electronics

MDPI

Article

Comparison of Machine Learning Approaches for Sentiment Analysis in Slovak

Zuzana Sokolová, Maroš Harahus *, Jozef Juhár, Matúš Pleva, Ján Staš and Daniel Hládek

Department of Electronics and Multimedia Telecommunications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 040 01 Košice, Slovakia; zuzana.sokolova@tuke.sk (Z.S.); jozef.juhar@tuke.sk (J.J.); matus.pleva@tuke.sk (M.P.); jan.stas@tuke.sk (J.S.); daniel.hladek@tuke.sk (D.H.)
* Correspondence: maros.harahus@tuke.sk

Abstract: The process of determining and understanding the emotional tone expressed in a text, with a focus on textual data, is referred to as sentiment analysis. This analysis facilitates the identification of whether the overall sentiment is positive, negative, or neutral. Sentiment analysis on social networks seeks valuable insight into public opinions, trends, and user sentiments. The main motivation is to enable informed decisions and an understanding of the dynamics of online discourse by businesses and researchers. Additionally, sentiment analysis plays a vital role in the field of hate speech detection, aiding in the identification and mitigation of harmful content on social networks. In this paper, studies on the sentiment analysis of texts in the Slovak language, as well as in other languages, are introduced. The primary aim of the paper, aside from releasing the "SentiSK" dataset to the public, is to evaluate our dataset by comparing its results with those of other existing datasets in the Slovak language. The "SentiSK" dataset, consisting of 34,006 comments, was created, specified, and annotated for the task of sentiment analysis. The proposed approach involved the utilization of three datasets in the Slovak language, with nine classification methods trained and compared in two defined tasks. For the first task, testing on the "SentiSK" and "Sentigrade" datasets involved three classes (positive, neutral, and negative). In the second task, testing on the "SentiSK", "Sentigrade", and "Slovak dataset for SA" datasets involved two classes (positive and negative). Selected models achieved an F1 score ranging from 75.35% to 95.04%.

Keywords: hate speech; Scikit-learn; sentiment analysis; SlovakBERT; Slovak language; social media

check for updates

Citation: Sokolová, Z.; Harahus, M.; Juhár, J.; Pleva, M.; Staš, J.; Hládek, D. Comparison of Machine Learning Approaches for Sentiment Analysis in Slovak. *Electronics* **2024**, *13*, 703. https://doi.org/10.3390/electronics13040703

Academic Editors: Arkaitz Zubiaga and Xi Peng

Received: 15 December 2023
Revised: 20 January 2024
Accepted: 5 February 2024
Published: 9 February 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Sentiment analysis (SA), a pivotal field in natural language processing (NLP), falls into the area of hate speech and offensive language detection. Pang et al. [1] described sentiment analysis as the process of identifying and evaluating emotional tones in text to understand and assess opinions and attitudes. SA has gained significant prominence in recent years for its ability to decipher and understand human emotions expressed in textual data. As communication increasingly permeates digital platforms, analyzing sentiments has become essential for businesses, researchers, and organizations aiming to comprehend public opinions, customer feedback, and societal trends. SA identifies the sentiment expressed in the text and then analyzes it. Sentiment analysis aims to find opinions, identify the feelings that people express, and then classify the polarity. The goal of sentence-level sentiment analysis is to classify the sentiment expressed in each sentence. Sentence-level sentiment analysis determines whether the sentence expresses positive or negative opinions.

A survey paper by Medhat et al. [2] provided an overview of recent updates in sentiment analysis algorithms and applications, categorizing and summarizing fifty-four recently published and cited articles. The articles contributed to various SA-related fields, emphasizing the ongoing research opportunities in enhancing sentiment classification and feature selection algorithms. The study identified Naïve Bayes and Support Vector Machines as frequently used machine learning algorithms for solving sentiment classification

problems and underscored the growing interest in languages other than English in SA research, emphasizing the need for resources and studies in these languages. Corso et al. [3] introduced an improved iteration of PArameter-free Solutions for Textual Analysis (PASTA), a distributed self-tuning engine. This enhanced version is applied to a collection of crisis tweets to automatically organize a corpus of tweets into cohesive and distinct clusters, requiring minimal intervention from analysts. The paper by Wankhade et al. [4] delved into sentiment analysis and its techniques, emphasizing the investigation of classification methods along with their pros and cons. It covered various levels of sentiment analysis, including procedures such as data collection and feature selection. The survey highlighted the prevalence of supervised machine learning methods, particularly Naïve Bayes (NB) and Support Vector Machines (SVM), due to their simplicity and high accuracy. It also addressed common application areas and explored the significance of challenges in sentiment analysis, including domain dependence. The paper by Jiang et al. [5] introduced a significant innovation by employing different data ratios for concurrent comparisons with multiple methods. This approach allowed for assessing performance across varying amounts of data. When dealing with limited data, machine learning demonstrated effective performance, while superior results were achieved with deep learning when larger datasets were utilized in their experiments. Notably, the use of Bidirectional Recurrent Neural Networks (BiRNNs) yielded the most favorable outcomes compared to the other methods under consideration. The study thus highlighted the nuanced impact of data ratios on the performance of different methods in the context of the research. The paper by Vigna et al. [6] aimed to address and prevent the spread of hate campaigns on public Italian pages, using Facebook as a reference. Consequently, the authors introduced hate categories to distinguish different types of hate. Comments were annotated by human annotators based on a defined taxonomy. Two classifiers for the Italian language, employing different learning algorithms, were then designed and tested for their effectiveness in recognizing hate speech on the first manually annotated Italian Hate Speech Corpus of social media text.

Quite often, sentiment analysis is used for marketing purposes. Several research findings indicate that stock market prices do not adhere to a random walk and, to some extent, can be predicted [7–11]. For instance, Gruhl et al. [12] demonstrated the correlation between online chat activity and the prediction of book sales. Liu et al. [13] utilized a Probabilistic Latent Semantic Analysis (PLSA) model to extract sentiment indicators from blogs in order to forecast future product sales. In addition, Mishne and Rijke [14] employed evaluations of blog sentiment to anticipate the sales of movies. This meant that the company had a portal and could access reviews under the products analyzed. This way, the company could find out what people thought about its products, what it needed to improve, etc. Politicians use a similar tactic when they analyze the comments under their posts, for example, on social networks, finding out what the voters' preferences and opinions are [15–17].

In our study, we conducted a comprehensive comparison of sentiment analysis methodologies in the Slovak language, assessing precision, accuracy, F1 scores, recall rates, and training times across diverse language models. We examined rule-based systems, traditional machine learning algorithms, and advanced deep learning models for their efficacy in handling Slovak linguistic nuances. The evaluation was performed on three distinct Slovak text datasets. In the introduction, we elucidated the diverse applications of sentiment analysis and its pivotal role in the broader field of natural language processing. Subsequently, we delve into a comprehensive review of related works, highlighting key advancements in sentiment analysis and providing a contextual backdrop for our study. Following this, we introduce the methodologies and datasets integral to our experimental framework, offering transparency regarding our chosen approaches and data sources. In the final stages of our study, we present a detailed examination of the achieved results and conduct a comparative analysis, providing valuable insights into the relative efficacy of the applied methods. Through this structured approach, our study contributes to the evolving discourse on sentiment analysis methodologies and their practical implications.

The paper's Introduction aimed to familiarize readers with the issue of sentiment analysis, and we provide a description of our motivation and goals in Section 1. Subsequently, Section 2 presents individual related works, while Section 3 characterizes the research methodology. In Section 4, the proposed approach, training datasets, and utilized methods are described. Section 5 provides a detailed account of the experiments and results for the two solved tasks. Following this, Section 6 features a discussion of the achieved results and a comparison. The final Section 7 draws conclusions and proposes the next direction for our work.

Motivation

Sentiment analysis is a technical tool through which we can better understand and respond to the thoughts, feelings, and opinions of individuals and societies [1,18]. It has far-reaching implications across various domains, offering improved decision-making, enhanced user experiences, and the potential to drive positive social change [19]. Therefore, investing in sentiment analysis is not only good but essential in our data-driven world [20,21].

To our knowledge, there are only a few works that have reported results in this area. In our opinion, there is still room for improvement in this task. The primary scientific contribution and motivation of this paper resides in the release of a novel dataset in the Slovak language. The dataset underwent annotation by a specifically chosen scientific cohort. Notably, it stands as the most extensive dataset documented to date within the domain of research centered on the Slovak language. The "SentiSK" dataset was systematically subjected to experimental evaluations, as were the other selected datasets in the study. For this reason, we examined publicly available datasets in the Slovak language and applied several of the classification methods that are most often used to classify texts for sentiment analysis tasks. Subsequently, we compared the achieved results with those of other studies that have dealt with this issue for the Slovak language as well as other languages.

The aim of this study, excluding the publication of the "SentiSK" dataset to the public, was to compare the results of our dataset with those of other existing datasets in the Slovak language. This comparison was intended to determine how our data stand and whether the direction taken in data collection was correct.

