и применимость инструментов анализа настроений. подходит не только казахскому, но и другим языкам со сложной языковой структурой. Эта работа расширила возможности НЛП для недостаточно представленных языков и имеет способствовал развитию более инклюзивных инструментов языковой обработки.

Abbreviations

ED - Edit Distance
KZ - Kazakh Language
LR - Logistic Regression
ML - Machine Learning
NLP - Natural Language Processing
RB - Rule-Based
RF - Random Forest
SA - Sentiment Analysis
TF-IDF - Term Frequency-Inverse Document Frequency
XGBoost - Extreme Gradient Boosting

Table of Contents

Chapter 1

Introduction

1.1 Introduction

In the domain of natural language processing (NLP), sentiment analysis has become an essential tool for interpreting public sentiment across various media and industries. This thesis is particularly focused on the unique challenges posed by the Kazakh language, whose complex linguistic structure and scarce representation in computational linguistics research present unique opportunities for innovation. The primary aim of this research is to evaluate and compare the effectiveness of rule-based methods versus machine learning (ML) approaches in the sentiment analysis of Kazakh language texts, incorporating a preliminary exploration of edit distance algorithms to enhance the rule-based techniques.

Aim

The primary aim of this thesis is to assess and compare the effectiveness of rule-based and machine learning methods in the sentiment analysis of texts in the Kazakh language, while also exploring the potential of edit distance algorithms to enhance the accuracy of rule-based techniques. This research seeks to identify which approach, or combination of approaches, provides the most reliable and efficient means of analyzing sentiment in a context where linguistic resources are scarce and the language structure is complex.

Through this comparative analysis, the study aims to:

1. Determine the strengths and weaknesses of rule-based versus machine learning approaches in sentiment analysis specific to the Kazakh language.
2. Enhance the rule-based method by incorporating edit distance algorithms to correct typographical errors and normalize variations in the text, thereby improving the method's robustness and accuracy.
3. Contribute to the field of computational linguistics by providing insights and practical approaches that can be applied to other underrepresented languages, thus bridging the gap in NLP tools and resources available for such languages.

Ultimately, this thesis will not only advance the understanding of sentiment analysis in the Kazakh language but also explore innovative ways to improve the

precision and applicability of NLP techniques in similar linguistic contexts.

Relevance of the Research

The relevance of this research lies in its focus on the Kazakh language, which is notably underrepresented in the field of natural language processing (NLP). Kazakh, like many other less-studied languages, presents unique challenges due to its agglutinative structure and sparse linguistic resources. These challenges have historically limited the development and application of advanced computational tools and techniques that are readily available for more widely spoken languages.

This study is particularly relevant because it addresses these limitations by:

1. Adapting and testing NLP tools specifically for the Kazakh language, thereby expanding the technological reach and applicability of sentiment analysis. This adaptation is crucial for developing local digital resources and tools, which can significantly benefit cultural preservation, educational initiatives, and local business landscapes.
2. Exploring the integration of edit distance algorithms in rule-based sentiment analysis methods. This exploration is aimed at enhancing the accuracy and reliability of these methods, which are particularly vulnerable to errors due to the morphological complexity of Kazakh. Improving error handling can lead to more robust NLP applications, which are essential in areas such as social media monitoring, market analysis, and public opinion assessment.
3. Setting a precedent for similar research in other underrepresented languages, which can leverage the findings and methodologies developed through this study. This has the potential to catalyze further research and development in the field of NLP, expanding the benefits of technology across linguistic borders.

By addressing these points, the research not only contributes to filling a critical gap in the NLP field but also enhances the linguistic inclusivity of sentiment analysis tools, thereby supporting more global and culturally diverse applications of technology.

Significance of the Study

This research significantly advances the field of natural language processing (NLP) for the Kazakh language, an underrepresented linguistic area in computational linguistics. The integration of edit distance algorithms into rule-based sentiment analysis methods introduces a novel approach to improve accuracy and robustness, addressing Kazakh's unique morphological complexities. This has practical implications across various domains:

- Technological Inclusivity: Enhances the linguistic resources available for Kazakh, promoting technological inclusivity and reducing the digital divide.
- Hybrid NLP Models: Paves the way for hybrid NLP models that combine rule-based and machine learning approaches, offering potential improvements in sentiment analysis accuracy and adaptability.
- Business and Public Policy: Improved sentiment analysis tools can aid busi-

nesses and government agencies in better understanding and engaging with the Kazakh-speaking community.

- Educational and Cultural Benefits: Supports educational initiatives and cultural preservation by improving digital access to Kazakh language resources.

Overall, this study not only enriches the academic and practical understanding of sentiment analysis in Kazakh but also sets a precedent for similar advancements in other underrepresented languages.

Research Objectives

This thesis sets forth several specific objectives designed to support its primary aim of evaluating and comparing rule-based and machine learning methods for sentiment analysis of Kazakh texts, and to explore the incorporation of edit distance algorithms. These objectives are structured to provide a comprehensive analysis of the methods' effectiveness, identify potential enhancements, and contribute broadly to the field of computational linguistics for underrepresented languages:

1. Evaluate the Accuracy and Efficiency of Rule-Based and Machine Learning Methods: Implement and assess both rule-based and machine learning sentiment analysis models across diverse Kazakh text datasets. The evaluation will focus on key performance metrics including accuracy, precision, recall, and computational efficiency. This objective will help determine which method is more effective in handling the linguistic characteristics unique to Kazakh.
2. Explore the Integration of Edit Distance Algorithms: Investigate how the inclusion of edit distance algorithms can improve the performance of rule-based methods. This will involve modifying these methods to incorporate error correction and normalization processes, aimed at enhancing the handling of typographical and morphological variations in Kazakh texts.
3. Assess Improvements Brought by Edit Distance Algorithms: Quantitatively measure the impact of edit distance enhancements on the rule-based methods' accuracy and robustness. This will provide empirical evidence of whether these preliminary explorations yield significant improvements over traditional approaches.
4. Compare and Contrast the Adaptability of the Analyzed Methods: Beyond performance metrics, evaluate how well each method adapts to the complexities of the Kazakh language. This involves analyzing their scalability, ease of integration into existing frameworks, and their flexibility in accommodating the language's morphological richness.
5. Propose Methodological Enhancements Based on Comparative Analysis: Based on the outcomes of the evaluations, propose refinements and optimizations for both rule-based and machine learning methods. This objective aims to develop recommendations that can enhance sentiment analysis tools not only for Kazakh but also for other similar languages.
6. Contribute to the Field of NLP for Underrepresented Languages: Synthesize the findings from this research to offer insights and practical approaches that can be utilized in the development of NLP applications for other underrep-

resented languages. This will help bridge the gap in technology availability and encourage more inclusive research in the field.

Each objective is intended to build upon the findings of the previous one, creating a layered investigation that not only assesses but also seeks to improve how sentiment analysis can be conducted on Kazakh texts. This structured approach ensures that each phase of the research is purposeful and contributes directly to the thesis's overall goals.

Hypothesis

The hypothesis of this study posits that "the integration of edit distance algorithms into rule-based sentiment analysis methods will significantly enhance the accuracy and robustness of these methods when applied to the Kazakh language." This hypothesis is based on the assumption that by correcting typographical errors and normalizing lexical variations, the refined rule-based methods will outperform traditional rule-based approaches and potentially match or exceed the performance of machine learning models in specific scenarios.

Research Question

The central question of this research is: "How do rule-based methods, enhanced by a preliminary application of edit distance algorithms, compare to machine learning approaches in conducting accurate sentiment analysis of Kazakh language texts across various domains?"

This question seeks to explore not only the comparative effectiveness of these methodologies but also to evaluate the specific contributions of edit distance algorithms to the rule-based analysis process.

Research Novelty

This research introduces several novel elements to the field of sentiment analysis, specifically targeting the Kazakh language. These innovations not only enhance the capabilities of existing sentiment analysis methods but also pave the way for more accurate and effective NLP tools for underrepresented languages.

Integration of Edit Distance Algorithms in Rule-Based Methods

The primary innovation of this thesis lies in the integration of edit distance algorithms into traditional rule-based sentiment analysis methods. This approach is particularly novel because:

- Error Correction and Normalization: By incorporating edit distance algorithms, the research aims to improve the accuracy of sentiment detection in Kazakh texts by automatically correcting typographical errors and normalizing lexical variations. This is crucial for Kazakh, where morphological complexity can lead to frequent typographical errors that traditional rule-based methods might not handle effectively.
- Adaptability to Agglutinative Languages: The edit distance feature is tai-

lored to address the specific challenges of agglutinative languages like Kazakh, which involve complex word formations. This adaptation makes it one of the first attempts to systematically apply such a linguistic tool in the sentiment analysis of Kazakh, setting a precedent for other similar languages.

Comparative Analysis of Rule-Based and Machine Learning Approaches Another innovative aspect of this study is the comprehensive comparative analysis between enhanced rule-based methods and standard machine learning models. This comparison is unique due to:

- Broad Application Spectrum: It evaluates these methods across various types of text, from news articles and literary works to online product reviews, providing a broad understanding of their effectiveness in different contexts.
- Insights into Machine Learning and Rule-Based Synergies: By comparing these approaches, the research not only identifies which is more effective but also explores how they might be combined to leverage the strengths of both. This could lead to the development of hybrid models that are both robust and adaptable.

Methodological Contributions

This thesis also makes methodological contributions that are novel in the context of Kazakh sentiment analysis:

- Bespoke Metrics and Evaluation: The use of specific metrics designed to evaluate the performance of sentiment analysis tools in handling the Kazakh language's unique features is another novelty. These metrics consider linguistic nuances that generic sentiment analysis metrics might overlook.
- Resource Creation: The development of a dataset annotated specifically for sentiment analysis in Kazakh, including the establishment of a lexicon of sentiment-laden Kazakh words and phrases, contributes a valuable resource to the field.

Broader Implications

The methodologies and findings of this thesis are expected to have broad implications:

- Expansion to Other Languages: The approaches developed could be adapted for other underrepresented languages with similar linguistic features, providing a model for extending advanced NLP tools to a wider array of languages.
- Enhancement of Linguistic Tools: By demonstrating the effectiveness of integrating edit distance algorithms into sentiment analysis, this research could inspire further innovations in linguistic processing tools, potentially impacting other areas of NLP such as machine translation and text summarization.

