Educational Administration: Theory and Practice
2024, 30(4), 6377-6381
ISSN: 2148-2403
https://kuey.net/

Research Article

The Role Of Naïve Bayes, SVM, And Decision Trees In Sentiment Analysis

Dhamayanthi N[1], B. Lavanya B.[2*]

[1,2] * Department of Computer Science, University of Madras Chennai, India dhamayanthin@outlook.com, lavanmu@gmail.com

Citation: B. Lavanya B et. Al. (2024), The Role Of Naïve Bayes, SVM, And Decision Trees In Sentiment Analysis, *Educational Administration: Theory and Practice*, *3*(4), 6377-6381, Doi: 10.53555/kuey.v30i4.2392

ARTICLE INFO

ABSTRACT

Sentiment analysis is a critical area of study within natural language processing that aims to systematically identify and categorize opinions expressed in text data. This paper evaluates the performance of three prominent machine learning algorithms, Naïve Bayes, Support Vector Machines (SVM), and Decision Trees in their ability to conduct sentiment analysis. Through empirical testing on datasets composed of online product reviews, we compare the accuracy, efficiency, and applicability of each algorithm. Our results indicate that SVMs provide the highest accuracy (85%), effectively managing high-dimensional data and complex linguistic structures. However, Naïve Bayes offers unparalleled speed, making it ideal for real-time applications, while Decision Trees excel in interpretability, despite their susceptibility to overfitting. The study highlights significant challenges, including sarcasm detection, contextual dependency, and data bias, suggesting future research directions such as enhanced contextual analysis and the development of multimodal sentiment analysis systems. This research contributes to the ongoing advancement of sentiment analysis technologies, providing insights that can aid in the selection of appropriate algorithms.

Key Words: Sentiment Analysis, Machine Learning, Natural Language Processing, Text Mining, Opinion Mining

Introduction

In the age of information, understanding human sentiments through digital communications has become crucial for businesses, policymakers, and individuals alike. Sentiment analysis, a subfield of natural language processing (NLP), focuses on the computational identification and categorization of opinions expressed in text data. This technology enables the automated analysis of customer feedback, social media comments, and other forms of written expression to gauge public opinion, monitor brand reputation, and enhance customer service.

With advancements in artificial intelligence and machine learning, it is now possible to analyse vast amounts of textual data with increasing accuracy. This research aims to compare the effectiveness of different machine learning models, Naïve Bayes, Support Vector Machines, and Decision Trees in accurately classifying and predicting sentiments expressed in text.

Literature Survey

The exploration of sentiment analysis across diverse sectors has been well-documented in recent research. Starting with the political sphere, Ansari et al. utilized LSTM networks to scrutinize sentiments on Twitter during the 2019 Indian General Elections, suggesting predictive insights into election outcomes [1]. Meanwhile, in the educational domain, Dalipi et al. performed a comprehensive review of sentiment analysis applications on MOOC feedback from 2015 to 2021, pinpointing critical areas for further exploration [2]. The commercial potential of sentiment analysis is also evident, as demonstrated by Dhamayanthi and Lavanya, who applied ensemble machine learning algorithms to analyze sentiments on Amazon product reviews, with stochastic gradient boosting showing superior results [3]. In a multilingual context, Kanclerz et al. introduced language-agnostic techniques to enable sentiment analysis across languages, validating their model with texts translated into eight different languages [4].

Expanding into social media, Nhan et al. delved into the challenges of applying deep learning to analyze sentiments, comparing various models' effectiveness [5]. The technical aspects of sentiment analysis were further explored by Palomino and Aider, who demonstrated how strategic pre-processing significantly enhances the accuracy of Naïve Bayes classifiers in Twitter sentiment analysis [6]. The optimization of sentiment analysis using machine learning classifiers was methodically studied by Singh et al., who underscored the importance of feature selection and model tuning in enhancing analysis outcomes [7]. Sohangir et al. integrated deep learning with big data to enhance financial sentiment analysis on platforms like StockTwits, revealing the high efficacy of convolutional neural networks [8].