2. Related Work

In this section, we provide an in-depth review of the most recent scholarly papers closely tied to sentiment analysis, both within the context of the Slovak language and across various languages. Our exploration encompasses cutting-edge contributions, methodologies, and insights, aiming to offer a comprehensive overview of the latest developments in sentiment analysis research. By synthesizing these contemporary works, our objective is to contribute to the ongoing academic discourse and provide readers with an updated perspective on sentiment analysis advancements in diverse linguistic contexts.

2.1. Related Works Focused on Other Languages

Wankhade et al. [4] conducted a comprehensive exploration of sentiment analysis and related techniques, aiming to assess and augment classification methods in sentiment analysis by outlining their pros and cons. The study covered multiple levels of sentiment analysis, providing a brief overview of essential procedures such as data collection and feature selection. The authors specifically emphasized the application of sentiment analysis in analyzing e-shop reviews. The classification methods employed included NB (89.05% accuracy in K-fold cross-validation), SVM (81.01% accuracy), Relief (82.03% accuracy), and hybrid algorithms (approaching 84% accuracy).

Recently, Liang et al. [22] introduced Sentic GCN, a graph convolutional network for aspect-based sentiment analysis over dependency trees. Their model utilized SenticNet to capture aspect-specific sentence affective dependencies, integrating affective knowledge to enhance sentence dependency graphs. The improved affective graph model considered

dependencies between context and aspect words, along with affective information between opinion words and aspects. The experimental results obtained using public datasets demonstrated the superior performance of their approach compared to existing state-of-the-art methods.

Habernal et al. [23] conducted research on the sentiment analysis of Czech social media using controlled machine learning methods. They built a sizable Facebook dataset with 10,000 posts, each annotated by humans. Through a comprehensive evaluation of state-of-the-art features, classifiers, language-specific preprocessing techniques, and feature selection algorithms, they surpassed the baseline (unigram feature without preprocessing). The combination of features (unigrams, bigrams, POS features, emoticons, and character n-grams) and preprocessing techniques (unsupervised stemming and phonetic transcription) led to an impressive F score of 69% in three-class classification.

Karthika et al. [24] focused on classifying customer-provided star ratings for products, employing the Random Forest algorithm for the classification process. They compared the accuracy of Random Forest (RF) with the Support Vector Machine (SVM) algorithm, finding that RF achieved superior accuracy at 97% compared to SVM's 92%. Additionally, they measured the Receiver Operating Characteristic (ROC) curve for multiclass classification. The proposed system utilized a dataset sourced from Kaggle.com, comprising 20,000 records.

Chetviorkin and Loukachevitch [25] conducted a sentiment analysis study focusing on the Russian language. Their report stemmed from two evaluations of sentiment analysis systems conducted during the period 2011–2012 as part of the Russian seminar on information retrieval, ROMIP. The research addressed three specific tasks. The first task involved classifying the sentiment of user reviews in three domains (movies, books, and digital cameras) using various sentiment scales. The second task centered on the sentiment classification of news-based opinions, extracted from direct or indirect speech fragments within news articles. The third and final task focused on the query-based retrieval of opinionated blog posts across three domains (movies, books, and digital cameras). In the sentiment analysis task, they achieved an accuracy of 61.6% and an F1 score of 62.1% for the classification of three classes.

Rotim and Šnajder [26] tackled the challenge of short-text sentiment classification in Croatian, employing two straightforward yet effective methods: word embeddings and string kernels. In their experiment, they utilized multiple SVM models with varied preprocessing techniques and kernels, comparing them across three datasets with distinct characteristics. Their dataset, comprising around 12,824 tweets categorized as positive, negative, and neutral, served as the basis for evaluation. The researchers concluded that word embeddings proved to be the method of choice for short-text sentiment classification in Croatian. They also noted that in situations where word embeddings were not applicable, a bag of words with simple stemming emerged as the preferred alternative.

Kapočiūtė-Dzikienė et al. [27] addressed the task of the sentiment analysis of Internet comments in Lithuanian, employing two distinct approaches: knowledge-based and machine learning. They highlighted that, for the knowledge-based approach, adjectives, nouns, and verbs (excluding adverbs) were identified as the most significant sentiment indicators. The dataset utilized consisted of 58,129 comments. In their experiment, the highest accuracy of 67.9% was achieved through the implementation of Multinomial Naïve Bayes. This model utilized token unigrams and bigrams as features, with diacritics replacement applied as a preprocessing technique. A summary of this research, as well as others in Section 2.1, can be found in Table 1.

Table 1. Summary of papers mentioned in Section 2.1.

Paper	Strengths	Weaknesses	Opportunities	Challenges
[4]	Comprehensive exploration of sentiment analysis techniques. High accuracy in classification methods. Application to e-shop reviews.	Limited to specific techniques and domains. May not address all linguistic nuances.	Further refinement of classification methods. Application in different domains and languages.	Generalizing methods to diverse datasets and languages.
[22]	Introduction of Sentic GCN, a novel model. Integration of affective knowledge. Superior performance on public datasets.	Complexity of the model may limit wider application. Specific to aspect-based sentiment analysis.	Extending the model to other types of sentiment analysis. Application to different languages.	Ensuring model adaptability to various contexts.
[23]	Focused on Czech social media. Comprehensive feature evaluation. Effective combination of features and preprocessing.	Specific to Czech language and social media. May not generalize to other domains.	Adapting methodology to other languages and platforms.	Balancing feature selection and preprocessing for different languages.
[24]	High accuracy with Random Forest. Comparison with SVM. Utilization of ROC curve for evaluation.	Limited to product star ratings. Dataset sourced from a specific platform (Kaggle).	Expanding the approach to other types of reviews and datasets.	Ensuring the robustness of the model across diverse product categories.
[25]	Focus on Russian language. Multiple sentiment analysis tasks. Evaluation of sentiment in diverse domains.	Limited to specific domains and tasks. Accuracy and F1 score might be improved.	Applying methodology to other languages. Exploration of different sentiment scales.	Enhancing accuracy and adaptability to different domains.
[26]	Focus on Croatian short-text classification. Effective use of word embeddings and string kernels. Comparison of different SVM models.	Specific to short texts and Croatian language. Limited to three datasets.	Extending the methods to longer texts and other languages.	Adapting the models to diverse linguistic features and datasets.
[27]	Two distinct approaches: knowledge-based and machine learning. Focus on Lithuanian Internet comments. High accuracy with Multinomial Naïve Bayes.	Limited to Internet comments. Specific to Lithuanian language.	Expanding approaches to other types of content and languages.	Adapting methodologies to capture linguistic nuances in different languages.

2.2. Related Works Focused on Slovak Language

Krchnavy and Simko [28] addressed the challenge of conducting sentiment analysis on social media posts in Slovak. The authors aptly noted various difficulties inherent to working with the Slovak language, such as high inflection, complex morphology, and syntax. Additionally, the user-generated content on social networks introduced further complications, including variability in diacritics, inconsistent style, and a high error rate. As part of their study, the authors proposed a machine-learning-based method for the sentiment analysis of Facebook posts, incorporating multilevel text preprocessing. They curated their own "Sentigrade" dataset, detailed in Section 4.1. Their results demonstrated an achievement comparable to sentiment analysis in other languages worldwide.

Mojžis et al. [29] established robust foundations for document-level sentiment analysis in languages with extensive inflection, such as Czech and Slovak. Their research highlighted the necessity of eliminating a substantial number of duplicates from the dataset prior to analysis. Additionally, the authors clarified that features associated with part-of-speech tagging do not negatively impact sentiment analysis. Furthermore, they contested the suitability of the chi-squared metric for feature selection in the context of sentiment analysis.

Pecar et al. [30] tackled the problem of sentiment classification for the Slovak language, which suffers mainly from low-resource datasets. They introduced several neural model architectures employing state-of-the-art techniques for sentiment analysis. In their experiment, they combined different types and sizes of recurrent layers (1 LSTM, 1 Bi-LSTM) with or without the use of the attention layer. They alternated four different word representations (randomly initialized embedding layer—LookUp; deep contextualized word representations—ELMo; fastText; and word2vec). They worked with two different datasets:

Reviews3, which contained 5320 customer reviews and achieved an F1 score of 81.32%, and Twitter [31], which contained 50,710 tweets and achieved an F1 score of 69.78%.

Machová et al. [32] employed an optimization method to iteratively label all words in a lexicon, assessing the efficacy of opinion classification with the lexicon until optimal labels for the words were identified. They also addressed the limitation of lexicon-based approaches in classifying opinions in texts devoid of lexicon words. To overcome this, they introduced an auxiliary approach based on machine learning. Their findings indicated that the hybrid approach successfully classified over 99% of texts, outperforming the original lexicon-based method. The study utilized the "Slovak dataset for sentiment analysis", as discussed in Section 4.1.