Chapter 2

Background and Literature Review

2.1 Definition and Importance

Sentiment analysis, also known as opinion mining, is a technique used to analyze and interpret people's opinions, emotions, and attitudes within text data. It involves the use of natural language processing (NLP), text analysis, and computational linguistics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis focuses on determining the sentiment polarity (positive, negative, or neutral) of textual data, which can be derived from various sources such as social media posts, product reviews, news articles, and political speeches. This analysis can be conducted at different levels, including document level (assessing the overall sentiment of an entire document), sentence level (evaluating the sentiment of individual sentences), and aspect level (determining the sentiment towards specific aspects or features mentioned within a text).

The importance of sentiment analysis spans multiple domains, impacting businesses, political entities, social media platforms, and more.

In the business world, sentiment analysis is utilized to gauge customer satisfaction, identify emerging trends, and monitor brand reputation. This valuable insight can inform decision-making processes and help companies tailor their strategies to better meet consumer needs. For instance, analyzing customer reviews can reveal common complaints and areas for improvement, enabling businesses to address issues proactively and enhance customer satisfaction. Bele et al. [1] discuss how businesses use political sentiment mining as an intelligence tool for strategy formulation. By understanding political sentiments, businesses can anticipate regulatory changes and adjust their strategies accordingly to maintain compliance and competitive advantage.

In the realm of politics, sentiment analysis plays a crucial role in understanding public sentiment towards policies, politicians, and current events. By analyzing the tone and emotions expressed in public discourse, political entities can gain valuable insights into public opinion, allowing them to craft more effective messaging and policies. This can lead to more targeted and responsive political campaigns, as well as better alignment of policies with public interests. Adwan et al. [2] emphasize the importance of sentiment analysis on platforms like Twitter, where public

sentiment towards political events can be quickly gauged. This real-time analysis helps political entities react promptly to changing public opinions and adjust their communication strategies.

Social media platforms utilize sentiment analysis to monitor user sentiment towards products, services, and trending topics. This information is invaluable for businesses looking to understand customer preferences and concerns, as well as for social media platforms seeking to enhance user experience and engagement. By analyzing user-generated content, platforms can identify trending topics, detect potential crises, and improve content recommendations. Pandey et al. [3] discuss various aspects of sentiment analysis, highlighting its application in social media monitoring. They note that sentiment analysis helps in tracking public opinion and sentiment dynamics, which is crucial for both businesses and social media platforms.

The application of sentiment analysis extends beyond these domains, permeating various aspects of modern society and providing valuable insights into human emotions and behaviors. For example, in healthcare, sentiment analysis can be used to analyze patient feedback and improve healthcare services. In finance, financial sentiment analysis can predict market trends based on public sentiment towards economic events or companies. In education, educational institutions can use sentiment analysis to assess student feedback and improve educational experiences.(Table 2.1)

Table 2.1 - Applications of Sentiment Analysis Across Domains

Domain	**Applications**	**Benefits**
Business	Customer satisfaction, brand monitoring, trend analysis	Informed decision-making, strategy formulation
Politics	Public sentiment towards policies and politicians	Effective messaging, responsive policy-making
Social Media	User sentiment towards products and topics	Enhanced user experience, crisis detection
Healthcare	Patient feedback analysis	Improved healthcare services
Finance	Market trend prediction	Better investment decisions
Education	Student feedback assessment	Enhanced educational experiences

Sentiment analysis is a powerful tool that offers profound insights into public opinion and sentiment across various domains. By leveraging advanced NLP techniques and computational models, entities can derive actionable insights that drive strategic decisions and enhance user experiences. The continuous evolution of sentiment analysis methodologies, as highlighted in the surveyed papers, under-

scores its growing significance and potential in understanding and responding to human emotions and attitudes effectively.

2.2 Historical Development of Sentiment Analysis

Sentiment analysis, also known as opinion mining, has evolved significantly over the years. One of the key milestones in its development was the emergence of machine learning and natural language processing techniques, which allowed for more accurate and efficient analysis of textual data. Early approaches to sentiment analysis relied heavily on manual lexicon creation and simple statistical methods. These methods, while foundational, had limitations in scalability and context understanding[4].

The advent of machine learning introduced more sophisticated techniques that improved the accuracy of sentiment analysis. Algorithms such as Naive Bayes, Support Vector Machines (SVM), and decision trees allowed for more nuanced sentiment classification by learning from labeled data. These methods marked a significant step forward from rule-based systems, enabling the handling of larger datasets and more complex linguistic patterns.

In recent years, the use of deep learning models has further revolutionized sentiment analysis. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have enhanced the capability to capture complex linguistic patterns and contextual information. CNNs are particularly effective in identifying spatial hierarchies in data, making them suitable for sentence-level sentiment analysis. RNNs, especially Long Short-Term Memory (LSTM) networks, excel in handling sequential data, capturing dependencies and contextual information over longer text spans[5].

The development of transformer models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) has significantly advanced sentiment analysis systems. BERT's ability to consider the context from both directions in a sentence (bidirectional context) has made it a powerful tool for various NLP tasks, including sentiment analysis. GPT, with its generative capabilities, has also been influential in understanding and generating human-like text, further enhancing sentiment analysis accuracy and application[5].

Another important development in sentiment analysis is the integration of domain-specific knowledge and context. By incorporating industry-specific terminology and understanding the cultural nuances of different languages, sentiment analysis models have become more adaptable and reliable in capturing the sentiment of diverse user groups. For instance, sentiment analysis in the healthcare domain benefits from integrating medical terminologies, while financial sentiment analysis incorporates economic and market-specific terms[6].

Looking ahead, the development of sentiment analysis is likely to continue on an upward trajectory. Researchers and practitioners are exploring new techniques and applications in fields such as social media monitoring, customer feedback analysis, and market research. As the volume of textual data continues to grow, sentiment analysis will play an increasingly vital role in understanding and responding to the sentiments of individuals and communities. The focus is also shifting towards

the explainability and interpretability of sentiment analysis models. This is particularly important in critical decision-making domains where understanding the inner workings of the models is crucial for transparency and accountability[7].

As sentiment analysis continues to evolve, the focus on explainability and interpretability of the models used becomes essential. This is particularly important in critical decision-making domains where understanding the inner workings of the sentiment analysis models is crucial for transparency and accountability[8].

The field of sentiment analysis has evolved significantly over the past few decades. Early approaches relied heavily on manual lexicon creation and simple statistical methods. However, with the advent of machine learning and natural language processing (NLP), more sophisticated techniques have been developed. The introduction of deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has further revolutionized sentiment analysis by improving the accuracy and efficiency of text analysis. Additionally, the development of pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) has significantly advanced the capabilities of sentiment analysis systems. (Table 2.2)

Table 2.2 - Key Developments in Sentiment Analysis

Time Period	**Development**	**Description**
Early Years	Manual Lexicon Creation	Initial methods relied on manually created lexicons and simple statistical techniques to classify sentiment.
2000s	Machine Learning Techniques	Introduction of algorithms like Naive Bayes, SVM, and decision trees improved sentiment classification by learning from labeled data.
2010s	Deep Learning Models	Models like CNNs and RNNs, particularly LSTMs, enhanced the ability to capture complex linguistic patterns and context in text.
Late 2010s	Transformer Models	BERT and GPT models significantly advanced the field by providing bidirectional context understanding and generative capabilities, respectively.
Recent Years	Domain-Specific Integration	Incorporation of industry-specific terminology and cultural nuances improved the adaptability and reliability of sentiment analysis models.
Ongoing	Explainability and Interpretability	Growing focus on making sentiment analysis models transparent and interpretable, especially in critical decision-making domains.

2.3 Methodologies in Sentiment Analysis

There are several key methodologies in sentiment analysis that are commonly used to understand and interpret textual data. One popular approach is the use of lexicon-based methods, which involve the creation and utilization of sentiment dictionaries to assign scores to words based on their emotional connotations. Another approach is machine learning, which involves training models on labeled data to classify text as positive, negative, or neutral. Additionally, deep learning techniques, such as recurrent neural networks and convolutional neural networks, have also shown promising results in sentiment analysis tasks. These methodologies, along with others, provide a diverse toolkit for analyzing and understanding sentiments expressed in text[4].

Lexicon-based sentiment analysis relies on predefined lists of words (sentiment lexicons) that have associated sentiment scores. These scores indicate the emotional valence (positive, negative, or neutral) of each word. Lexicon-based methods are straightforward and interpretable, making them suitable for quick sentiment analysis tasks and for applications where transparency is crucial. However, they may struggle with context-specific meanings and can miss nuances that require a deeper understanding of language. The simplicity of lexicon-based approaches makes them accessible and easy to implement, but their effectiveness heavily depends on the comprehensiveness and accuracy of the sentiment lexicons used[4].

Rule-based sentiment analysis techniques rely on predefined rules and patterns to interpret and classify text based on sentiment. These techniques involve the use of linguistic rules, such as identifying negations, intensifiers, and context-specific words to determine the sentiment of a given text. By applying these rules, the sentiment of the text can be inferred and categorized as positive, negative, or neutral. Rule-based techniques offer an interpretable and transparent way to analyze sentiment and can be particularly effective for domains with specialized language or specific sentiment expressions. For instance, in legal or medical texts, where specific terminology and context are crucial, rule-based systems can be fine-tuned to provide accurate sentiment analysis[9].

In addition to lexicon-based methods and machine learning, rule-based techniques play a significant role in sentiment analysis, providing a complementary approach for understanding and interpreting textual data. These methodologies collectively contribute to a comprehensive toolkit for sentiment analysis across diverse domains and applications. The flexibility of rule-based systems allows them to be adapted to various languages and cultural contexts, making them versatile tools for global applications[10].

Machine learning approaches are increasingly popular in sentiment analysis due to their ability to automatically learn and adapt to different types of textual data. These approaches typically involve the use of algorithms and statistical models to analyze and classify sentiment in text. By training on labeled data, machine learning models can effectively capture complex patterns and nuances in language, enabling more accurate sentiment classification. This adaptability makes machine learning approaches suitable for a wide range of applications, from social media monitoring to customer feedback analysis[11].

