International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
Article Received: *25 June 2023* ***Revised:*** *12 July 2023* ***Accepted:*** *20 August 2023*

Comparative Analysis of Machine Learning Techniques for Opinion Mining in Web Texts Using Artificial Intelligence

Ram Chandra Pal [1] and Dr. Suresh S. Asole[2]
[1,2]Department of Computer Science & Engineering, Dr. A. P. J. Abdul Kalam University, Indore, MP, India
rcpal07@gmail.com[1] suresh_asole@yahoo.com[2]

Abstract : This paper presents a comparative analysis of machine learning techniques for opinion mining in web texts using artificial intelligence. Opinion mining, also known as sentiment analysis, involves extracting and classifying opinions from textual data found on the web, such as reviews and blogs. The study evaluates the effectiveness of traditional data mining classifiers Naïve Bayes, k-Nearest Neighbor, and Random Forest and neural network classifiers LVQ, Elman, and FFNN. A novel approach combining Naïve Bayes and SVM in an ensemble method is proposed to enhance classification accuracy. The KINN algorithm, a neural network-based model, is introduced, demonstrating improved performance over existing methods. Experimental results using a dataset from Kaggle show that the proposed methods achieve higher accuracy in sentiment classification, offering valuable insights for improving product and service quality based on customer feedback.

Keywords: *KINN, FFNN, NLP, AI*

INTRODUCTION

Natural Language Processing (NLP) has revolutionized the way textual data is processed, enabling the transformation of text into machine-readable formats. When combined with Artificial Intelligence (AI), this transformation allows for the extraction of opinions from textual data on the web, a process known as opinion mining or sentiment analysis. Opinion mining encompasses a range of tasks such as sentiment classification, feature-based sentiment classification, and opinion summarization.

Importance of Opinion Mining

The proliferation of user-generated content on platforms like social media, blogs, and review sites has made opinion mining crucial for businesses and organizations. By analyzing these opinions, companies can gauge public sentiment towards their products, services, and brand image. This feedback loop is vital for strategic decision-making and improving customer satisfaction.

Challenges in Opinion Mining

Opinion mining involves several challenges:

- **Ambiguity in Text**: Human language is inherently ambiguous, with words and phrases often having multiple meanings depending on context.
- **Subjectivity Detection**: Distinguishing between subjective and objective statements can be difficult.
- **Feature Extraction**: Identifying and extracting relevant features from text requires sophisticated algorithms.

These challenges necessitate the use of advanced machine learning techniques to accurately classify and summarize opinions.

Machine Learning Approaches

Various machine learning algorithms are employed in opinion mining, each with its strengths and limitations. Techniques such as Naive Bayes, Support Vector Machines (SVM), and ensemble methods combine multiple algorithms to improve classification accuracy. These methods leverage large datasets to train models capable of understanding and predicting sentiment with high precision.

RESEARCH OBJECTIVES

This thesis aims to:

- Analyze different machine learning algorithms used for opinion extraction.
- Evaluate the performance of these algorithms in terms of accuracy, efficiency, and scalability.
- Propose improvements or new approaches to enhance opinion mining techniques.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
Article Received: *25 June 2023* ***Revised:*** *12 July 2023* ***Accepted:*** *20 August 2023*

REVIEW OF LITERATURE

Chen and Qi (2011) report about the job of informal organization in online customer's choice procedure when they scan for an unpractised item to purchase. They get a lot of framework derivation and coordinated into three-phase framework engineering. For choice stages, straight chain contingent irregular field-based social-assessment mining calculation is utilized and recognized its viability against gauge.

Generally sentence extremity is recognized by either a solitary word or mix of words which are unigram, bigram and trigram. Relies upon area and characterization strategies, specialists utilize any of the above said 'grams' since the mix has critical part in arrangement execution. In view of the information corpus, the grouping model may take any kind of text, for example, unigram, bigram or trigram. To use the total connection between words, such sets are vital. Ganu et.al. (2012) propose techniques that utilization text-based café audits to improve suggestions. This literary data gives preferred forecast over numerical star appraisals. Yang and Ko (2011) utilize changed Term recurrence (TF) and Inverted Document Frequency (IDF) through word check to feature the significance of a bigram.

METHODOLOGY

The web has become a vital platform for online learning, thought exchange, and product/service reviews due to the accessibility of high transmission capacity wireless networks. Tracking and understanding user sentiment on the web has become challenging due to the vast number of reviews available. Opinion mining, a burgeoning research area, summarizes users' product/service reviews to determine whether opinions are positive, negative, or neutral.