The same dataset was also used in the study by Mikula et al. [33]. Their approach involved an automated annotation of a dictionary derived from English, utilizing Particle Swarm Optimization (PSO) and Bare-bones Particle Swarm Optimization (BBPSO) to assign polarity values to the dictionary words. Two dictionaries were created for the experiment: the first was generated, translated, and refined by a human, while the second was derived from six English dictionaries and translated into Slovak using the first dictionary. Both dictionaries underwent labeling by a human annotator, PSO, and BBPSO. The labeled versions were then applied to sentiment analysis in the Slovak language, and their performances were compared. The results showed that the versions annotated by PSO and BBPSO surpassed the human-labeled version, indicating better polarity values for the words in the dictionaries. They achieved an F1 score of 77.52% for the first dataset and 74.27% for the second dataset.

Pikuliak et al. [34] developed a language model named SlovakBERT, utilizing the RoBERTa architecture [35] and trained on a web-crawled corpus. This model addressed various tasks including part-of-speech tagging, semantic textual similarity, sentiment analysis, and document classification. Notably, the model achieved a high accuracy of 98.37% for part-of-speech tagging. The dataset incorporated data from diverse sources such as Wikipedia, Open Subtitles, the OSCAR corpus, and Slovak websites, totaling 19.35 GB of clean text without HTML tags. A BPE tokenizer with a vocabulary size of 50,264 was employed, and the model underwent training for 300k steps with a batch size of 512, limiting samples to a maximum of 512 tokens. The Adam optimization algorithm and Hugging Face Transformers were used in the training process. For sentiment analysis, the model achieved an F1 score of 67.2% using a dataset comprising 41,084 tweets, including 11,160 negative, 6668 neutral, and 23,256 positive samples. SlovakBERT demonstrated state-of-the-art performance in these tasks, and the authors released fine-tuned models for the Slovak community. They highlighted that while existing multilingual models may yield comparable results in certain tasks, they are less efficient in terms of memory and computation resources.

Koncz and Paralic [36] provided an overview of prevailing approaches in sentiment analysis, with a particular focus on feature selection methods. They introduced a novel approach to feature selection, which underwent experimental evaluation using a movie review dataset comprising 2000 documents. The results indicated that their proposed method demonstrated computational efficiency while only marginally sacrificing accuracy, achieving accuracies ranging from 50% to 90%. In their subsequent studies [37,38], these authors summarized methods aimed at enhancing the effectiveness of dataset annotation, crucial for sentiment analysis. Given the dependency of sentiment analysis on manually annotated datasets, these datasets play a vital role in training and evaluating machine-learning-based methods, as well as evaluating dictionary-based methods. The authors employed techniques such as automatic corpora creation, active learning, and annotation suggestions to improve the annotation process.

Mozetič et al. [31] conducted an analysis of an extensive set of sentiment-labeled tweets categorized into three groups: negative, neutral, or positive. Their primary objective was to utilize consistent evaluation measures to assess both the quality of human annotations and the effectiveness of classification models. They argued that the performance of a

classification model is predominantly constrained by the quality of the labeled data, and this quality can be gauged through agreement among human annotators. In their study, they opted for various Support Vector Machine (SVM) variants as classification models, with Naïve Bayes serving as a reference. The research encompassed two corpora, with particular interest in the first dataset, which comprised 1.6 million annotated tweets in multiple languages, including Slovak; English; Albanian; German; Portuguese; Hungarian; Ser/Cro/Bos (a combined set of Serbian, Croatian, and Bosnian tweets, which are challenging to distinguish on Twitter); Russian; Bulgarian; Polish; Spanish; Swedish; and Slovenian. Focusing on the Slovak subset of this dataset containing 70,425 tweets, the achieved accuracy was 76.2%, with an F1 score of 77.2%. A summary of this research, as well as others in Section 2.2, can be found in Table 2.

Table 2. Summary of papers mentioned in Section 2.2.

Paper	Strengths	Weaknesses	Opportunities	Challenges
[28]	Addressed the unique challenges of the Slovak language in sentiment analysis. Proposed a machine learning method with multilevel text preprocessing.	Limited to social media posts, specifically Facebook.	Potential to adapt the method for other platforms and types of Slovak language content.	Managing high inflection, complex morphology, and syntax in Slovak.
[29]	Established foundations for document-level sentiment analysis in inflection-rich languages.	Study was specific to Czech and Slovak languages.	Extending research to other inflection-rich languages.	Balancing the removal of duplicates and maintaining data integrity.
[30]	Introduced neural model architectures for Slovak sentiment classification. Achieved high F1 scores.	Faced challenges due to low-resource datasets in Slovak.	Development of more advanced neural models for under-resourced languages.	Ensuring the effectiveness of models with limited data resources.
[32]	Used an optimization method for lexicon labeling and introduced a machine learning auxiliary approach.	Limited by the constraints of lexicon-based methods.	Opportunities for hybrid approaches combining lexicon-based and machine learning methods.	Overcoming limitations of lexicon-based sentiment analysis.
[33]	Automated annotation of dictionaries using optimization algorithms, improving polarity values.	Reliance on dictionaries translated from English to Slovak.	Application of these methods to other language pairs and broader sentiment analysis contexts.	Ensuring accuracy and relevance in automated dictionary translation.
[34]	Developed SlovakBERT with high accuracy in various tasks. Utilized a comprehensive and diverse training corpus.	Focused on Slovak language, limiting broader applicability.	Adaptation of SlovakBERT for other languages or more specialized tasks.	Managing resource efficiency compared to multilingual models.
[36]	Provided an overview of sentiment analysis approaches. Introduced an efficient feature selection method.	The proposed method had a range of accuracy (50–90%).	Potential for further refinement and application of their feature selection methods in other sentiment analysis contexts.	Balancing computational efficiency with accuracy.
[31]	Conducted extensive analysis of sentiment-labeled tweets in multiple languages. Demonstrated the importance of quality in labeled data.	The research was limited to Twitter data, which may not generalize.	Applying the evaluation methods to other platforms and types of sentiment data.	Ensuring the quality of data labeling across diverse languages.

2.3. Related Works Focused on Slovak Language in Our Research

In our initial study [39], we scrutinized the "SentiSK" dataset created for the purpose of sentiment analysis in the Slovak language. Our custom implementation of text representation in Python incorporated a convolutional neural network with elements of a recursive neural network. Additionally, we introduced and implemented an alternative approach for identifying the most hateful comments, utilizing a lexicon of expressive words. On a smaller test dataset, we attained a satisfactory accuracy of 61.32%. This outcome was

noteworthy when compared to contemporaneous research such as [40–42]. Upon finalizing the "SentiSK" dataset and fine-tuning the neural network parameters, we achieved an accuracy of 52.83%. The lower accuracy was attributed to the substantial representation of negative comments in the dataset.

Subsequently, in [43], our attention was directed toward the most recent advancements in hate speech and offensive language detection. We delineated the envisioned characteristics of our ongoing work in a new dataset. This upcoming dataset is intended to exhibit improved class balance, heightened accuracy, and more comprehensive annotation with additional subcategories. It is designed to incorporate emoticons and hashtags, given their significant role in discerning people's emotions and moods. Furthermore, the dataset aims to identify explicit language, encompassing the detection of hate speech, offensive language, toxicity, and related categories.

In our paper [44], we explored the application of the Fairseq toolkit for machine translation in the context of the Slovak language. Machine translation presents an opportunity to achieve substantial time savings by swiftly translating entire text documents. Additionally, it offers the advantage of significantly reduced costs due to decreased reliance on human involvement. Furthermore, machine translation exhibits the ability to memorize key terms, facilitating their reuse in various contexts. This tool has the potential to assist in diverse tasks such as text analysis, corpus creation, and the identification of specific text features.

Following this, our work in [45] delineated the process and criteria for systematically crafting a dataset in Slovak. This dataset was specifically designed to address tasks such as sentiment analysis and the identification of hate speech or offensive language in the Slovak language. We emphasized the imperative need to tailor the dataset creation methodology to the unique characteristics of each language. Languages typically exhibit variations in inflections, word forms, punctuation, complexity, verb tenses, and numerous other features. These distinctions necessitate careful consideration when dealing with text in a particular language.

In our recent paper [46], we conducted a successful comparison of classifier performance using datasets in Slovak, revealing noteworthy distinctions between the original and translated models. While translated datasets yielded relatively satisfactory results, the original Slovak datasets enabled the models to make more accurate predictions. This finding underscores the crucial significance of crafting and utilizing original datasets tailored to specific languages. Translated datasets, while useful in situations where original data are unavailable, may not fully capture the nuances and intricacies of the target language. A summary of this research, as well as others in Section 2.3, can be found in Table 3.

Table 3. Summary of papers mentioned in Section 2.3.