One common machine learning approach in sentiment analysis is the use of supervised learning algorithms, such as support vector machines (SVM), logistic regression, and random forests. These algorithms are trained on labeled datasets, where each piece of text is associated with a sentiment label (positive, negative, or neutral). Through the process of learning from these labeled examples, the machine learning model can generalize and make predictions on new, unseen text data. The performance of these models can be further enhanced by using feature engineering techniques to extract meaningful features from the text data, such as n-grams, part-of-speech tags, and syntactic dependencies[12].

Unsupervised learning techniques, such as clustering and topic modeling, also play a role in sentiment analysis by identifying patterns and groupings within textual data. These approaches can be valuable for uncovering underlying sentiment trends and themes in large volumes of unstructured text. Unsupervised methods do not require labeled data, making them suitable for exploratory analysis where labeling large datasets might be impractical[12].

Incorporating machine learning approaches into sentiment analysis can enhance the accuracy and scalability of sentiment classification, particularly in scenarios where rule-based or lexicon-based methods may not capture the full complexity of language. However, it is important to consider the potential challenges associated with machine learning, such as the need for large labeled datasets, model interpretability, and potential biases in training data. Despite these challenges, the integration of machine learning techniques enriches the toolkit for sentiment analysis, enabling more robust and sophisticated analysis of textual sentiments. Furthermore, advancements in transfer learning and the availability of pre-trained models have reduced the data requirements, making machine learning more accessible[11].

Deep learning techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown promising results in sentiment analysis tasks. CNNs are particularly effective in identifying spatial hierarchies in data, making them suitable for sentence-level sentiment analysis. RNNs, especially Long Short-Term Memory (LSTM) networks, excel in handling sequential data, capturing dependencies and contextual information over longer text spans. The introduction of transformer models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) has significantly advanced sentiment analysis systems. BERT's ability to consider the context from both directions in a sentence (bidirectional context) has made it a powerful tool for various NLP tasks, including sentiment analysis. GPT, with its generative capabilities, has also been influential in understanding and generating human-like text, further enhancing sentiment analysis accuracy and application[5].

In recent years, there has been a growing interest in developing hybrid models that combine the strengths of multiple sentiment analysis methodologies. These hybrid models aim to leverage the advantages of different approaches to achieve more accurate and nuanced sentiment classification. One approach to building hybrid sentiment analysis models involves integrating rule-based techniques with machine learning algorithms. By combining the explicit rules and pattern recognition of rule-based methods with the learning capabilities of machine learning,

these hybrid models can effectively capture both the structured linguistic rules and the complex contextual patterns present in textual data. This integration enables the model to benefit from the interpretability of rule-based techniques while also harnessing the predictive power of machine learning[13].

Another direction in hybrid sentiment analysis involves leveraging both lexicon-based methods and deep learning techniques. Lexicon-based approaches provide a foundation for understanding sentiment at the word level, while deep learning models excel in capturing intricate relationships and dependencies within text. By merging these two approaches, hybrid models can effectively capture the broad semantic understanding provided by lexicon-based methods and the nuanced contextual understanding offered by deep learning techniques. This combination enhances the overall performance of sentiment analysis systems, making them more versatile and robust across various applications[4].

The development and implementation of hybrid models in sentiment analysis represent an exciting frontier in the field, as they offer the potential to address the limitations of individual methodologies and achieve more comprehensive sentiment understanding. As research continues to explore the integration of diverse techniques, hybrid models are poised to play a key role in advancing the accuracy and flexibility of sentiment analysis across diverse domains and applications.

Table 2.3 - Key Sentiment Analysis Methodologies

Methodology	**Description**	**Strengths and Weaknesses**
Lexicon-Based Methods	Utilizes predefined sentiment dictionaries to assign scores to words	**Strengths:** Simple, interpretable **Weaknesses:** Context-insensitive, limited nuance
Rule-Based Techniques	Relies on predefined rules and patterns to classify sentiment	**Strengths:** Transparent, effective for specialized domains **Weaknesses:** Limited scalability, rigid
Machine Learning Approaches	Uses algorithms to learn from labeled data for sentiment classification	**Strengths:** Captures complex patterns, adaptable **Weaknesses:** Requires large labeled datasets, potential biases
Deep Learning Techniques	Employs neural networks to understand context and dependencies in text	**Strengths:** High accuracy, captures intricate relationships **Weaknesses:** Computationally intensive, less interpretable
Hybrid Models	Combines strengths of multiple methodologies for enhanced sentiment analysis	**Strengths:** Comprehensive, flexible **Weaknesses:** Complex implementation, potential overfitting

This table (Table 2.3) summarizes the key methodologies in sentiment analysis, highlighting their descriptions, strengths, and weaknesses. Including such a table can help provide a clear and concise overview of the different approaches used

in sentiment analysis, enhancing the reader's understanding of the methodologies discussed in the extended text.

2.4 Challenges in sentiment analysis

There are several challenges in sentiment analysis that need to be addressed. One of the major challenges is the ambiguity of language. Natural language is complex and often filled with sarcasm, irony, and figurative speech, making it difficult for sentiment analysis algorithms to accurately interpret the true sentiment behind the words. For instance, a sentence like "Great job! You really messed that up" can be interpreted as positive due to the word "Great," but the context indicates sarcasm and a negative sentiment. The challenge is further compounded by the subtleties and variances in human communication that machines struggle to interpret[14].

Another challenge is the cultural and contextual differences in language. Sentiment analysis models trained on one type of language or cultural context may not generalize well to other languages or cultures, leading to inaccuracies in analysis. For example, a phrase that is considered positive in one culture might be neutral or even negative in another. This cultural variance requires models to be adaptable and sensitive to these differences to accurately interpret sentiment across diverse user bases[15].

Furthermore, the dynamic nature of language and the constant evolution of slang and expressions pose challenges for sentiment analysis models to stay relevant and up to date. New slang terms, idiomatic expressions, and context-specific jargon continuously emerge, requiring models to be frequently updated. This dynamic aspect of language evolution means that a model trained on data from even a few years ago might struggle with contemporary language use, leading to outdated or inaccurate sentiment detection[16].

In addition, the presence of mixed sentiments within a single text or the use of negation can further complicate sentiment analysis accuracy. A single text can contain both positive and negative sentiments, making it challenging to classify the overall sentiment accurately. For example, a review might say, "The product quality is excellent, but the customer service was terrible." Such sentences convey mixed sentiments that simple classification models might not handle well. Similarly, negations like "not bad" can invert the sentiment, making it difficult for models to accurately interpret the true sentiment without sophisticated context understanding[16].

Addressing these challenges in sentiment analysis calls for the development of more sophisticated algorithms that can better understand the nuances of human language and adapt to cultural and contextual variations. It also requires continuous updates and retraining of models to keep up with the ever-changing landscape of language use. To improve sentiment analysis, researchers are exploring advanced techniques such as deep learning and natural language processing. These approaches aim to capture more nuanced language features and contextual clues to better discern the true sentiment of text. For example, transformer models like BERT and GPT-3 have shown significant promise in understanding context and handling the intricacies of human language[17]. Moreover, utilizing domain-

specific sentiment lexicons and incorporating knowledge graphs can enhance the accuracy of sentiment analysis by providing domain-specific insights and understanding relationships between entities and sentiments. Domain-specific lexicons can be tailored to the unique language and jargon of specific fields, such as medical or financial texts, leading to more accurate sentiment detection. Knowledge graphs, on the other hand, can provide context by linking related entities and concepts, helping models understand the broader context of the sentiment expressed.

Overcoming the challenges in sentiment analysis necessitates the integration of advanced technologies, diverse datasets, and continuous adaptation to the evolving landscape of human language. Diverse datasets that include various languages, dialects, and cultural contexts can help create more robust and generalizable models. Additionally, the development of transfer learning techniques allows models to leverage knowledge from related tasks or domains, improving their performance on new and diverse datasets[15].

Furthermore, ensuring model interpretability and transparency is crucial, especially in applications where decisions based on sentiment analysis have significant implications. Models need to provide explanations for their predictions to build trust and allow users to understand how decisions are made. Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can be used to explain the output of sentiment analysis models, helping users understand the factors contributing to the sentiment classification[16].

By addressing these obstacles, sentiment analysis can become a more robust and reliable tool for understanding public perception and feedback across diverse linguistic and cultural contexts. Continuous research and development in this field aim to create more adaptable, accurate, and interpretable models that can handle the complexities of human language. As sentiment analysis technologies advance, they will play an increasingly vital role in various applications, from customer feedback analysis and market research to social media monitoring and public opinion tracking.

This table (Table 2.4) summarizes the key challenges in sentiment analysis, their descriptions, and the impact and potential solutions. Including such a table can help provide a clear and concise overview of the challenges faced in sentiment analysis and the approaches being taken to address them.

Table 2.4 - Challenges in Sentiment Analysis

Challenge	**Description**	**Impact and Solutions**
Ambiguity of Language	Complex language elements like sarcasm, irony, and figurative speech	Difficult for algorithms to interpret true sentiment; advanced NLP and context-aware models needed
Cultural and Contextual Differences	Variations in sentiment across different cultures and contexts	Models trained on one context may not generalize well; requires adaptable and culturally sensitive models
Dynamic Nature of Language	Constant evolution of slang and expressions	Models need frequent updates to stay relevant; incorporation of up-to-date language data is crucial
Mixed Sentiments	Presence of both positive and negative sentiments in a single text	Complicates overall sentiment classification; advanced models and context understanding required
Negation Handling	Phrases with negation can invert sentiment	Accurate sentiment detection needs sophisticated handling of negations; context-aware models necessary
Domain-Specific Language	Unique terminology in specific fields	General models may struggle with domain-specific jargon; use of tailored lexicons and knowledge graphs enhances accuracy
Model Interpretability	Need for transparent and explainable models	Crucial for trust and understanding; techniques like SHAP and LIME can aid in explaining model decisions
Data Diversity	Inclusion of various languages, dialects, and cultural contexts	Enhances robustness and generalizability of models; diverse datasets and transfer learning techniques beneficial

2.5 Sentiment Analysis in Underrepresented Languages

While sentiment analysis has gained significant attention in well-resourced languages such as English, Chinese, and Spanish, conducting sentiment analysis in underrepresented languages presents unique challenges. One of the primary challenges is the scarcity of labeled datasets and resources in these languages. Without an adequate amount of labeled data, training sentiment analysis models becomes difficult, leading to poorer accuracy and performance. The lack of extensive corpora, sentiment lexicons, and annotated datasets in underrepresented languages hampers the development and evaluation of effective sentiment analysis systems. This scarcity results in models that are less robust and generalizable, struggling to capture the intricacies and nuances of these languages[15].