Moreover, Talaat introduced innovative deep learning models combining RoBERTa and DistilBERT to refine sentiment analysis on social media, aiming to better understand public opinion for strategic decision-making [9]. Tang et al. innovatively integrated sentence and document representations using convolutional and gated recurrent networks for document-level sentiment classification, showing improved performance on IMDB and Yelp datasets [10]. Pagolu et al. investigated the relationship between Twitter sentiments and stock market trends, employing advanced machine learning techniques to predict market movements [11]. Wang et al. introduced a hybrid convolutional-recurrent neural network for text classification, demonstrating its superiority over existing models [12]. Wankhade et al. provided a detailed overview of sentiment analysis, discussing its applications and challenges, particularly in analyzing public opinions derived from social media [13]. Lastly, Yadav et al. enhanced sentiment analysis techniques for financial news by modifying semantic orientation calculation methods, finding that noun-verb combinations were particularly effective [14].

These studies collectively underscore the broad applicability and effectiveness of sentiment analysis across different platforms and languages, highlighting its critical role in shaping decisions in political, educational, commercial, and financial sectors.

Machine Learning Algorithms for Sentiment Analysis

Machine learning (ML) is a branch of artificial intelligence that focuses on building systems that can learn from and make decisions based on data. In the context of sentiment analysis, machine learning algorithms interpret and classify human sentiments from textual data, learning to recognize complex patterns in language that indicate positive, negative, or neutral tones.

Key Algorithms

1. **Naïve Bayes**
   - Description: Naïve Bayes classifiers are probabilistic models that apply Bayes' Theorem, assuming independence between predictors. Simple to implement and fast in processing, they are particularly effective for large datasets.
   - Application in Sentiment Analysis: Often used for binary classification problems such as distinguishing between positive and negative reviews.
   - Strengths: Efficiency and speed.
   - Limitations: The assumption of independent predictors is often unrealistic in language processing, which can affect accuracy.

2. **Support Vector Machines (SVM)**
   - Description: SVMs are powerful classifiers that find the optimal hyperplane that best separates different classes in a high-dimensional space.
   - Application in Sentiment Analysis: Effective in handling non-linear relationships and high-dimensional data, making them suitable for complex sentiment analysis tasks.
   - Strengths: Accuracy in high-dimensional spaces and robustness to overfitting in large feature sets.
   - Limitations: Requires careful parameter tuning and can be computationally intensive, especially with large datasets.

3. **Decision Trees**
   - Description: Decision Trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance.
   - Application in Sentiment Analysis: Used for feature importance evaluation to determine which terms are most indicative of sentiment.
   - Strengths: Easy to interpret and can handle both numerical and categorical data.
   - Limitations: Prone to overfitting and can be biased toward attributes with more levels.

Methodology

Data Collection
To conduct sentiment analysis, we have collected the data from publicly available datasets consisting of product reviews from online marketplaces. These datasets are chosen for their diversity in language use and sentiment expression, providing a robust testing ground for our algorithms.

Data Pre-processing
Data pre-processing is crucial for effective sentiment analysis. The following steps are undertaken:
- Text Cleaning: Removing HTML tags, special characters, and numbers, and converting all text to lowercase to standardize the input data.
- Tokenization: Breaking down paragraphs into sentences or words to simplify analysis.
- Stop Words Removal: Eliminating common words (e.g., "and", "the") that do not contribute to sentiment analysis.
- Stemming/Lemmatization: Reducing words to their base or root form to consolidate different forms of the same word.

Algorithm Implementation
1. **Naïve Bayes**
   - Training: The model is trained using a bag-of-words approach where frequency counts of words are used as features. The Naïve Bayes classifier then uses these features to estimate the probabilities of a text belonging to positive, negative, or neutral categories based on the Bayes theorem.
   - Testing: The classifier is tested on a separate dataset to ensure the model generalizes well to new data.

2. **Support Vector Machines (SVM)**
   - Training: SVMs are trained using both linear and non-linear kernels to best capture the high-dimensional feature space typical in text data. Features are TF-IDF (term frequency-inverse document frequency) scores.
   - Testing: Similar to Naïve Bayes, SVM models are evaluated using a distinct set of data to check for overfitting and to assess generalization capabilities.