Paper	Strengths	Weaknesses	Opportunities	Challenges
[39]	Utilized convolutional and recursive neural networks for Slovak sentiment analysis. Implemented an approach for identifying hateful comments.	Lower accuracy in final dataset due to a high proportion of negative comments.	Further refinement of the model to improve accuracy.	Balancing the representation of different sentiments in the dataset.
[43]	Focused on advancements in hate speech and offensive language detection. Proposed characteristics for an improved dataset.	Ongoing work; dataset development and refinement are in progress.	Creation of a more balanced and comprehensive dataset with additional subcategories.	Integrating and accurately annotating emoticons, hashtags, and explicit language.
[44]	Explored Fairseq toolkit for machine translation in Slovak, offering potential time savings and reduced human involvement.	Focused primarily on machine translation, which might not directly address sentiment nuances.	Application of machine translation in text analysis, corpus creation, and feature identification.	Ensuring accurate translation while retaining key terms and context.

Table 3. *Cont.*

Paper	Strengths	Weaknesses	Opportunities	Challenges
[45]	Detailed the process of crafting a Slovak-specific dataset for sentiment analysis and hate speech detection.	Challenges in tailoring the dataset to the unique linguistic features of Slovak.	Opportunity to create language-specific datasets that better capture linguistic nuances.	Managing the complexity and variability of Slovak language features.
[46]	Conducted a comparison of classifier performance using Slovak datasets, highlighting the importance of original datasets.	Translated models might not capture all nuances of the target language.	Demonstrating the value of original language datasets for more accurate predictions.	Balancing the use of original vs. translated datasets and their respective accuracies.

3. Research Methodology

Sentiment analysis is a vibrant field with different methodologies, each dealing with different levels or aspects of sentiment analysis. Several of the most relevant of these methods are listed below.

- Aspect-Level Sentiment Analysis: This approach focuses on identifying sentiments related to specific entities or aspects in a document. It provides information about sentiment in different domains. The current solutions are categorized based on aspect detection, sentiment analysis, or both, using different algorithmic approaches in each case [47].
- Clustering-Based Approach: This sentiment analysis methodology uses techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) weighting and voting mechanisms. It differs from symbolic and supervised learning techniques in its efficiency and minimal human involvement [48].
- KNN Classifier-Based Approach for Multiclass Sentiment Analysis: This method uses a K-nearest neighbors algorithm to classify sentiment into multiple categories. It has been shown to be effective, especially when analyzing data from social media platforms such as X (Twitter), Facebook, and YouTube [49].
- Proximity-Based Sentiment Analysis: Proximity sentiment analysis is a natural language processing (NLP) technique that aims to understand sentiment in text by focusing on the proximity of words. It analyzes how the arrangement and proximity of words affect the overall sentiment expressed. One of the key aspects of this approach is the proximity distribution, which examines the relative distribution of words in a text. The frequent proximity of certain words can greatly affect the sentiment of a phrase. Another essential feature is the mutual awareness between different types of verbal proximity, for example, the proximity of adjectives to nouns or adverbs to verbs. It helps to understand how different word relationships contribute to the overall sentiment [50,51].
- Machine Learning Methodologies on Social Networks: This approach involves sentiment analysis on social networks using various machine learning techniques such as Naïve Bayes, Entropy max, and Support Vector Machines. It is particularly useful for sentiment analysis on large-scale social network data streams [52].
- Deep Transfer Learning for Sentiment Analysis: Deep Transfer Learning for Sentiment Analysis is a methodology that integrates advanced deep learning techniques such as BERT (Bidirectional Encoder Representations from Transformers) to perform sentiment analysis, especially on mental health texts. This approach excels in its ability to understand and represent emotional feelings along separate valence axes, which is highly effective for specialized domains such as mental health. By harnessing the power of deep learning, it can understand the nuances and complexities of emotional expression in texts that standard sentiment analysis methods might miss [53].
- Lexicon-Based Analysis on Multilingual Datasets: This method, which deals with sentiment analysis in multilingual datasets, involves the use of different lexicons adapted for each language. This approach is crucial for accurately capturing sentiment in different linguistic contexts, as each language has unique semantic and syntactic

structures that influence the way sentiment is expressed. The method is based on the principle that a one-size-fits-all approach is inadequate for multilingual sentiment analysis due to the large differences between languages [54].

This research focused on applying machine learning methods to social networks. Various methods were employed, as detailed in Section 4.2, across three datasets in the Slovak language, as outlined in Section 4.1.

4. Proposed Approach

The primary objective of our research was to conduct a comparison of existing datasets in the Slovak language for sentiment analysis, employing various classification algorithms and models. For the experimental solution, we used three datasets in the Slovak language. These datasets were chosen because, at the time of our experiential solution, there were no other datasets publicly accessible in the Slovak language. Additionally, we aimed to compare datasets that were not machine-translated and were very similar in structure and annotation. The selected datasets are described in Section 4.1. We used classification algorithms and models, which we describe in Section 4.2.

4.1. Training Datasets

As we mentioned earlier, we used three different datasets in our experimental solution. The first one was our dataset in the Slovak language, which we annotated manually. We named this dataset "SentiSK", and it contained 34,006 comments from the Facebook social network. The comments were obtained using a Python tool for scraping data from websites, and they were comments under the contributions of three Slovak politicians. The preprocessing of the data consisted of clearing the text of unwanted characters and deleting blank lines, empty spaces, dots, etc. For the preprocessing, we used the NLTK library. We annotated the data with the Prodigy annotation tool provided by our Department of Electronics and Multimedia Communications. We marked each comment as negative, neutral, or positive. In total, there were 20,655 negative, 9573 neutral, and 3778 positive comments in the dataset. Since we downloaded the data from the posts of Slovak politicians, there were a lot of negative comments. For this reason, the dataset was class-imbalanced.

Upon a brief exploration, we identified a freely accessible dataset [28] in the Slovak language available on the Internet. Its creators developed and utilized the dataset for a task involving multilevel text preprocessing to address the intricacies of user-generated social content. The original dataset was categorized into five groups. For ease of comparison with our dataset, we consolidated the "strongly negative" category with the "negative" category and merged the "strongly positive" with the "positive" category. The "neutral" category remained unchanged. In the final structure, the dataset comprised a total of 1584 comments, with 709 comments labeled as "negative", 573 as "positive", and 306 as "neutral". Termed "Sentigrade" by the authors, this dataset was curated from Seesame's Facebook pages and the associated comments on these pages.

The third dataset in this experiment was the "Slovak dataset for SA", which was created by Machova during her research [32]. The dataset is available online on the website [55]. The dataset contains 2669 negative and 2573 positive comments.

It is possible to notice that the datasets we obtained from the Internet were significantly smaller in terms of the number of comments. On the other hand, we can see that these datasets were more class-balanced. We describe the selected datasets in more detail in Table 4.

Table 4. Specifications of datasets.

Datasets	SentiSK [39]	Sentigrade [28]	Slovak Dataset for SA [55]
Number of comments	34,006	1584	5242
Number of categories	3	5	2

Table 4. *Cont.*

Datasets	SentiSK [39]	Sentigrade [28]	Slovak Dataset for SA [55]
Categories	- positive neutral negative -	strongly positive positive neutral negative strongly negative	- positive - negative -
Number of negative comments	20,668	709	2669
Number of neutral comments	9581	306	-
Number of positive comments	3779	573	2573
Number of words	356,300	32,788	155,522
Data source	Facebook	Facebook	E-shop reviews

4.2. Used Methods

We used the Scikit-learn library to train the models. Scikit-learn, often referred to as sklearn, is a widely used open-source machine learning library designed for the Python programming language. This library contains efficient implementations of many popular machine learning algorithms, like Support Vector Machines, Random Forests, and k-nearest neighbors [56]. In the following paragraphs, we briefly describe all the methods that we applied in our experimental solution.

The Random Forest Classifier (RFC) is as a machine learning algorithm tailored for classification tasks. During training, the algorithm constructs multiple decision trees and outputs the mode of classes (for classification) or an average prediction (for regression) derived from individual trees. Essentially, it amalgamates predictions from multiple decision trees to create a more accurate and robust model compared to a single decision tree. The algorithm operates on the principle of randomly selecting a subset of features and data samples for each tree, training them independently. This introduction of randomness mitigates overfitting, enhancing the model's generalization capability. In prediction, the algorithm aggregates the results from all trees to generate a final prediction [57].

- Strengths: RFCs are highly adaptable and can be utilized for diverse tasks like classification, regression, and outlier detection. Compared to other algorithms like decision trees, RFCs are more resistant to overfitting. This resilience stems from the aggregation of multiple tree predictions, thereby lowering the model's variance. Due to their proficiency in handling non-linear data, RFCs are adept at modeling complex variable relationships, a significant advantage in scenarios with non-linearly separable data. One of the key capabilities of RFCs is their ability to pinpoint crucial data features. This is beneficial for both understanding variable interconnections and for the process of feature selection.
- Weaknesses: In terms of interpretability, RFCs lag behind simpler algorithms such as decision trees. The complexity of being an ensemble model obscures the clarity of their predictive processes. When dealing with high-dimensional data, RFCs tend to be slower in both the training and prediction phases. The necessity of constructing numerous trees contributes to this computational burden. RFCs demand more extensive datasets for training in comparison to algorithms like decision trees. The requirement to build multiple trees means that they need more data to achieve effective generalization [58].