To overcome the scarcity of labeled data, researchers have explored various

strategies such as transfer learning, semi-supervised learning, and active learning. These approaches involve leveraging existing labeled data from well-resourced languages and using it to supplement the limited labeled resources in underrepresented languages. Transfer learning, in particular, has shown promise in adapting sentiment analysis models from one language to another, thereby reducing the reliance on abundant labeled data for each specific underrepresented language. For instance, a model trained on English data can be fine-tuned with a smaller set of labeled data from an underrepresented language, effectively transferring the learned features and improving the model's performance in the target language[18].

While transfer learning has shown promise in mitigating the scarcity of labeled data in underrepresented languages, another effective approach is collaboration among researchers and organizations. By pooling together resources and expertise, collaborative efforts can contribute to the creation of labeled datasets for sentiment analysis in underrepresented languages. Such collaborations can involve sharing data, tools, and best practices, as well as organizing joint annotation efforts to build comprehensive and high-quality datasets. Collaborative initiatives can significantly accelerate the progress in developing sentiment analysis systems for underrepresented languages by leveraging the collective knowledge and resources of the research community[19].

In addition to collaboration, data augmentation techniques can also be valuable in expanding the available labeled data. Methods such as back translation, synonym replacement, and paraphrasing can help generate diverse and augmented datasets for training sentiment analysis models in underrepresented languages. These techniques involve creating new training examples by modifying existing ones, thus increasing the size and diversity of the training data. Back translation, for example, involves translating a sentence from the target language to a pivot language (e.g., English) and then back to the target language, resulting in a paraphrased version of the original sentence. Such augmentation methods can help capture a broader range of language expressions and nuances, enhancing the performance and robustness of sentiment analysis models[15].

Despite these advancements, challenges remain in ensuring the quality and consistency of augmented data. It is crucial to maintain the original sentiment and context during the augmentation process to avoid introducing noise and inaccuracies. Furthermore, continuous efforts are needed to develop more sophisticated and automated data augmentation techniques that can effectively handle the diverse linguistic features of underrepresented languages.

The development of sentiment analysis systems for underrepresented languages also requires a focus on creating language-specific resources such as sentiment lexicons, syntactic parsers, and word embeddings. These resources can provide valuable linguistic insights and support the training of more accurate and context-aware models. Additionally, integrating cultural and contextual knowledge into sentiment analysis models can help address the unique challenges posed by different languages and regions.

In conclusion, overcoming the challenges in sentiment analysis for underrepresented languages necessitates a multifaceted approach that combines advanced technologies, collaborative efforts, and innovative data augmentation techniques.

By addressing these challenges, sentiment analysis can become a more inclusive and effective tool for understanding public perception and feedback across a diverse range of languages and cultures.

This table (Table 2.5) summarizes the strategies for enhancing sentiment analysis in underrepresented languages, highlighting their descriptions, impacts, and solutions. Including such a table can provide a clear and concise overview of the approaches being taken to address the challenges in this area, enhancing the reader's understanding of the methodologies and their effectiveness.

Table 2.5 - Strategies for Enhancing Sentiment Analysis in Underrepresented Languages

Strategy	**Description**	**Impact and Solutions**
Transfer Learning	Leveraging existing labeled data from well-resourced languages	Reduces reliance on abundant labeled data for underrepresented languages; improves model adaptation
Collaboration	Pooling resources and expertise among researchers and organizations	Enhances dataset creation through joint efforts; accelerates development of robust models
Data Augmentation	Techniques like back translation, synonym replacement, and paraphrasing	Expands labeled datasets; captures diverse expressions and nuances; improves model robustness
Semi-Supervised Learning	Combining labeled and unlabeled data for model training	Utilizes available data more effectively; improves model performance with limited labeled data
Active Learning	Iterative process of labeling most informative data points	Optimizes labeling efforts; enhances dataset quality and model accuracy
Creating Language-Specific Resources	Developing sentiment lexicons, syntactic parsers, and word embeddings	Provides linguistic insights; supports accurate and context-aware models
Integrating Cultural and Contextual Knowledge	Incorporating regional and cultural nuances into models	Addresses unique linguistic challenges; improves sentiment accuracy and relevance

2.6 Edit Distance Algorithms

Edit distance algorithms, such as Levenshtein distance and Jaccard distance, play a crucial role in computational linguistics and are highly relevant to sentiment analysis. These algorithms measure the similarity between two strings and are widely used in natural language processing tasks. The Levenshtein distance, for example, calculates the minimum number of operations (insertions, deletions, or substitutions) required to transform one string into another. This measure of similarity is fundamental in various NLP applications, providing a quantitative approach to compare textual data[20].

In computational linguistics, edit distance algorithms are employed in tasks such as spell checking, text summarization, and machine translation. By calculating the minimum number of operations required to transform one string into another, these algorithms help in identifying linguistic patterns and similarities. For instance, in spell checking, edit distance can identify the closest correct word to a misspelled word by finding the word with the smallest edit distance. Similarly, in machine translation, these algorithms assist in aligning parallel corpora by matching phrases with minimal differences, thereby improving translation accuracy[20].

Moreover, in sentiment analysis, edit distance algorithms can be utilized to compare and analyze textual data, enabling the extraction of valuable insights from social media posts, customer reviews, and other forms of user-generated content. By quantifying the differences between textual expressions, these algorithms contribute to understanding the sentiment and emotions conveyed in the language. For example, they can be used to cluster similar sentiment expressions or detect changes in sentiment over time by comparing different versions of a text[21].

Overall, the concept of edit distance algorithms is fundamental in computational linguistics and holds significant relevance in the context of sentiment analysis, contributing to the advancement of language processing and understanding human emotions through textual data. Furthermore, edit distance algorithms have found application in information retrieval systems, where they are used to measure the similarity between search queries and documents. This enables more accurate retrieval of relevant information based on the closeness of the query and the document content. In information retrieval, edit distance helps in ranking documents by their relevance to the search query, improving the user's search experience[22].

In computational linguistics, these algorithms are also utilized in the field of named entity recognition, aiding in the identification and classification of entities such as names of people, organizations, and locations within a given text corpus. Named entity recognition benefits from edit distance algorithms by improving the accuracy of entity matching and disambiguation. Additionally, edit distance algorithms have been integrated into machine learning models for tasks such as text classification and clustering. By leveraging the similarity metrics provided by these algorithms, machine learning algorithms can better understand and process textual data for various applications. For instance, in text classification, edit distance can enhance feature extraction by quantifying the similarity between text samples, leading to better classification performance[21].

In the context of sentiment analysis, researchers are exploring the combina-

tion of edit distance algorithms with other natural language processing techniques to enhance the accuracy and depth of sentiment classification models, leading to more nuanced analysis of textual expressions. This combination can improve the detection of subtle sentiment changes and the handling of complex linguistic phenomena such as sarcasm and irony. For example, integrating edit distance with deep learning models can help in fine-tuning sentiment analysis by providing additional features that capture textual similarity and variation[23].

The versatility and applicability of edit distance algorithms across different domains underscore their significance in computational linguistics and their potential to further enrich the capabilities of sentiment analysis and language processing technologies. These algorithms provide a robust framework for various NLP tasks, enhancing the overall understanding and processing of human language.

This table (Table 2.6) summarizes the key applications of edit distance algorithms in computational linguistics, highlighting their descriptions, impacts, and solutions. Including such a table can provide a clear and concise overview of the diverse uses of edit distance algorithms, enhancing the reader's understanding of their relevance and applicability in various NLP tasks.

To further illustrate the concept of edit distance, here is an example calculation of Levenshtein distance between two strings:

		s	i	t	t	i	n
g							
	0	1	2	3	4	5	6
7							
k	1	1	2	3	4	5	6
7							
i	2	2	1	2	3	4	5
6							
t	3	3	2	1	2	3	4
5							
t	4	4	3	2	1	2	3
4							
e	5	5	4	3	2	2	3
4							
n	6	6	5	4	3	3	2
3							

Figure 2.6.1 - Levenshtein distance matrix for "kitten" and "sitting".

Example of Levenshtein Distance Calculation:

Calculate the Levenshtein distance between the words *kitten* and *sitting*. The Levenshtein distance between *kitten* and *sitting* is 3.

This example (Figure 2.6.1) shows the step-by-step calculation of the Levenshtein distance between two words, highlighting the operations required to transform one word into the other. Including such examples can help readers better understand the practical application and calculation of edit distance algorithms in NLP tasks.

Table 2.6 - Applications of Edit Distance Algorithms in Computational Linguistics

Application	**Description**	**Impact and Solutions**
Spell Checking	Identifies the closest correct word to a misspelled word	Enhances accuracy of spelling correction by finding minimal edit distance
Text Summarization	Helps in comparing and summarizing texts by identifying key differences	Improves summary quality by focusing on minimal changes
Machine Translation	Assists in aligning parallel corpora by matching similar phrases	Enhances translation accuracy through minimal edit transformations
Sentiment Analysis	Compares and analyzes textual data to extract sentiment insights	Provides quantitative measures for sentiment comparison and clustering
Information Retrieval	Measures similarity between search queries and documents	Improves relevance ranking and retrieval accuracy
Named Entity Recognition	Aids in identifying and classifying entities in text corpora	Enhances entity matching and disambiguation accuracy
Text Classification	Integrates similarity metrics for better feature extraction and classification	Improves classification performance by quantifying textual similarity

2.7 Sentiment Analysis for the Kazakh Language

Comparative Analysis: Rule-Based vs. Machine Learning in Kazakh Language

When comparing rule-based and machine learning approaches for the Kazakh language, it is essential to consider the specific characteristics of the language and the intended application. Rule-based systems rely on handcrafted linguistic rules to analyze and process the language, while machine learning methods use large datasets to learn patterns and make predictions.