Data Collection

Data for this study was collected from various online review sites and social blogs where users express their opinions. The reviews were retrieved using data retrieval technologies to gather sentiment data on products/services.

Methodology

1. **Data Preprocessing:**
   - Extract reviews and opinions from collected data sources.
   - Clean the data by removing unnecessary information such as HTML tags, stop words, and special characters.
   - Tokenize the text into individual words or phrases.
2. **Feature Extraction:**
   - Identify and extract relevant features that contribute to the sentiment of the text.
   - Use Natural Language Processing (NLP) techniques to identify sentiment-related words and phrases.
3. **Sentiment Analysis:**
   - Implement machine learning algorithms to classify the sentiment of the reviews as positive, negative, or neutral.
   - Utilize supervised learning methods for sentiment classification, employing labeled datasets to train the model.
4. **Evaluation:**
   - Evaluate the performance of the sentiment analysis model using metrics such as accuracy, precision, recall, and F1-score.
   - Compare the results with baseline models to ensure the effectiveness of the proposed approach.

Tools and Technologies

- **Natural Language Processing (NLP):** Used for text processing and feature extraction.
- **Machine Learning Algorithms:** Employed for sentiment classification.
- **Data Retrieval Technologies:** Used for collecting data from online sources.

This methodology provides a systematic approach to analyze user opinions on the web, offering valuable insights for improving product/service quality based on customer feedback.

KINN (KIrubakaran and Nisha Jebaseeli Neural Network) Algorithm

The KINN algorithm, named after the research guide Dr. Kirubakaran and the researcher Prof. Nisha Jebaseeli, is a neural network-based algorithm designed to improve classification accuracy. The key features of the KINN algorithm include:

1. **Extra Weights from Input Layer to Output Layer**:
   - KINN introduces additional weights connecting the input layer directly to the output layer. This architecture enhances the network's efficiency compared to traditional Feedforward Neural Networks (FFNNs) with a similar number of processing elements.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
Article Received: *25 June 2023* ***Revised:*** *12 July 2023* ***Accepted:*** *20 August 2023*

2. **Improved Backpropagation Algorithm**:

- The KINN algorithm uses an enhanced backpropagation method to optimize the learning process. This involves calculating the summation and activation functions within the hidden layer using a Sigmoidal activation function. The process is iterated until a threshold value is encountered or the number of epochs is reached.

3. **Entropy and Information Gain**:

- The algorithm calculates entropy for each feature and uses Information Gain to select the top five positive and top five negative values for further processing. This selection process helps in improving the classification accuracy by focusing on the most informative features.

4. **Architecture and Parameters**:

- The KINN architecture includes one hidden layer with 20 neurons, an input layer with 57 neurons, and an output layer with 3 neurons. The learning rate is set at 0.1, momentum at 0.5, and the number of epochs at 500.

5. **Efficiency**:

- KINN has demonstrated higher efficiency and classification accuracy compared to other ANN classifiers such as LVQ, Elman, and FFNN. For example, in experiments with the M-Learning dataset, KINN achieved a classification accuracy of 83.33%, outperforming other algorithms.

In summary, KINN is a neural network algorithm that leverages additional input-to-output connections and an improved backpropagation technique to enhance classification accuracy. It employs entropy and information gain for feature selection and has been validated to perform better than traditional neural network models.

RESULT AND DISCUSSION

Nostalgic investigation has recently been utilized by nearly every online firm due to the increase in customers buying products online, prompting businesses to enhance product qualities. Nowadays, many customers who use websites, blog sites, and online shopping platforms tend to review products. These reviews significantly influence other customers' purchasing decisions during their product search. Businesses have thus identified the key aspects customers look for in the right product based on user reviews using customer sentiment analysis. Sentiment analysis is a data analysis concept that collects, analyzes, processes, and recommends these reviews to users. These reviews are extensive, consisting of several content paragraphs. The data used for this research work is gathered from the Kaggle website, which provides Amazon product reviews. Initially, these reviews are preprocessed to remove noisy data (e.g., stop words, verbs, conjunctions). Using Naive Bayes and SVM algorithms, the preprocessed data is classified via the training dataset. These current machine learning algorithms have given insufficient precision. Therefore, an ensemble approach is applied to improve the accuracy of the review classification. The ensemble represents a classification approach by combining two or more algorithms and determining the mode value for each algorithm used, based on the vote reference. Naive Bayes and Support Vector Machine algorithms are combined in this research work for an ensemble approach, leading to greater accuracy compared to the existing algorithms in the proposed method. Once accuracy is determined, the item is recommended based on the reviews.