The Multilayer Perceptron (MLP) classifier is a machine learning algorithm for analyzing tasks. The algorithm is a type of neural network. It processes data in one direction, from

input to output. The MLP classifier consists of one or more nodes, each of which performs a weighted sum of inputs followed by a non-linear activation function. The output of each node is then transferred as input to the next layer until the final output layer is reached [59].

- Strengths: MLPs are capable of learning complex relationships between input and output variables, even when the relationships are non-linear. This makes them well-suited for tasks such as image recognition and natural language processing. MLPs can be trained on large datasets, which can help them to capture more complex patterns in the data.
- Weaknesses: MLPs can be prone to overfitting, which means that they may learn the training data too well and perform poorly on new data. This can be a problem when training on small datasets. They also require significant computational resources. MLPs can be computationally expensive to train, especially when using large datasets. This can make them impractical for some applications. The performance of an MLP can be sensitive to the choice of hyperparameters, such as the number of hidden layers and the number of neurons in each layer. This can make it difficult to tune MLPs to work well on a particular task [60].

Logistic Regression (LR) is a machine learning algorithm. The algorithm is based on the probability that an event will occur. Logistic Regression uses a logistic function (also known as a symbol function) to map the input characteristics to probability values between 0 and 1. If the probability is higher than the threshold value (usually 0.5), the event is expected to occur, and if it is not, the event is not expected to occur [61].

- Strengths: LR stands out for its simplicity and efficiency, making it easy to implement and train while being computationally light. It provides probabilistic outcomes, offering insights into the likelihood of each class for specific data points, which is particularly valuable in tasks like risk assessment. LR is effective in determining the relationship between input features and the target variable.
- Weaknesses: LR is built on the premise of a linear relationship between input and output variables, which can lead to subpar results in scenarios with non-linear relationships. In cases where input variables and the output have a non-linear relationship, LR may not perform as well as other algorithms capable of modeling non-linear relationships, such as decision trees or Support Vector Machines. LR's applicability is primarily limited to binary classification, making it less suitable for multiclass classification tasks where alternatives like Multinomial Logistic Regression or Random Forests may be more effective [62].

A Support Vector Classifier (SVC) searches for the best hyperplane to split input data into different classes. The hyperplane is selected to maximize the margin (i.e., the distance between the hyperplane and the closest data point of each class). The data points on the margin are called support vectors.

- Strengths: SVCs excel in high-dimensional spaces, outperforming traditional linear classifiers. Their efficiency in these complex spaces comes from the kernel trick, which effectively maps data into a higher-dimensional space for linear separation. They demonstrate remarkable versatility by handling both linear and non-linear separable data. The kernel trick equips SVCs with the flexibility to adapt to the data's structure, enhancing their resilience to noise and outliers. Known for their strong generalization capabilities, SVCs often perform well with unseen data. This strength is rooted in their training optimization process, which aims to balance minimizing training errors and model complexity.
- Weaknesses: Training SVCs can be computationally demanding, particularly for large datasets, due to the complex quadratic optimization problem involved in their training. The efficacy of SVCs heavily depends on the selection of the kernel function and hyperparameters, necessitating meticulous tuning for optimal results. While adept at binary classification, SVCs are not inherently designed for multiclass classification

tasks, making other algorithms like multiclass SVMs or Random Forests more suitable for such scenarios [63].

The K-Neighbors Classifier (KNN) classifies new data points based on the k-nearest neighbor class in the training data. During training, the K-Neighbors Classifier stores trained data points in a data structure that allows an efficient search for the nearest neighbor [64].

- Strengths: KNN is renowned for its simplicity, both in understanding and implementation, lacking a complex training phase, which makes it ideal for beginners in machine learning. As an instance-based learning algorithm, KNN does not build a conventional model but instead stores the entire dataset for classification tasks. This attribute allows it to scale well with large datasets. KNN's ability to adjust to new data and shifts in data distribution is a significant advantage, especially in scenarios where data are continually evolving.
- Weaknesses: Classifying new data points with KNN can be computationally heavy, particularly with larger datasets, due to the necessity of comparing each new point with all points in the training set. The algorithm's performance is highly sensitive to how the features are scaled. Incorrect scaling can severely impair KNN's effectiveness because it relies on distance measurements for determining similarity between data points. In high-dimensional spaces, KNN's efficacy diminishes. The exponential increase in potential neighbors as dimensions grow complicates the identification of the most relevant neighbors, often leading to suboptimal classification results [65].

The Multinomial NB (MNB) is based on the Bayes theorem. It assumes that the probability of a property for a class is independent of other properties. It models the probability of each class (e.g., different categories of text documents) based on the frequency of occurrence of different features (e.g., words or phrases) in the input data. It assumes that the frequency of different characteristics follows a multinomial distribution [66].

- Strengths: MNB stands out for its efficiency in both training and classifying data, making it particularly suitable for handling large datasets. It is highly effective for text classification tasks, capable of managing extensive feature spaces and adeptly capturing relationships between words within documents. MNB's ability to deal with large feature spaces is a critical advantage in text classification, where the number of unique words (features) can be exceedingly high.
- Weaknesses: A fundamental limitation of MNB is its assumption of feature independence. In real-world data, this assumption is often not met, potentially leading to subpar classification performance. The algorithm can struggle with imbalanced datasets where one class significantly outweighs others. MNB's reliance on feature frequency in class probability assignment can skew results in favor of the more prevalent class. MNB is primarily designed for categorical data, with features having a limited set of values. Its effectiveness diminishes with continuous data, which encompass a broader range of feature values [67].

Multilingual BERT (mBERT) is a variant of BERT that is pretrained on a large corpus of texts in multiple languages, allowing it to perform well on multilingual natural language processing tasks. mBERT is trained to jointly model language-specific and common features of multiple languages, allowing it to transfer knowledge between languages and improve performance in low-resource languages. The mBERT architecture is based on a transformational model. The transformation model uses attention mechanisms that allow the model to focus on different parts of the input sequence at each layer. The mBERT model has 110 million parameters [68].

- Strengths: mBERT is a versatile multilingual neural network model, proficient in processing and understanding text across multiple languages. This capability renders it highly valuable for applications like machine translation, cross-lingual sentiment analysis, and question answering in various languages. The model excels in grasping deep contextual relationships within sentences, enabling it to comprehend the contextual meaning of words. This feature is crucial for complex tasks such as natural language

understanding and effective machine translation. mBERT offers the flexibility to be fine-tuned for specific tasks and domains, allowing for customization to meet the specific requirements of different applications.

- Weaknesses: One major drawback of mBERT is its size and complexity, necessitating substantial computational resources for training and fine-tuning. This aspect can limit its practicality in some scenarios. The complexity of mBERT can also pose challenges in terms of understanding and implementation, potentially making it less accessible for certain users. The effective fine-tuning of mBERT demands specialized knowledge in machine learning and natural language processing, which can be a barrier for users without such expertise [68].

SlovakBERT is a variant of BERT that was specially trained for the Slovak language. SlovakBERT is trained to capture the context of Slovak words in sentences and can perform well in various natural language processing tasks. SlovakBERT has 110 million parameters for the largest version and 30 million for the basic version. SlovakBERT is an open-source model and can be fine-tuned to specific natural language processing tasks using relatively small amounts of data for specific tasks [34].

- Strengths: SlovakBERT's specialization in processing Slovak text, thanks to its training on a vast corpus of the Slovak language, enables it to adeptly capture linguistic nuances. This attribute makes it particularly effective for tasks like sentiment analysis, named entity recognition, and machine translation involving the Slovak language. The model is designed to understand contextual meanings within sentences, a key feature for comprehensive natural language understanding and accurate machine translation in Slovak. SlovakBERT offers adaptability, allowing for fine-tuning to cater to specific tasks and domains. This customization potential enables it to meet the unique needs of various applications involving the Slovak language.
- Weaknesses: Its exclusive focus on the Slovak language limits its applicability; SlovakBERT is not suitable for tasks involving languages other than Slovak. Due to its size and complexity, SlovakBERT demands significant computational resources for training and fine-tuning, which may render it impractical for certain applications or settings with limited resources. Achieving optimal performance on specific tasks with SlovakBERT requires meticulous fine-tuning, a process that can be both time-intensive and necessitates expertise in machine learning and natural language processing [34].

The purpose of the experiment was to compare the accuracy, precision, recall, F1 score, and running time of classification algorithms. The test machines included a CPU Core i7920 with 2.67 GHz, 32 GB of RAM, 2 NVIDIA GeForce 1080, and 12 GB of VRAM. We used Scikit to train statistics classifiers. Hugging-face transformers were used to fine-tune the neural language model.

5. Description of Experiments and Achieved Results

As part of the experimental solution, we worked on two tasks. In the first task (see Section 5.1), we compared our "SentiSK" dataset with the "Sentigrade" dataset. At first glance, we could see that the "SentiSK" dataset was much larger in terms of the number of comments. However, it was also evident that it was much less class-balanced than the "Sentigrade" dataset. In the second task (see Section 5.2), we compared all three Slovak datasets but did not include neutral comments, because there were no neutral comments available in the "Slovak dataset for SA". The specifications of the datasets are given in Table 4.