In the context of the Kazakh language, rule-based approaches may be advantageous for tasks that require precise grammatical and syntactic analysis. These systems can be tailored to the unique features of Kazakh, such as its agglutinative nature and complex morphology. However, they often require extensive manual effort and may struggle with capturing the full range of linguistic variation present in natural language[24].

On the other hand, machine learning models have the potential to automatically learn patterns from data, making them well-suited for tasks like text classification, sentiment analysis, and machine translation. With the availability of large corpora of Kazakh texts, machine learning approaches can leverage this data to make accurate predictions and generalize to new inputs.

In a comparative analysis, it is important to consider the specific requirements of the application, the availability of linguistic resources, and the desired balance between precision and scalability. Ultimately, the choice between rule-based and machine learning approaches in the context of the Kazakh language will depend on the specific use case and the resources available for development and training.

Impact of Technological Advancements in Sentiment Analysis Research for Kazakh and Similar Languages

The impact of technological advancements in sentiment analysis research for the Kazakh language and similar languages has been substantial. With the increasing availability of language-specific datasets and the development of advanced natural language processing algorithms, sentiment analysis for Kazakh has seen significant improvements.

Machine learning techniques, such as deep learning models and transformer-based architectures, have shown promising results in sentiment analysis for Kazakh. These approaches can effectively capture the nuances and complexities of sentiment expression in the language, leading to more accurate classifications of text data. Furthermore, the availability of pre-trained language models has expedited the development of sentiment analysis systems for Kazakh, allowing for quicker deployment and adaptation to specific domains[25].

Additionally, the emergence of sentiment lexicons and annotated corpora tailored to Kazakh has provided valuable resources for training sentiment analysis models. These resources enable the fine-tuning of machine learning models to better understand sentiment cues specific to the language, leading to improved overall performance[26].

As technological advancements continue to drive innovation in sentiment analy-

sis research, it is expected that the accuracy and applicability of sentiment analysis systems for Kazakh and similar languages will further improve, ultimately enhancing the understanding of sentiment in diverse linguistic contexts.

2.8 New Insights and Advancements in Sentiment Analysis

Ontological and Rule-Based Approaches for Kazakh Sentiment Analysis

Recent developments in sentiment analysis for the Kazakh language highlight the application of ontological models coupled with rule-based methods. [27] present a significant advancement through the creation of a specialized dictionary for Kazakh, which underpins an ontological framework designed to enhance the granularity with which sentiment can be analyzed in local literature. This framework systematically categorizes sentiments and aligns them with morphological rules, thereby enabling precise sentiment interpretation. Preliminary empirical assessments have demonstrated an accuracy rate of 83% in detecting sentiments in short phrases, suggesting a high potential for effective sentiment analysis in domains requiring nuanced understanding of Kazakh.

The use of ontologies in sentiment analysis allows for a more structured representation of knowledge. Ontologies define a set of concepts and categories in a subject area or domain, showing their properties and the relations between them. By incorporating ontological models, sentiment analysis systems can achieve a higher level of abstraction and reasoning, improving the accuracy and depth of sentiment detection. This is particularly useful for processing agglutinative languages like Kazakh, where words can have complex morphological structures.

Enhancing Text Classification in Low-Resource Agglutinative Languages

In response to the challenges posed by morphological diversity and the scarcity of training data in agglutinative languages, Li Z. [28] introduced AgglutiFiT, a novel fine-tuning methodology for low-resource language models. This approach focuses on refining pre-trained language models through morphological analysis and stem extraction, creating a low-noise dataset that enhances the model's ability to discern relevant syntactic and semantic features. The refined attributes are then employed in subsequent text classification tasks, demonstrating an innovative use of attention mechanisms to improve the performance of language models in processing complex language structures.

AgglutiFiT addresses the issue of data scarcity by optimizing the use of existing linguistic resources. It fine-tunes pre-trained models by focusing on the morphological features specific to agglutinative languages, which often concatenate multiple morphemes into single words. By doing so, it enhances the model's understanding of syntactic and semantic nuances, leading to better performance in text classification tasks. This methodology represents a significant step forward in the development of NLP tools for low-resource languages, providing a more robust framework for processing complex linguistic data.

Big Data Techniques in Sentiment Analysis

The integration of big data technologies has revolutionized sentiment analysis, particularly through the application of scalable frameworks like Hadoop. Kurian D. [29] explored the efficacy of using Hadoop to process large-scale Twitter datasets for sentiment analysis. Their approach not only increased the speed and accuracy of sentiment detection but also showcased the potential of big data technologies in managing vast amounts of unstructured data. This development is crucial for organizations aiming to harness the full spectrum of user-generated content across social media platforms, thereby enriching the strategic insights gained from big data sentiment analysis.

Big data frameworks like Hadoop enable the processing and analysis of massive datasets that would be computationally infeasible with traditional methods. By distributing the processing load across multiple nodes, Hadoop allows for faster data processing and more scalable analysis. This capability is particularly valuable for sentiment analysis, where large volumes of data from social media and other sources can be processed efficiently, leading to more timely and accurate insights.

Preprocessing Techniques for Robust Sentiment Analysis

The complexity of user-generated content, especially from diverse linguistic backgrounds, necessitates robust preprocessing methods to ensure accurate sentiment analysis. Niyazmetova K. [30] emphasized the importance of preprocessing in their study of sentiment analysis for Tashkent restaurant reviews on Google Maps. They employed advanced preprocessing techniques such as stemming, which is particularly tailored for agglutinative languages like Uzbek, to standardize text input and enhance the emotional granularity that can be achieved. Their use of logistic regression models further illustrates the integration of robust statistical methods to improve the reliability of sentiment categorization, making a significant contribution to the field by adapting traditional methods to unique linguistic contexts.

Preprocessing is a critical step in sentiment analysis that involves cleaning and transforming raw text data into a structured format suitable for analysis. Techniques such as stemming, lemmatization, and stop-word removal help reduce noise and standardize text input. For agglutinative languages, where words can have multiple affixes, stemming helps by reducing words to their base or root form, thereby simplifying the analysis. By enhancing the preprocessing stage, researchers can improve the accuracy and reliability of sentiment analysis models.

Morphological Normalization in Kazakh Language Processing

Tussupov J. [31] have made strides in developing normalization algorithms and morphological models that are specifically designed for the Kazakh language. Their work involves creating guidelines for synthesizing normalized forms and identifying word bases, which are critical in managing non-dictionary terms and evolving language use. The development of a Kazakh thesaurus for scientific and technical terms underscores the versatility and reliability of their approach, particularly in specialized fields. This research not only enhances text processing capabilities but also provides valuable tools for linguistic adaptation in morphologically rich

languages.

Normalization in NLP involves converting different forms of a word into a single canonical form. For languages with rich morphology like Kazakh, normalization helps in handling variations in word forms due to inflections, derivations, and other morphological processes. By developing comprehensive normalization algorithms, researchers can improve the consistency and accuracy of text processing, making it easier to analyze and interpret large corpora of text in these languages.

Challenges and Strategies in Indirect Translation

The study by Zhumabekova A. [32] delves into the complexities of translating sentiments from English to Kazakh, with Russian serving as an intermediary language. They explore the semantic shifts and stylistic adaptations necessary to maintain the integrity and cultural nuances of translated texts. The authors propose strategies to overcome these translation challenges, emphasizing the need for linguistic competence and cross-cultural sensitivity. This work enriches the understanding of indirect translation processes and offers practical insights for translators and researchers working within multilingual and multicultural frameworks.

Indirect translation, which involves translating from one language to another through a third language, introduces additional challenges in maintaining the original sentiment and context. Semantic shifts can occur due to differences in cultural norms, idiomatic expressions, and stylistic preferences. Strategies to address these challenges include maintaining close communication with native speakers, using back-translation techniques to verify accuracy, and employing culturally aware translators who can adapt the text to fit the target audience's context.

Probabilistic Modeling in Sentiment Analysis

Exploring probabilistic models, Surya, P. [33] applied the Naive Bayes classifier to the Amazon product review dataset, demonstrating the model's effectiveness in distinguishing between sentiments in textual data. Their study underscores the importance of probabilistic approaches in sentiment analysis, offering a detailed examination of how these models can be calibrated to improve accuracy in real-world applications.

Probabilistic models like Naive Bayes offer a simple yet powerful approach to text classification. By estimating the probability of each class given the input features, these models can effectively distinguish between different sentiment classes. The application of Naive Bayes to large datasets, such as Amazon product reviews, demonstrates its scalability and robustness in handling diverse and noisy data. Calibration techniques, such as smoothing and feature selection, can further enhance the model's performance.

Naive Bayes and K-Means in E-commerce Sentiment Analysis

Hariguna, T. [34] combined the Naive Bayes classifier with K-means clustering to analyze sentiment in customer reviews on Shopee. This hybrid approach not only categorized sentiments with significant accuracy but also highlighted the prevalence of negative feedback in certain product categories. Their findings illus-

trate the practical implications of sentiment analysis in e-commerce, where understanding consumer sentiment can directly influence product strategies and brand perception.