3. **Decision Trees**
   - Training: Decision trees are trained to create models that classify sentiments based on rules inferred from the data features, like word occurrences and syntax patterns. The trees make decisions by creating branches that lead to the most dominant features determining sentiment.
   - Testing: The performance of decision trees is tested using separate validation data to ensure that the model does not merely memorize the training data.

Results and Discussion

Results Overview
The performance of each machine learning algorithm was evaluated on a diverse set of text data from product reviews. The key findings are summarized as follows:

1. **Naïve Bayes**
   - Accuracy: Achieved an accuracy of approximately 78%.
   - Strengths: The algorithm was particularly efficient with large datasets and quick in delivering results, making it suitable for real-time sentiment analysis.
   - Weaknesses: Struggled with handling sarcasm and subtleties in language due to its simplistic assumption of feature independence.

2. **Support Vector Machines (SVM)**
   - Accuracy: Demonstrated superior accuracy, reaching up to 85%.
   - Strengths: Excelled in handling non-linear relationships in data and was robust against overfitting, especially in high-dimensional feature spaces.
   - Weaknesses: The main drawback was its computational intensity, which required substantial resources for training and tuning, particularly with larger datasets.

3. **Decision Trees**
   - Accuracy: Recorded an accuracy of around 80%.
   - Strengths: Provided clear insights into which features were most influential in sentiment determination, facilitating an easy interpretation of the decision-making process.
   - Weaknesses: Was prone to overfitting, particularly when dealing with data that had numerous features, which often led to overly complex models.

Comparative Analysis

The comparative analysis reveals that SVMs generally provide the best accuracy among the three tested algorithms, due to their effectiveness in managing the complexity and variety within the text data. However, the choice of an algorithm can depend significantly on the specific application requirements such as the need for real-time analysis (favouring Naïve Bayes for its speed) or the requirement for interpretable results (favouring Decision Trees).

The results underscore the importance of choosing the right machine learning algorithm based on the specific characteristics of the sentiment analysis task at hand. Each algorithm's performance varied based on the nature of the dataset and the computational resources available. Furthermore, the limitations observed, such as the handling of sarcasm and feature interdependence, highlight areas for future improvement in algorithm development. Enhancements in natural language processing techniques and deeper integration of contextual understanding could potentially improve the performance of these algorithms.

Challenges and Future Directions

Current Challenges

Despite significant advancements in machine learning for sentiment analysis, several challenges persist:

- Handling Sarcasm and Irony: Current models often struggle to correctly interpret sentiments when faced with sarcastic or ironic expressions, as these require an understanding of context beyond the literal meanings of words.
- Contextual Dependency: Sentiments can be heavily dependent on context, which can shift dramatically within a single document or even a sentence. Current algorithms may fail to capture these subtle shifts, leading to inaccuracies.
- Ambiguity and Polysemy: Words with multiple meanings (polysemy) can lead to confusion in sentiment classification, especially when out of context clues are minimal.
- Data Bias: Machine learning models can inadvertently learn and perpetuate biases present in the training data, affecting their fairness and objectivity.

Future Directions

To address these challenges and push the boundaries of what machine learning can achieve in sentiment analysis, future research could focus on several key areas:

- Enhanced Contextual Analysis: Developing models that better understand the broader context of conversations, possibly through the integration of world knowledge and longer memory spans in neural networks.
- Improved Sarcasm Detection: Leveraging more sophisticated NLP techniques and incorporating additional linguistic features that may indicate sarcasm or irony.
- Bias Mitigation: Implementing techniques to detect and correct biases in training data and model predictions to ensure fairness and inclusivity in sentiment analysis.
- Cross-lingual and Multimodal Sentiment Analysis: Expanding research to include multiple languages and modalities (e.g., text, audio, video) to build more robust and universally applicable models.
- Unsupervised and Semi-supervised Learning Methods: Exploring less supervised methods that can learn from unlabelled data, reducing the dependence on extensively annotated datasets and broadening the applicability of models to more diverse datasets.