5.1. "SentiSK" vs. "Sentigrade"

In Table 5, we show the results achieved for the first task. We compared our "SentiSK" dataset with the "Sentigrade" dataset. The highest F1 scores were achieved using the SlovakBERT model, namely 71.53% for the "Sentigrade" dataset and 67.68% for the "SentiSK" dataset.

In contrast, the lowest F1 score in both datasets was achieved using the SVC and KNN classifiers. In terms of training speed, the KNN classifier was trained the fastest for both datasets. The data splitting in this code was implemented using the train_test_split function from the sklearn.model_selection library. This function is used to produce two sets of data: processed_features, which represent input features, and labels, which represent labels or target variables. The parameter test_size = 0.2 specifies that 20% of the total data will be split off and used as the test set. Cross-verification was not implemented. The precision–recall curves for the KNN classifier can be seen in Figure 1. The training of the mBERT model took the longest for both datasets. The accuracy, precision, recall, F1 score, and time values are presented in Table 5. In Table 6, we provide the results for each class for the "SentiSK" and "Sentigrade" datasets and SlovakBERT and mBERT models.

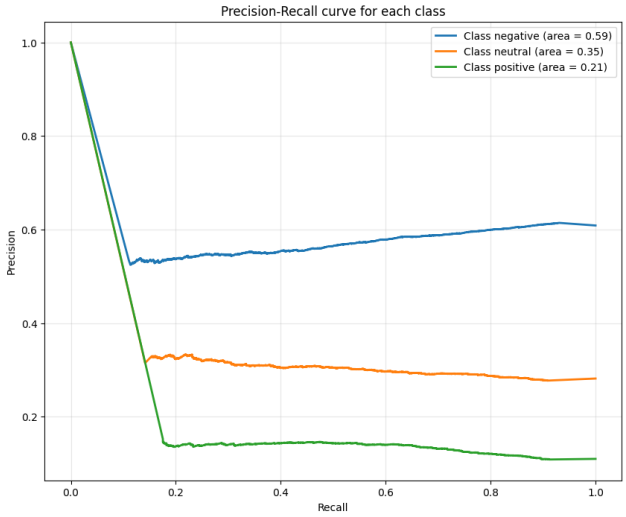

Figure 1. Precision–recall curve for KNN classifier.

Table 5. Overall results of the experiments for "SentiSK" and "Sentigrade" datasets.

Dataset	Measure	RFC	MLP	LR	SVC	KNN	MNB	mBERT	SlovakBERT
SentiSK [39]	**Accuracy (%)**	64.33	63.14	65.46	65.60	58.99	66.07	67.38	69.26
	Precision (%)	84.54	85.26	86.50	86.64	83.39	86.85	86.66	88.51
	Recall (%)	64.33	63.14	65.46	65.60	59.01	66.07	67.51	69.16
	F1 score (%)	60.47	60.79	61.38	60.26	58.51	60.85	65.70	67.68
	Time (s)	417.37	262.65	21.62	1136.17	0.05	0.30	4774.00	3562.00
Sentigrade [28]	**Accuracy (%)**	56.92	60.69	59.43	58.49	54.09	60.69	58.80	71.07
	Precision (%)	61.39	60.38	62.74	63.59	58.41	64.21	80.28	86.46
	Recall (%)	54.54	60.69	59.43	58.49	54.09	60.69	70.76	66.35
	F1 score (%)	54.54	55.83	54.61	54.09	55.48	56.77	59.57	71.53
	Time (s)	0.47	4.48	0.92	0.18	0.00	0.01	193.00	174.00

Table 6. Detailed results of the experiments for each class for "SentiSK" and "Sentigrade" datasets and SlovakBERT and mBERT models.

Dataset	Measure		SlovakBERT	mBERT
SentiSK [39]	F1 score	negative (%)	78.73	77.05
		neutral (%)	49.71	45.67
		positive (%)	54.22	51.01
	Precision	negative (%)	75.14	72.82
		neutral (%)	56.33	51.77
		positive (%)	54.47	54.79
	Recall	negative (%)	82.69	81.79
		neutral (%)	44.49	40.86
		positive (%)	53.98	47.72

Table 6. *Cont.*

Dataset	Measure		SlovakBERT	mBERT
Sentigrade [28]	F1 score	negative (%)	79.75	74.23
		neutral (%)	10.17	0.00
		positive (%)	77.39	66.67
	Precision	negative (%)	78.26	61.80
		neutral (%)	60.00	0.00
		positive (%)	66.45	77.12
	Recall	negative (%)	81.29	92.90
		neutral (%)	5.56	0.00
		positive (%)	92.66	58.72

5.2. "SentiSK" vs. "Sentigrade" vs. "Slovak Dataset for SA"

In the second experiment, we compared the achieved accuracy, precision, recall, F1 score, and training time for all three datasets. When testing the "SentiSK" dataset, we achieved the highest F1 score of 88.16% using the SlovakBERT model, and, conversely, the lowest F1 score (83.99%) was achieved using the KNN classifier (see Table 7). Using the "Sentigrade" dataset, the highest F1 score of 75.35% was achieved by the SlovakBERT model and also mBERT, while the lowest F1 score of 40.70% was achieved by the MLP classifier. The last dataset tested was the "Slovak dataset for SA", and the highest F1 score of 95.04% was achieved using the mBERT model. The lowest F1 score (80.34%) for this dataset was achieved using the KNN classifier. Taking into account the training speed, the mBERT and SlovakBERT model training took the longest for all three datasets. These models also achieved the best results. On the other hand, the KNN, MNB, and LR algorithm classifiers were trained the fastest. However, these achieved the lowest results. The accuracy, precision, recall, F1 score, and time values are presented in Table 7. In Table 8, we provide the results for each class for the "SentiSK", "Sentigrade", and "Slovak dataset for SA" datasets and SlovakBERT and mBERT models.

Table 7. Overall results of the experiments for each algorithm on different datasets.

Dataset	Measure	RFC	MLP	LR	SVC	KNN	MNB	mBERT	SlovakBERT
SentiSK [39]	**Accuracy (%)**	86.42	86.77	87.83	88.02	85.15	87.93	88.85	88.71
	Precision (%)	61.39	60.38	62.74	63.59	58.41	64.21	67.40	65.00
	Recall (%)	86.42	85.66	87.83	88.02	85.15	87.93	87.91	88.14
	F1 score (%)	84.89	85.66	85.69	86.28	83.99	85.58	86.63	88.16
	Time (s)	354.76	147.14	5.27	203.45	0.04	0.12	715.00	2533.00
Sentigrade [28]	**Accuracy (%)**	70.82	56.42	71.98	71.21	60.70	67.70	75.00	77.04
	Precision (%)	54.97	50.08	55.39	58.18	52.51	64.26	54.00	71.00
	Recall (%)	70.82	56.42	71.98	71.21	60.70	67.70	78.99	82.88
	F1 score (%)	70.54	40.70	71.84	71.14	60.40	66.34	75.35	75.35
	Time (s)	0.33	0.79	0.93	0.08	0.00	0.01	530.00	328.00
Slovak dataset for SA [55]	**Accuracy (%)**	81.51	84.65	85.89	85.41	80.36	86.46	88.08	93.90
	Precision (%)	82.47	82.28	86.50	83.22	73.80	85.52	90.14	94.00
	Recall (%)	81.51	84.65	85.89	85.41	80.36	86.46	90.85	94.28
	F1 score (%)	81.50	84.66	85.87	85.40	80.34	86.45	95.04	90.65
	Time (s)	7.45	62.50	1.61	12.14	0.01	0.05	714.00	563.00

Table 8. Detailed results of the experiments for each class for "SentiSK", "Sentigrade", and "Slovak dataset for SA" datasets and SlovakBERT and mBERT models.

Dataset	Measure		SlovakBERT	mBERT
SentiSK [39]	F1 score	negative (%)	93.10	92.58
		positive (%)	62.84	58.70
	Precision	negative (%)	92.56	91.62
		positive (%)	64.74	62.34
	Recall	negative (%)	93.61	93.56
		positive (%)	61.04	55.46
Sentigrade [28]	F1 score	negative (%)	90.91	81.36
		positive (%)	87.56	74.89
	Precision	negative (%)	95.07	85.71
		positive (%)	82.61	70.09
	Recall	negative (%)	87.10	77.42
		positive (%)	93.14	80.39
Slovak dataset for SA [55]	F1 score	negative (%)	93.64	89.73
		positive (%)	93.77	90.06
	Precision	negative (%)	94.00	90.61
		positive (%)	93.42	89.22
	Recall	negative (%)	93.28	88.87
		positive (%)	94.13	90.91

6. Discussion

Overall, we can say that our results for the Slovak language were satisfying. The experimental results in the first task (Section 5.1) were slightly lower than in the second experimental task (Section 5.2). We managed to outperform the results of other studies focused on the Slovak language, which we described in Section 2.2, in most cases. There are not many tools available to preprocess comments from social networks in the Slovak language, unlike other languages (mainly English). Since the processing of Slovak texts is significantly more demanding than processing for the English language, we consider our results to be very favorable. The employed methodologies were straightforward. However, the primary objective was to conduct a comparative analysis between the outcomes derived from our dataset and those of other existing datasets. In pursuit of objectivity, we systematically scrutinized the selected datasets by means of our proprietary source code, ensuring a rigorous evaluation process. Also, the double or triple checking of the data could be introduced, as the data were annotated by a specifically chosen scientific cohort, and each sentence was annotated only once, indicating a lack of back-checking. Additionally, methodologies based on the embedding and vectorization of the data could be explored. In our upcoming work, we will emphasize embedding to more effectively identify the words influencing the sentiment in a sentence. Our primary criterion for making this determination will be the frequency of the word's occurrence.