The combination of Naive Bayes and K-means clustering leverages the strengths of both methods: the probabilistic classification of Naive Bayes and the unsupervised clustering capabilities of K-means. This hybrid approach enables the identification of sentiment clusters within the data, providing deeper insights into consumer behavior and preferences. In e-commerce, such insights are valuable for optimizing product offerings, improving customer service, and enhancing overall customer satisfaction

Table 2.7 - Recent Advancements in Sentiment Analysis

Advancement	**Description**	**Impact and Solutions**
Ontological and Rule-Based Approaches for Kazakh Sentiment Analysis	Specialized dictionary and ontological framework for Kazakh	Enhances sentiment interpretation, achieves 83% accuracy in detecting sentiments in short phrases
Enhancing Text Classification in Low-Resource Agglutinative Languages	AgglutiFiT: fine-tuning pre-trained models with morphological analysis	Improves model performance in processing complex language structures
Big Data Techniques in Sentiment Analysis	Use of Hadoop for large-scale Twitter sentiment analysis	Increases speed and accuracy of sentiment detection, showcases potential of big data
Preprocessing Techniques for Robust Sentiment Analysis	Advanced preprocessing like stemming for Uzbek	Enhances emotional granularity, improves reliability of sentiment categorization
Morphological Normalization in Kazakh Language Processing	Normalization algorithms and morphological models for Kazakh	Enhances text processing, provides tools for linguistic adaptation
Challenges and Strategies in Indirect Translation	Translating sentiments from English to Kazakh via Russian	Maintains integrity and cultural nuances, offers insights for multilingual translation
Probabilistic Modeling in Sentiment Analysis	Naive Bayes applied to Amazon product reviews	Effective in distinguishing sentiments, improves accuracy
Naive Bayes and K-Means in E-commerce Sentiment Analysis	Combining Naive Bayes with K-means for Shopee reviews	Categorizes sentiments accurately, highlights e-commerce implications

This table (Table 2.7) summarizes the recent advancements in sentiment analysis, highlighting their descriptions, impacts, and solutions. Including such a table can provide a clear and concise overview of the various innovative approaches and their contributions to the field, enhancing the reader's understanding of the latest

developments in sentiment analysis.

Chapter 3

Methods and Materials

3.1 Dataset

This study employed three separate datasets to perform an extensive analysis, each described as follows:

1. **D. Chapaev's Sentiment Analysis Dataset:**
   - **Source:** Available on Dauren Chapaev's GitHub page [35].
   - **Content:** This dataset includes 20,014 sentences from various news outlets, with sentiment classifications of 5,993 negative, 4,422 positive, and 9,599 neutral sentences.
2. **Serek's Agglutinative Language Sentiment Dataset:**
   - **Authors:** Azamat Serek and colleagues' project on sentiment analysis is presented in the work titled "Bir atanyn balasy" (1973) by Mukhtar Makhauin [36].
   - **Content:** Comprising 732 sentences, this dataset is segmented into 231 positive, 228 negative, and 273 neutral sentiments, based on the emotional intensity of the literary content, which ranges from negative emotions like anger and sadness to positive ones like joy and euphoria, with remaining sentiments categorized as neutral.
3. **Kaggle Amazon Sentiment labeled Dataset:**
   - **Source:** This dataset was sourced from Kaggle and involves sentences initially in English that were translated into Kazakh for the study 'From Group to Individual Labels using Deep Features' [37].
   - **Content:** It consists of sentences from imdb.com, amazon.com, and yelp.com, each contributing 500 positive and 500 negative sentences. However, only Amazon.com sentences were selected for this analysis.

For the rule-based sentiment analysis, a compilation of Kazakh language adjectives was sourced from Sozdik Qor [38], a robust platform providing access to a vast array of words and phrases from various dictionaries. This platform is particularly useful for its comprehensive search features covering synonyms, antonyms, homonyms, and phraseology. The adjectives were initially untagged; sentiment scores ranging from -1 to 1 were manually assigned later. The resulting dataset includes 5,539 adjectives with sentiment tags distributed as 1,902 near positive, 1,657 close to negative, and 1,980 near neutral.

3.2 Rule-Based Sentiment Analysis

This section outlines the established guidelines for assessing sentiment in Kazakh language phrases:

1. **Adjective Tonality:** Define the tonality of an adjective Ta as directly influencing the phrase's sentiment (Tp), such that $Ta = Tp$.
   Example: "jaqsy adam" translates to a Positive sentiment (tonality is 1).
   Explanation:
   - "jaqsy" is a positive adjective with a sentiment score of 1.
   - "adam" is a noun.
2. **Adverb Influence:** If an adverb (A) modifies an adjective, it amplifies the adjective's tonality. Let Ta be the adjective's initial tone, and Tp the phrase's tonality, then $Tp = 2 \times Ta$.
   Example: "ote jaqsy adam" also translates to a Positive sentiment (tonality is 2).
   Explanation:
   - "ote" is an adverb.
   - "jaqsy" is a positive adjective with a score of 1.
   - "adam" is a noun.
3. **Negation Impact:** The presence of a negation (N) following an adjective inverses its tonality. If Ta is the adjective's tone, then the phrase's tonality $(Tp) = -Ta$.
   Example: "jaqsy adam emes" results in a Negative sentiment (tonality is -1).
   Explanation:
   - "jaqsy" is a positive adjective with a score of 1.
   - "adam" is a noun.
   - "emes" represents negation.
4. **Cumulative Adjective Tonality:** For phrases with multiple adjectives, the overall sentiment is calculated by summing the individual tonalities. For example, the cumulative tone from "ademi +0.7" and "ote jaman -2" is -1.3, indicating a Negative sentiment.

These rules form the basis for systematically analyzing sentiment in Kazakh, enabling a nuanced interpretation of textual emotions through the interplay of adjectives, adverbs, and negations.

3.3 Integration of Edit Distance

Edit distance, specifically the Levenshtein distance, is a metric for measuring the similarity between two text strings. It quantifies the minimum number of single-character edits (insertions, deletions, substitutions) required to change one word into another. This measurement is particularly relevant in text processing for identifying and correcting typographical errors that can significantly alter the sentiment conveyed by text. By integrating edit distance into sentiment analysis algorithms, we can enhance the accuracy of sentiment detection in the Kazakh

language, which is prone to such inconsistencies due to its morphological complexity.

Mathematical Formulation of Edit Distance

The Levenshtein distance between two strings, s and t, measures the minimum number of single-character edits required to change one string into the other. These edits can be insertions, deletions, or substitutions. The formula for calculating the Levenshtein distance is defined recursively as follows:

$$d(s,t) = \begin{cases} |s| & \text{if } |t| = 0, \\ |t| & \text{if } |s| = 0, \\ d(\text{tail}(s), \text{tail}(t)) & \text{if } \text{head}(s) = \text{head}(t), \\ 1 + \min \left\{ \begin{array}{l} d(\text{tail}(s), t), \\ d(s, \text{tail}(t)), \\ d(\text{tail}(s), \text{tail}(t)) \end{array} \right\} & \text{otherwise.} \end{cases} \tag{3.3.1}$$

Case Analysis

Base Cases:

- If $|t| = 0$: The distance $d(s,t)$ is $|s|$, the length of s. This means if t is an empty string, the distance is the number of deletions required to remove all characters from s, making s empty as well.
- If $|s| = 0$: Similarly, if s is empty, the distance $d(s,t)$ is $|t|$, representing the number of insertions needed to transform s into t by adding all characters of t.

Recursive Cases:

- If the first characters are equal ($\text{head}(s) = \text{head}(t)$): If the first characters of both strings are the same, the function recursively calculates $d(\text{tail}(s), \text{tail}(t))$, i.e., the distance between the rest of the strings after removing the first character from both. This situation does not require an edit operation because the first characters already match.
- Otherwise: If the first characters are not the same, the formula considers the minimum of three possible edit operations:
  1. **Deletion:** Remove the first character from s, hence the recursive call $d(\text{tail}(s), t)$, and add 1 to the result (counting this deletion).
  2. **Insertion:** Add the first character of t to s, prompting the calculation $d(s, \text{tail}(t))$, and increment the count by 1 for this insertion.
  3. **Substitution:** Replace the first character of s with that of t and proceed to calculate $d(\text{tail}(s), \text{tail}(t))$, adding 1 to account for this substitution.

Justification for Word Combination Choices

The choice of specific word combinations such as:

- **[ADJECTIVE] + [Negation] + [NOUN]**
- **[ADJECTIVE] + [NOUN] + [Negation]**
- **[ADJECTIVE] + [VERB] + [Negation]**
- **[ADVERB] + [ADJECTIVE]**

- **[ADJECTIVE] + [NOUN]**
- **[NOUN] + [ADJECTIVE]**
- **[ADJECTIVE] + [NEGATION]**
- **[ADJECTIVE] + [VERB]**
- **[ADJECTIVE]**

are strategically designed to address the syntactic structures unique to the Kazakh language, which influence sentiment expression. These combinations are significant as they mirror common linguistic constructs that critically affect the overall sentiment of phrases, thereby enabling more nuanced and accurate sentiment analysis.

Role of Edit Distance in Enhancing Rule-Based Methods

Edit distance algorithms enhance rule-based sentiment analysis by ensuring that words are compared in their intended form, free from typographical errors. This precision is crucial for maintaining the integrity of rule-based processing, which relies heavily on exact matches for word forms and their associated sentiment scores. By correcting these errors, edit distance algorithms significantly improve the overall reliability and accuracy of sentiment assessments.

Application Limits of Edit Distance

The application of edit distance in this research is strategically focused on individual word comparisons within predefined syntactic combinations rather than across entire sentences. This targeted approach is driven by the need for computational efficiency and the recognition that the most impactful errors in text sentiment analysis occur at the word level, especially in morphologically rich languages like Kazakh.

Detailed Applications for Each Word Combination

For each predefined syntactic combination:

- **[ADJECTIVE] + [Negation] + [NOUN]**: Edit distance corrects any misspellings in adjectives or negations, ensuring that their inverse impact on the noun's sentiment is accurately calculated.
- **[ADJECTIVE] + [NOUN] + [Negation]**: This setup ensures that when a negation follows a noun, the sentiment assigned by the adjective is appropriately negated, reflecting the intended sentiment of the phrase.

Additional combinations follow a similar rationale, with the edit distance providing a mechanism to ensure that the key elements of each combination are correctly identified and their relational sentiment impacts are accurately computed.

3.4 Data Processing and Machine Learning Approaches

Data Preprocessing

Before using the machine learning models, the text was carefully cleaned to guarantee that the data entering the models was in the best possible condition.

This cleaning involves a number of steps:

- **Lowercasing:** Change all text to lowercase to provide uniformity and eliminate case-related discrepancies.
- **HTML Tag Removal:** To get rid of any extraneous text, remove HTML tags.
- **Special Character, Number, and Punctuation Removal:** Remove punctuation, special characters, and number values to make your writing more focused and lucid.
- **Stopword Removal:** Remove often-used stop words to reduce noise and emphasize meaningful words.
- **Stemming:** To allow more effective feature extraction, use stemming to reduce words to their most basic form.