Conclusion

This research explored the effectiveness of various machine learning algorithms, Naïve Bayes, Support Vector Machines (SVM), and Decision Trees in performing sentiment analysis. Our findings demonstrate that while each algorithm has its strengths, SVMs generally outperformed the others in terms of accuracy, achieving up to 85% on diverse datasets. Naïve Bayes offered speed and efficiency, making it suitable for applications requiring real-time analysis. Decision Trees provided valuable insights into the features most influential in determining sentiments, though they were prone to overfitting.

The practical implications of these findings are significant for fields such as marketing, customer service, and social media monitoring, where understanding consumer sentiment is crucial. By selecting the appropriate machine learning algorithm, organizations can better gauge public opinion, tailor services to consumer needs, and enhance overall customer satisfaction.

References

1. Ansari M.Z., Aziz M.B., Siddiqui M.O., Mehra H. & Singh K.P. (2020). Analysis of Political Sentiment Orientations on Twitter. Procedia Computer Science, 167, pp. 1821-1828. https://doi.org/10.1016/j.procs.2020.03.201

2. Dalipi F, Zdravkova K, & Ahlgren F. (2021). Sentiment Analysis of Students' Feedback in MOOCs: A Systematic Literature Review. Frontiers in Artificial Intelligence. Sep 9.4.728708.https://www.frontiersin.org/articles/10.3389/frai.2021.728708
3. Dhamayanthi, N., Lavanya, B. (2021). Sentiment Analysis Framework for E-Commerce Reviews Using Ensemble Machine Learning Algorithms. In: Bhateja, V., Satapathy, S.C., Travieso-González, C.M., Aradhya, V.N.M. (eds) Data Engineering and Intelligent Computing. Advances in Intelligent Systems and Computing, vol 1407. Springer, Singapore. https://doi.org/10.1007/978-981-16-0171-2_34
4. Kanclerz K., Milkowski P., & Kocon J. (2020). Cross-lingual deep neural transfer learning in sentiment analysis. Procedia computer Science, 176, pp. 128-137. https://doi.org/10.1016/j.procs.2020.08.014
5. Nhan C.D, Maria N.M, and Fernando D.L.P (2020). Sentiment Analysis Based on Deep Learning: A Comparative Study. Electronics, 9(3), 483. https://doi.org/10.3390/electronics9030483
6. Palomino, Marco A., & Farida Aider. (2022). Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis. Applied Sciences, 12(17). https://doi.org/10.3390/app12178765
7. Singh, J., Singh, G. & Singh, R. (2017). Optimization of sentiment analysis using machine learning classifiers. Human-centric Computing and Information Sciences. 7, 32. https://doi.org/10.1186/s13673-017-0116-3
8. Sohangir, S., Wang, D., Pomeranets, A. *et al.* (2018). Big Data: Deep Learning for financial sentiment analysis. Journal of Big Data, 5. https://doi.org/10.1186/s40537-017-0111-6
9. Talaat, A.S. (2023). Sentiment analysis classification system using hybrid BERT models. Journal of Big Data 10, 110. https://doi.org/10.1186/s40537-023-00781-w
10. Tang D., Qin B., & Liu T. (2015). Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1422–1432, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/https://doi.org/10.18653/v1/D15-1167
11. V. S. Pagolu, K. N. Reddy, G. Panda & B. Majhi. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In: International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES). pp. 1345-1350. https://doi.org/10.1109/SCOPES.2016.7955659
12. Wang R., Li Z., Cao J., Chen T. & Wang L. (2019). Convolutional Recurrent Neural Networks for Text Classification. *2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 2019, pp. 1-6, https://doi.org/10.1109/IJCNN.2019.8852406
13. Wankhade, M., Rao, A.C.S. & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55, pp. 5731–5780. https://doi.org/10.1007/s10462-022-10144-1
14. Yadav A., Jha C.K., Sharan A., & Vaish V. (2020). Sentiment analysis of financial news using unsupervised approach. Procedia Computer Science, 167, pp. 589-598. https://doi.org/10.1016/j.procs.2020.03.325