In comparison with [29], where the authors achieved an accuracy of 76.22% for Support Vector Machine (SVM) and two-class classification, we achieved a higher accuracy (88.20%). On the other hand, for three-class classification, we achieved an accuracy of 65.60%, and the above authors had a very similar accuracy of 66.25%.

Further, in [30], the authors achieved an F1 score of 69.78% in three-class classification by the BiLSTM model. With the SVM model, they achieved an F1 score of 68.40%, and we achieved an F1 score of 60.26%.

Ref. [32] achieved an accuracy of 74.30% by the lexicon approach using BBPSO labeling representation and an accuracy of 80.70% by the lexicon approach in combination with Naïve Bayes, also using BBPSO labeling representation (both for two-class classification). We achieved similar or better results using the RFC, LR, SVC, mBERT, and SlovakBERT models for two-class classification. It should be noted that using the same data, but different training methods, we achieved better results by up to 24%.

In [36], the highest F1 score in task sentiment analysis was 67.20% for two-class classification and 70.50% for three-class classification using the SlovakBERT model. By using the same model, we achieved an F1 score of 71.53% for two-class classification and 90.65% for three-class classification.

It is also interesting to observe how the results for individual classes (see Tables 6 and 8) correlate with the class balance of individual datasets (see Table 4). Consequently, it is crucial to pay attention to the balance of datasets. We will certainly apply these observations in our further research.

7. Conclusions

In this paper, we described the latest research in the field of sentiment analysis on social networks for the Slovak language. As part of our experimental solution, we tested three Slovak datasets that contained comments from social networks. For the first task, where we tested three classes (positive/neutral/negative), we achieved F1 scores ranging from 67.68% to 71.53% for the SlovakBERT model on the "SentiSK" and "Sentigrade" datasets. In the second task, where we tested only two classes (positive/negative), we achieved F1 scores ranging from 75.35% to 95.04% for the mBERT and SlovakBERT models on the "SentiSK", "Sentigrade", and "Slovak dataset for SA" datasets. Eight algorithms or models were used for testing the datasets, namely RFC, MLP, LR, SVC, KNN, Multinomial NB, mBERT, and SlovakBERT. Our primary objective did not entail the introduction of novel insights into the domain of sentiment analysis. Our central aim was to contribute by presenting and disseminating a dataset that underwent manual annotation. Subsequently, we subjected our annotated dataset to rudimentary classification methodologies, followed by a comparative analysis of the outcomes against existing freely accessible datasets in the Slovak language. In our work, we were limited by the lack of datasets in the Slovak language for comparing our results and by the imbalance of the corpus we created. Our paper is primarily aimed at a Slovak audience and Slovak researchers but is also relevant for researchers of other languages. After we publish the dataset, researchers can repeat and validate our experiments and try to achieve similar or better results. Researchers can also translate the datasets into another language, for example through machine translation.

In the near future, we plan to obtain new sentiment data and compare them with the values we obtained herein. We plan to publish both the dataset and the training codes used in this article so that other researchers can also try the experiments. We plan to create a web interface for sentiment detection. Additionally, we plan to concentrate on addressing the issue of hate speech in the Slovak language. Our approach involves creating a balanced dataset in Slovak, manually annotating it, and subsequently developing an automatic annotator specifically tailored for hate speech and offensive language in Slovak. In our subsequent research, we will explore a combination of multiple classifiers. Additionally, we aim to machine-translate an existing database into Slovak and compare the performance of various models across different languages.

Author Contributions: Conceptualization, Z.S., M.H. and J.S.; formal analysis, Z.S., M.H. and J.S.; funding acquisition, J.J., M.P. and D.H.; investigation, D.H.; methodology, Z.S., M.H., M.P. and J.S.; resources, M.H.; software, M.H. and D.H.; supervision, J.J., M.P. and D.H.; validation, Z.S., M.H., M.P. and J.S.; visualization, Z.S., M.H. and J.S.; writing—original draft, Z.S., M.H., M.P. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: The research in this paper was partially supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research projects VEGA 2/0165/21 and KEGA 049TUKE-4/2024; the Slovak Research and Development Agency under the project of bilateral cooperation APVV SK-TW-21-0002; the projects APVV-22-0414 and APVV-22-0261; and the Faculty of Electrical Engineering and Informatics, TU Košice, under the grant FEI-2023-95.

Institutional Review Board Statement: Ethical review and approval were waived for this study because the research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside the research context.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
2. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. [CrossRef]
3. Di Corso, E.; Ventura, F.; Cerquitelli, T. All in a twitter: Self-tuning strategies for a deeper understanding of a crisis tweet collection. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3722–3726.
4. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. [CrossRef]
5. Jiang, L.; Suzuki, Y. Detecting hate speech from tweets for sentiment analysis. In Proceedings of the 2019 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, 2–4 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 671–676.
6. Del Vigna, F.; Cimino, A.; Dell'Orletta, F.; Petrocchi, M.; Tesconi, M. Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, 17–20 January 2017; pp. 86–95.
7. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *J. Comput. Sci.* **2011**, *2*, 1–8. [CrossRef]
8. Gallagher, L.A.; Taylor, M.P. Permanent and temporary components of stock prices: Evidence from assessing macroeconomic shocks. *South. Econ. J.* **2002**, *69*, 345–362.
9. Qian, B.; Rasheed, K. Stock market prediction with multiple classifiers. *Appl. Intell.* **2007**, *26*, 25–33. [CrossRef]
10. Butler, K.C.; Malaikah, S.J. Efficiency and inefficiency in thinly traded stock markets: Kuwait and Saudi Arabia. *J. Bank. Financ.* **1992**, *16*, 197–210. [CrossRef]
11. Kavussanos, M.G.; Dockery, E. A multivariate test for stock market eYciency: The case of ASE. *Appl. Financ. Econ.* **2001**, *11*, 573–579. [CrossRef]
12. Gruhl, D.; Guha, R.; Kumar, R.; Novak, J.; Tomkins, A. The predictive power of online chatter. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, IL, USA, 21–24 August 2005; pp. 78–87.
13. Liu, Y.; Huang, X.; An, A.; Yu, X. ARSA: A sentiment-aware model for predicting sales performance using blogs. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23–27 July 2007; pp. 607–614.
14. Mishne, G.; De Rijke, M. Capturing Global Mood Levels using Blog Posts. In Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, Standford, CA, USA, 27–29 March 2006; Volume 6, pp. 145–152.
15. Ceron, A.; Curini, L.; Iacus, S.M.; Porro, G. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media Soc.* **2014**, *16*, 340–358. [CrossRef]
16. Wang, H.; Can, D.; Kazemzadeh, A.; Bar, F.; Narayanan, S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 System Demonstrations, Jeju, Republic of Korea, 10 July 2012; pp. 115–120.
17. Choy, M.J.; Cheong, M.L.F.; Ma, N.L.; Koo, P.S. A Sentiment Analysis of Singapore Presidential Election 2011 using Twitter Data with Census Correction. *arXiv* **2011**, arXiv:1108.5520.
18. Liu, B. *Sentiment Analysis and Opinion Mining*; Springer Nature: Berlin/Heidelberg, Germany, 2022.
19. Kauffmann, E.; Peral, J.; Gil, D.; Ferrández, A.; Sellers, R.; Mora, H. Managing marketing decision-making with sentiment analysis: An evaluation of the main product features using text data mining. *Sustainability* **2019**, *11*, 4235. [CrossRef]
20. Chowdhury, S.G.; Routh, S.; Chakrabarti, S. News analytics and sentiment analysis to predict stock price trends. *Int. J. Comput. Sci. Inf. Technol.* **2014**, *5*, 3595–3604.
21. Siering, M. "Boom" or "Ruin"—Does It Make a Difference? Using Text Mining and Sentiment Analysis to Support Intraday Investment Decisions. In Proceedings of the 2012 45th Hawaii International Conference on System Sciences, Maui, HI, USA, 4–7 January 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1050–1059.
22. Liang, B.; Su, H.; Gui, L.; Cambria, E.; Xu, R. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowl.-Based Syst.* **2022**, *235*, 107643. [CrossRef]
23. Habernal, I.; Ptáček, T.; Steinberger, J. Supervised sentiment analysis in Czech social media. *Inf. Process. Manag.* **2014**, *50*, 693–707. [CrossRef]
24. Karthika, P.; Murugeswari, R.; Manoranjithem, R. Sentiment analysis of social media network using random forest algorithm. In Proceedings of the 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 11–13 April 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
25. Chetviorkin, I.; Loukachevitch, N. Evaluating sentiment analysis systems in Russian. In Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria, 8–9 August 2013; pp. 12–17.