Vectorization Techniques

Two popular techniques were used to vectorize the processed text input in order to offer numerical properties for machine learning models:

- **TF-IDF Vectorizer:** The Frequency-Inverse Document Frequency (TF-IDF) technique is utilized to quantify the importance of a phrase with respect to a collection of documents. In addition to the term's frequency in a document, it considers the inverted document frequency throughout the whole dataset. The reason TF-IDF was chosen is because of its ability to highlight key phrases in a document while reducing the number of common terms. This improves the discriminating power of the sentiment analysis features [39].
  Mathematical Formulation: The calculation of TF-IDF for a term t in a document d is as follows:

$$TF - IDF(t, d) = TF(t, d) \times IDF(t) \tag{3.4.1}$$

  where TF is the term frequency and IDF is the inverse document frequency. This process highlights the importance of phrases in a document while downplaying common words.
- **Count Vectorizer:** Using count vectorization, the text is converted into a sparse matrix that shows the quantity of each phrase in the document. The selection of count vectorization is based on its ease of use and effectiveness in determining the word frequency in a document. It offers a simple depiction of word occurrences, which is advantageous for some kinds of emotion analysis assignments [40].
  Mathematical Formulation: A matrix element is used to indicate the count of each phrase in a text:

$$Count(t, d) = \text{number of occurrences of term } t \text{ in document } d \tag{3.4.2}$$

Machine Learning Models

Two widely used classifiers were combined with the vectorization algorithms:

- **Logistic Regression:** Logistic regression is a linear model that performs well in situations involving binary classification. It helps with sentiment analysis tasks and forecasts the probability that a sample will fall into a particular class. Logistic regression is a better solution because of its ease of use, efficiency, and interpretability. It often performs well in text classification tasks [41].
  Mathematical Formulation: Logistic regression uses the sigmoid function to forecast the likelihood that a sample will belong to a certain class.

$$P(y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)}} \tag{3.4.3}$$

  where $b_0, b_1, \ldots, b_n$ are coefficients and $x_0, x_1, \ldots, x_n$ are features.
- **Multinomial Naive Bayes:** Probabilistic classifier Bayes theorem serves as the foundation for multinomial Naive Bayes. Given that it assumes that the features are conditionally independent given the class, it performs particularly well in text classification applications. Based on its effectiveness and efficiency in handling high-dimensional, sparse data, Naive Bayes was chosen. It is well suited for tasks that include word frequencies in documents [42].
  Mathematical Formulation: Using a document's attributes as input, Naive Bayes determines the likelihood that it belongs in a class.

$$P(\text{class} \mid \text{features}) = P(\text{features} \mid \text{class}) \times P(\text{class})/P(\text{features}) \tag{3.4.4}$$

  The computation is made simpler by the conditional independence assumption.
- **Decision Trees:** A highly interpretable model for regression and classification that divides the dataset into branches. Textual data may be efficiently categorized using decision trees according to feature thresholds.
  Mathematical Formulation: To produce subsets that are as pure as feasible at each node, decision trees partition data based on criteria like information gain or Gini impurity.
- **Random Forest:** An ensemble of decision trees called Random Forest reduces overfitting and increases prediction accuracy. To create a prediction that is more reliable and accurate, it integrates the forecasts from several decision trees.
  Mathematical Formulation: The Random Forest model uses its decision trees to determine the majority vote for categorization problems. It averages the outcomes in regression tasks. The use of unpredictability in the model-building process is beneficial.
- **XGBoost:** eXtreme Gradient Boosting, or XGBoost, is a scalable and effective gradient boosting algorithm. In data contests, it is renowned for its effectiveness and quickness. When managing structured data for both regression and classification problems, XGBoost is quite helpful.
  Mathematical Formulation: To reduce these mistakes across all forecasts, XGBoost iteratively adds trees that anticipate the residuals or errors of previous trees in order to optimize a loss function.

Along with Logistic Regression and Multinomial Naive Bayes, Decision Trees, Random Forest, and XGBoost encompass a wide range of machine learning techniques, from basic to sophisticated models, which enhances the analysis. This broad collection of models was chosen with attention for qualities like interpretability, efficiency, and efficacy in managing high-dimensional data, based on their complementing capabilities in addressing various facets of sentiment analysis.

Data Splitting

Each dataset was divided into training and testing sets to make it easier to assess how well the suggested models performed using a variety of data sources. More specifically, 80% of each dataset was put aside for training and the remaining 20% was set aside for testing. All three datasets will have a consistent assessment framework thanks to our methodical splitting strategy.

Computational Environment

The research was conducted using Google Colab, a cloud-based platform for Python programming and machine learning. Google Colab provides a simple and scalable environment for executing code, which is particularly useful for resource-intensive tasks like machine learning. This platform makes cooperation simpler and guarantees that the approach given may be done with simplicity.

Chapter 4

Results and Discussion

4.1 Results

The findings are illustrated using the F1 score for the applied model on each unique dataset, providing a detailed view of the performance of various vectorizers, classifiers, and the rule-based approach (Tables 4.1, 4.2, and 4.3). It is important to note that these results were obtained without the application of edit distance.

Table 4.1 - 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 4.2 - 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 4.3 - Amazon Sentiment Labeled Sentences Dataset Results

#	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

Results of Sentiment Analysis on Dauren Chapaev's Dataset

When paired with TF-IDF vectorization, Logistic Regression and Multinomial Naive Bayes perform admirably on Dauren Chapaev's dataset, achieving F1 scores as high as 0.80. The performance of Decision Trees, Random Forest, and XGBoost varies, with Random Forest exhibiting the highest effectiveness. Although less successful in this dataset, the Rule-Based method demonstrates the variety of linguistic challenges present.

Results for Azamat Serek's Agglutinative Language Sentiment Dataset

The ensemble models—particularly Random Forest and Decision Trees—show a greater ability to manage the linguistic intricacies of Azamat Serek's dataset. This performance underscores the capability of these models to analyze texts with complex linguistic patterns. The sustained success of the Rule-Based method highlights its adeptness at interpreting the nuances of literary phrases.

Results from the Amazon Sentiment Labeled Sentences Dataset

Random Forest performs best in this dataset, underscoring the advantages of ensemble approaches. This resilience to the diverse linguistic patterns typical of internet reviews indicates their robustness. Alternatives like Decision Trees and XGBoost also showcase the variety of effective sentiment analysis methods available.

4.2 Discussion

Comparative Evaluation

- **Vectorization Techniques:** TF-IDF and Count Vectorizer consistently excel across all datasets, demonstrating their reliability in capturing textual features irrespective of linguistic context.
- **Classifier Performance:** The versatility of Multinomial Naive Bayes and Logistic Regression across different datasets suggests their general suitability for sentiment analysis tasks. The choice of model may depend on specific project needs related to interpretability and computational demands.
- **Dataset-Specific Observations:** The models demonstrate an ability to

adapt to the diverse linguistic patterns encountered, from the broad scope of Dauren Chapaev's dataset to the specific challenges posed by Azamat Serek's literary work. The latter scenario particularly demonstrates the effectiveness of the Rule-Based approach in contexts where capturing emotional depth is crucial.

- **Model Selection:** The study illustrates the effectiveness of both Rule-Based techniques and machine learning models in their respective domains. Ensemble methods, in particular, show promise due to their adaptability and robust performance across different types of text.

Exploratory Analysis: Impact of Edit Distance on Rule-Based Sentiment Analysis

In an exploratory extension of our existing methods, we incorporated edit distance to correct frequent typographical errors in the Kazakh text. This subsection discusses the potential improvements that edit distance could bring to the rule-based sentiment analysis if it were to be integrated systematically.

Corrective Actions Through Edit Distance

The application of edit distance specifically targeted common misspellings that could significantly affect sentiment interpretation. For instance:

- 'жаксы' was frequently misspelled, and using edit distance, it was consistently corrected to 'жақсы' (Edit distance: 1).
- These corrections were applied across several syntactic structures:
  - **[ADJECTIVE] + [Negation] + [NOUN]**: Corrected from 'жаксы емес адам' to 'жақсы емес адам' — a crucial adjustment ensuring the negative sentiment was accurately captured.
  - **[NOUN] + [ADJECTIVE]**: 'өте жаксы' corrected to 'өте жақсы' — maintained the positive sentiment by accurately identifying the adjective despite a higher edit distance (2).

Future Implications of Edit Distance Integration

While the current dataset analysis was conducted without edit distance, these preliminary results demonstrate its potential efficacy. Integrating edit distance more comprehensively in the future could address broader typographical and lexical variations, thereby increasing the robustness and accuracy of sentiment analysis, especially for under-resourced languages like Kazakh.

Considerations for Future Sentiment Analysis

The initial findings suggest significant promise for incorporating edit distance into rule-based methods to enhance their precision. As we consider expanding this approach to full-sentence analyses, this method may offer a valuable tool in refining sentiment analysis across diverse text types in Kazakh.

Considerations for Sentiment Analysis in Kazakh

This study highlights the critical role of model and preprocessing technique selection in sentiment analysis for the Kazakh language. The integration of Deci-

sion Trees, Random Forest, and XGBoost has demonstrated robust performance across a variety of text types, showcasing the effectiveness of ensemble methods in accommodating the linguistic diversity found in Kazakh literary and journalistic expressions.

Furthermore, the value of Rule-Based approaches, particularly when enhanced with linguistic-based corrections such as those provided by edit distance, cannot be underestimated. These approaches excel at capturing nuanced emotional content that is often vital for precise sentiment analysis. By synergizing the strengths of machine learning models with the precision of rule-based methods, more adaptable and nuanced analysis strategies can be developed.

This combined approach underlines the necessity for flexible and sophisticated analysis techniques to navigate the wide range of expressions encountered in the Kazakh language, from formal journalism to informal social media texts. As we consider future research directions, systematic integration of edit distance into sentiment analysis workflows promises to refine the accuracy of sentiment detection further and expand the linguistic capabilities of rule-based models.

Chapter 5

Conclusion and future work

5.1 Conclusions

The research presented in this thesis offers a comprehensive examination of sentiment analysis techniques for the Kazakh language, an underrepresented linguistic domain in computational linguistics. Through the successful integration of edit distance algorithms into rule-based methods, this work has not only enhanced the precision of these approaches, particularly in correcting typographical errors and normalizing text variations, but has also adeptly addressed the unique morphological challenges of Kazakh, providing a more reliable sentiment analysis framework.

The completion of the research objectives is clearly evidenced by the substantial advancements in our understanding and capability in applying both rule-based and machine learning methodologies. The objectives set out to assess the effectiveness, adaptability, and precision of these methods in processing diverse Kazakh texts, and these have been fully met as demonstrated by the empirical results. Ensemble methods like Random Forest and XGBoost, along with the improved rule-based approaches, have shown strong performance across various text types, thereby underlining their adaptability to the linguistic diversity of Kazakh texts.

However, this analysis also highlighted notable limitations in the performance of the rule-based model, particularly in contexts beyond literary texts, revealing the challenges in capturing a broad spectrum of sentiment expressions and language nuances. These findings underscore the need for ongoing enhancements and underscore that while the initial objectives have been achieved, the door is open for future advancements.

5.2 Future work

Future work will focus on expanding the application of edit distance to full-sentence analyses and exploring its integration with machine learning models to create hybrid approaches. Prospective studies might involve integrating sentiment dictionaries or domain-specific lexicons tailored to the diverse language styles found beyond literary domains. These initiatives are expected to further refine the efficacy of sentiment analysis tools, making them more adaptable to a wider range of

sentiment expressions and linguistic subtleties. This ongoing work will continue to promote linguistic inclusivity and broaden the applicability of sentiment analysis technology across various domains.

Bibliography

[1] N. Bele, P. K. Panigrahi, and S. K. Srivastava. "Political Sentiment Mining: A New Age Intelligence Tool for Business Strategy Formulation". In: *IGI Global* (Jan. 2017).

[2] O. Adwan et al. "Twitter Sentiment Analysis Approaches: A Survey". In: (Aug. 2020).

[3] H. Pandey, A. K. Mishra, and N. Kumar. "Various Aspects of Sentiment Analysis: A Review". In: *Social Science Research Network* (Jan. 2019).

[4] M. R. Yaakub, M. I. A. Latiffi, and L. S. Zaabar. "A Review on Sentiment Analysis Techniques and Applications". In: (Aug. 2019).

[5] C. N. Dang, M. N. M. García, and F. D. L. Prieta. "Sentiment Analysis Based on Deep Learning: A Comparative Study". In: *Electronics* 9.3 (Mar. 2020), pp. 483–483.

[6] E. Gogia. "A Review on Sentiment Analysis Techniques and Applications". In: (Dec. 2021).

[7] "The evolution of sentiment analysis - A review of research topics, venues, and top cited papers". In: (Feb. 2018).

[8] J. Cui et al. "Survey on sentiment analysis: evolution of research methods and topics". In: (Jan. 2023).

[9] R. Feldman. "Techniques and Applications for Sentiment Analysis". In: *ACM Transactions on Intelligent Systems and Technology* 4.1 (Apr. 2013). DOI: 10.1145/2436256.2436274. URL: https://dl.acm.org/doi/10.1145/2436256.2436274.

[10] Soujanya Poria et al. "Sentic Patterns: Dependency-based Rules for Concept-level Sentiment Analysis". In: *Knowledge-Based Systems* 69 (Oct. 2014). DOI: 10.1016/j.knosys.2014.05.018. URL: https://www.sciencedirect.com/science/article/abs/pii/S095070511400183X.

[11] H. R. Sankar and V. Subramaniyaswamy. "Investigating Sentiment Analysis Using Machine Learning Approach". In: *International Journal of Scientific Research in Network Security and Communication* 5.1 (Dec. 2017). DOI: 10.1109/iss1.2017.8389293. URL: https://doi.org/10.1109/iss1.2017.8389293.

[12] B. K. Bhavitha, A. P. Rodrigues, and N. N. Chiplunkar. "Comparative Study of Machine Learning Techniques in Sentiment Analysis". In: *2017 International Conference on Inventive Computation Technologies (ICICT)* (Mar. 2017). DOI: 10.1109/icicct.2017.7975191. URL: https://doi.org/10.1109/icicct.2017.7975191.

[13] "Sentiment Analysis Using Lexicon and Machine Learning-Based Approaches: A Survey". In: (Jan. 2018). URL: https://www.researchgate.net/publication/324626996_Sentiment_Analysis_Using_Lexicon_and_Machine_Learning-Based_Approaches_A_Survey.

[14] P. Soujanya et al. "Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research". In: (May 2020). DOI: 10.48550/arXiv.2005.00357.

[15] K. R. Mabokela, T. Celik, and M. Raborife. "Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape". In: *IEEE Access* 11 (Jan. 2023), pp. 15996–16020.

[16] S. Poria et al. "Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research". In: *IEEE Transactions on Affective Computing* 14.1 (Jan. 2023), pp. 108–132.

[17] T. Al-Moslmi et al. "Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review". In: *IEEE Access* 5 (Jan. 2017), pp. 16173–16192.

[18] R. Hameed, S. Ahmadi, and F. Daneshfar. "Transfer Learning for Low-Resource Sentiment Analysis". In: (Apr. 2023).

[19] X. Yu et al. "Beyond Counting Datasets: A Survey of Multilingual Dataset Construction and Necessary Resources". In: (Jan. 2022). DOI: 10.18653/v1/2022.findings-emnlp.273.

[20] N. L. France et al. *Computational Linguistics*. https://en.wikipedia.org/wiki/Computational_linguistics. Accessed: 2024-04-14. Sept. 2001.

[21] W. H. Gomaa and A. A. Fahmy. "A Survey of Text Similarity Approaches". In: *International Journal of Computer Applications* (Apr. 2013). DOI: 10.5120/11638-7118.

[22] D. Soyusiawaty and F. Rahmawanto. "Similarity Detector on the Student Assignment Document Using Levenshtein Distance Method". In: *International Conference on Information Systems for Research, Technology and Innovation (ISRITI)* (Nov. 2018). DOI: 10.1109/isriti.2018.8864339.

[23] Y. Wang, M. Wang, and W. Xu. "A Sentiment-Enhanced Hybrid Recommender System for Movie Recommendation: A Big Data Analytics Framework". In: *Big Data* (Jan. 2018). DOI: 10.1155/2018/8263704.

[24] Banu Yergesh et al. "Semantic hyper-graph based representation of nouns in the Kazakh language". In: *Computacion y Sistemas* 18.3 (2014), pp. 627–635.

[25] Narynov Sergazy Sakenovich and Arman Serikuly Zharmagambetov. "On one approach of solving sentiment analysis task for Kazakh and Russian languages using deep learning". In: *Computational Collective Intelligence: 8th International Conference, ICCCI 2016, Halkidiki, Greece, September 28-30, 2016. Proceedings, Part II 8*. Springer. 2016, pp. 537–545.

[26] Banu Yergesh, Gulmira Bekmanova, and Altynbek Sharipbay. "Sentiment analysis of Kazakh text and their polarity". In: *Web Intelligence*. Vol. 17. 1. IOS Press. 2019, pp. 9–15.

[27] Yergesh B. et al. "Ontology-based sentiment analysis of Kazakh sentences". In: *Computational Science and Its Applications–ICCSA 2017: 17th International Conference, Trieste, Italy, July 3-6, Proceedings, Part III*. Springer, 2017, pp. 669–677.

[28] Z. Li et al. "AgglutiFiT: efficient low-resource agglutinative language model fine-tuning". In: *IEEE Access* 8 (Jan. 2020), pp. 148489–148499.

[29] D. Kurian and et al. "Big data sentiment analysis using Hadoop". In: *International Journal of Innovative Research in Science and Technology* 1.11 (Jan. 2015), pp. 92–96.

[30] K. Niyazmetova et al. "Formation of a Database For Sentiment Analysis of Texts in the Uzbek Language". In: *Sci. Innov.* 2.C11 (Jan. 2023), pp. 20–23.

[31] J. Tussupov et al. "Development and implementation of a morphological model of Kazakh language". In: *Eurasian Journal of Mathematical and Computer Applications* 3.3 (Jan. 2015), pp. 69–79.

[32] Aigul K Zhumabekova and Leila Yu Mirzoyeva. "Peculiarities of indirect translation from English into Kazakh via Russian language". In: *TOJET* (2016).

[33] P. P. Surya and B. Subbulakshmi. "Sentimental analysis using Naive Bayes classifier". In: *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)* (Jan. 2019), pp. 1–5.

[34] T. Hariguna, W. M. Baihaqi, and A. Nurwanti. "Sentiment analysis of product reviews as a customer recommendation using the naive Bayes classifier algorithm". In: *International Journal of Informatics and Information Systems* 2.2 (Jan. 2019), pp. 48–55.

[35] Dauren Chapayev. *Open Access Kazakh News Sentiment Labeled Dataset*. https://github.com/chapayevdauren/sentiment-analysis-for-kz/blob/master/data/sample.csv. Year.

[36] A. Serek, A. Issabek, and A. Bogdanchikov. "Distributed sentiment analysis of an agglutinative language via Spark by applying machine learning methods". In: *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*. IEEE. 2019, pp. 1–4.

[37] Mark Lavelle. *Sentiment Labeled Sentences Data Set of Product Reviews*. https://www.kaggle.com/datasets/marklvl/sentiment-labelled-sentences-data-set?rvi=1.

[38] *Sozdikqor.kz: Comprehensive Kazakh Language Portal for Diverse Word Meanings and Phrases*. https://sozdikqor.kz/.

[39] H. D. Abubakar, M. Umar, and M. A. Bakale. "Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec". In: *SLU Journal of Science and Technology* 4.1 & 2 (2022), pp. 27–33.

[40] R. Goyal. "Evaluation of rule-based, CountVectorizer, and Word2Vec machine learning models for tweet analysis to improve disaster relief". In: *2021 IEEE Global Humanitarian Technology Conference (GHTC)*. IEEE. Oct. 2021, pp. 16–19.

[41] S. E. Saad and J. Yang. "Twitter sentiment analysis based on ordinal regression". In: *IEEE Access* 7 (2019), pp. 163677–163685.

[42] M. Abbas et al. "Multinomial Naive Bayes classification model for sentiment analysis". In: *IJCSNS Int. J. Comput. Sci. Netw. Secur* 19.3 (2019), p. 62.