26. Rotim, L.; Šnajder, J. Comparison of short-text sentiment analysis methods for croatian. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, Valencia, Spain, 4 April 2017; pp. 69–75.
27. Kapočiūtė-Dzikienė, J.; Krupavičius, A.; Krilavičius, T. A comparison of approaches for sentiment classification on lithuanian internet comments. In Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria, 8–9 August 2013; pp. 2–11.
28. Krchnavy, R.; Simko, M. Sentiment analysis of social network posts in Slovak language. In Proceedings of the 2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Bratislava, Slovakia, 9–10 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 20–25.
29. Mojžiš, J.; Krammer, P.; Kvassay, M.; Skovajsová, L.; Hluchỳ, L. Towards Reliable Baselines for Document-Level Sentiment Analysis in the Czech and Slovak Languages. *Future Internet* **2022**, *14*, 300. [CrossRef]
30. Pecar, S.; Šimko, M.; Bielikova, M. Improving sentiment classification in Slovak language. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, Florence, Italy, 2 August 2019; pp. 114–119.
31. Mozetič, I.; Grčar, M.; Smailović, J. Multilingual Twitter sentiment classification: The role of human annotators. *PLoS ONE* **2016**, *11*, e0155036. [CrossRef] [PubMed]
32. Machová, K.; Mikula, M.; Gao, X.; Mach, M. Lexicon-based sentiment analysis using the particle swarm optimization. *Electronics* **2020**, *9*, 1317. [CrossRef]
33. Mikula, M.; Gao, X.; Machová, K. Adapting sentiment analysis system from english to slovak. In Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, USA, 27 November–1 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–8.
34. Pikuliak, M.; Grivalskỳ, Š.; Konôpka, M.; Blšták, M.; Tamajka, M.; Bachratỳ, V.; Šimko, M.; Balážik, P.; Trnka, M.; Uhlárik, F. SlovakBERT: Slovak masked language model. *arXiv* **2021**, arXiv:2109.15254
35. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692
36. Koncz, P.; Paralic, J. An approach to feature selection for sentiment analysis. In Proceedings of the 2011 15th IEEE International Conference on Intelligent Engineering Systems, Poprad, Slovakia, 23–25 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 357–362.
37. Koncz, P.; Paralič, J. Effective corpora creation for sentiment analysis. In *Cognitive Traveling in Digital Space of the Web and Digital Libraries: Studies in Informatics and Information Technologies*; STU: Bratislava, Slovakia, 2013; pp. 92–94.
38. Koncz, P.; Paralič, J. Active learning enhanced document annotation for sentiment analysis. In Proceedings of the Availability, Reliability, and Security in Information Systems and HCI: IFIP WG 8.4, 8.9, TC 5 International Cross-Domain Conference, CD-ARES 2013, Regensburg, Germany, 2–6 September 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 345–353.
39. Sokolová, Z.; Staš, J.; Hládek, D. An Introduction to Detection of Hate Speech and Offensive Language in Slovak. In Proceedings of the 2022 12th International Conference on Advanced Computer Information Technologies (ACIT), Ruzomberok, Slovakia, 26–28 September 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 497–501.
40. Ouyang, X.; Zhou, P.; Li, C.H.; Liu, L. Sentiment analysis using convolutional neural network. In Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, UK, 26–28 October 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 2359–2364.
41. Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; Passonneau, R.J. Sentiment analysis of twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011), Portland, OR, USA, 23 June 2011; pp. 30–38.
42. Arras, L.; Montavon, G.; Müller, K.R.; Samek, W. Explaining recurrent neural network predictions in sentiment analysis. *arXiv* **2017**, arXiv:1706.07206.
43. Sokolová, Z.; Staš, J.; Juhár, J. Review of Recent Trends in the Detection of Hate Speech and Offensive Language on Social Media. *Acta Electrotech. Inform.* **2022**, *22*, 18–24. [CrossRef]
44. Harahus, M.; Hládek, D.; Juhár, J.; Sokolová, Z. Comparison of neural architectures for machine translation of the Slovak language using the Fairseq toolkit. In Proceedings of the 2023 IEEE 21st World Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia, 19–21 January 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 185–190.
45. Sokolová, Z.; Maroš, H.; Staš, J.; Juhár, J. Tvorba korpusu textov pre úlohy detekcie nenávistných prejavov, ofenzívneho jazyka a analýzy sentimentu. *Electr. Eng. Inform.* **2023**, *14*, 399–402.
46. Sokolová, Z.; Maroš, H.; Juhár, J.; Pleva, M.; Hládek, D.; Staš, J. Comparison of Sentiment Classifiers on Slovak Datasets: Original versus Machine Translated. *Int. Conf. Emerg. Elearning Technol. Appl.* **2023**, *21*, 485–492.
47. Schouten, K.; Frasincar, F. Survey on Aspect-Level Sentiment Analysis. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 813–830. [CrossRef]
48. Li, G.; Liu, F. Application of a clustering method on sentiment analysis. *J. Inf. Sci.* **2012**, *38*, 127–139. [CrossRef]
49. Hota, S.; Pathak, S. KNN classifier based approach for multi-class sentiment analysis of twitter data. *Int. J. Eng. Technol.* **2018**, *7*, 1372–1375. [CrossRef]
50. Hasan, S.S.; Adjeroh, D.A. Detecting Human Sentiment from Text using a Proximity-Based Approach. *J. Digit. Inf. Manag.* **2011**, *9*, 206–212.
51. Hasan, S.S.; Adjeroh, D.A. Proximity-based sentiment analysis. In Proceedings of the Fourth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2011), Stevens Point, WI, USA, 4–6 August 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 106–111.

52. Atmakur, V.K.; Kumar, S. A prototype analysis of machine learning methodologies for sentiment analysis of social networks. *Int. J. Eng. Technol. (UAE)* **2018**, *7*, 963–967. [CrossRef]
53. Shickel, B.; Heesacker, M.; Benton, S.; Rashidi, P. Automated emotional valence prediction in mental health text via deep transfer learning. In Proceedings of the 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), Cincinnati, OH, USA, 26–28 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 269–274.
54. Mathews, D.M.; Abraham, S. Lexicon based document level sentiment analysis on the multilingual dataset. In Proceedings of the 2nd International Conference on Advanced Computing and Software Engineering (ICACSE), Sultanpur, India, 8–9 February 2019.
55. Machová, K. Slovak Dataset for Sentimet Analysis. Available online: https://kristina.machova.website.tuke.sk/useful/ (accessed on 14 December 2023).
56. Kramer, O.; Kramer, O. Scikit-learn. In *Machine Learning for Evolution Strategies*; Springer: Berlin/Heidelburg, Germany, 2016; pp. 45–53.
57. DataCamp. Random Forest Classifier in Python. Available online: https://www.datacamp.com/tutorial/random-forests-classifier-python (accessed on 14 December 2023).
58. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
59. Crabbé, A.; Cahy, T.; Somers, B.; Verbeke, L.; Van Coillie, F. Neural Network MLP Classifier. 2020. Available online: https://kuleuven.limo.libis.be/discovery/fulldisplay?docid=lirias3345825&context=SearchWebhook&vid=32KUL_KUL:Lirias&lang=en&search_scope=lirias_profile&adaptor=SearchWebhook&tab=LIRIAS&query=any,contains,LIRIAS3345825&offset=0%soft (accessed on 14 December 2023).
60. Hornik, K. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
61. Bisong, E.; Bisong, E. Logistic regression. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*; Springer: Berline/Heidelburg, Germany, 2019; pp. 243–250.
62. Hosmer, D.W., Jr.; Lemeshow, S. *Logistic Regression;* Dover Publications: Mineola, NY, USA, 1989; Volume 39, pp. 300–313.
63. Cortes, C.; Vapnik, V. Support-vector networks. In *Machine Leaming*; Kluwer Academic Publishers: Boston, MA, USA, 1995; pp. 273–280.
64. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
65. Cover, T.M.; Hart, P.E. An empirical study of the k nearest neighbor rule. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1957; pp. 510–532.
66. McCallum, A.; Nigam, K. A comparison of event models for Naive Bayes text classification. In Proceedings of the AAAI-98 Workshop on Learning for Text Categorization, Madision, WI, USA, 26–27 July 1998; Volume 752, pp. 41–48.
67. Lewis, D.D.; Gale, W.A. Naive Bayes text classification. *Mach. Learn.* **1998**, *37*, 1–44.
68. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.