Table 1: Product Review Dataset

S.No	Name	Description
1.	Product Title	Details about the Product Name and its specifications
2.	Brand	Brand Details with specifications
3.	Price	Price of the product
4.	Rating	Rating of the product between 1 to 5
5.	Review	Review comments about the product
6.	Review Votes	Number of people who found the review helpful

Neural networks have become essential in data mining due to their high accuracy and ability to handle noisy data. In Knowledge Discovery in Databases (KDD), they are used for order processing. The rise of sentiment-rich resources like online reviews and blogs has led to advancements in sentiment mining and analysis, which processes emotions and opinions in text. Originally studied by NLP researchers, sentiment mining classifies text as positive or negative and is widely used in product reviews, business intelligence, and recommender systems. This computational method aids decision-making by analyzing large sets of opinion data. Various classifiers, such as Naive Bayesian and Support Vector Machine, have been explored. This research introduces a novel neural network classification algorithm, the Input Output Positive Negative Weight Feed Forward Neural Network, to enhance classification accuracy.

A Preprocessing Algorithm to Improve the Classification Accuracy

The order precision of three data mining classification algorithms and three neural network classification algorithms were compared using ten-fold cross-validation for training. The algorithms compared include Naïve Bayes, k-Nearest Neighbor, Random Forest, LVQ, Elman, and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
Article Received: *25 June 2023* ***Revised:*** *12 July 2023* ***Accepted:*** *20 August 2023*

FFNN classifiers. Sentiment extraction and classification play a crucial role in achieving classification accuracy. This study proposes extracting sentiment words based on their frequency across documents, creating a word frequency matrix by eliminating common and rare words, and ranking extracted opinions using SVD. Among the algorithms, FFNN provided the best classification accuracy. Results comparing data mining classifiers and the proposed KN preprocessing are shown in figure, with Naïve Bayes, kNN, and Random Forest accuracy rates listed.

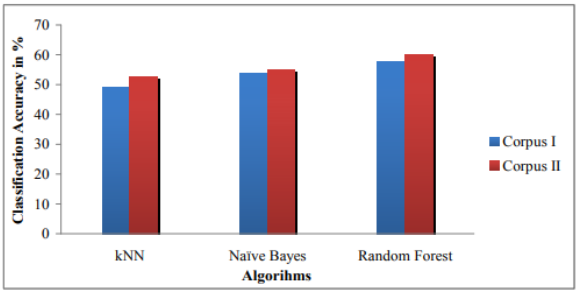

Figure 1: Classification Accuracies with Data Mining Classification Algorithms

Figure-1, diagrammatically shows the classification accuracies. Classification Accuracies obtained through explorations reveal an improvement of a mean of 2.78%.

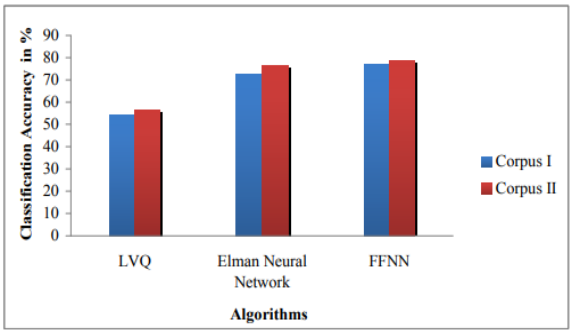

Figure 2: Classification Accuracies with Neural Network Based Algorithms

Supposition mining classification, The proportion of audits ordered effectively to the total number of surveys described is used to measure accuracy. Order Classifiers aren't judged on accuracy alone. Precision, Recall, and F-measure assess characterisation exactnesses to identify its surroundings. Accuracy and recall tests might reveal a paired classifier's presentation characteristics. Classifier accuracy measures precision. Higher accuracy reduces fake positives, whereas lesser exactness increases them. Classifier effectiveness is assessed via review. Higher reviews indicate fewer false negatives, while lower reviews indicate more. Improving review can reduce accuracy since it makes it harder to be precise as test space grows. One measurement fashioned by merging Precision and Recall is F-measure, the weighted symphonious mean of accuracy and review. The F-measure reflects review versus accuracy importance.

CONCLUSION

This research underscores the importance of advanced machine learning techniques in opinion mining for web texts. By comparing traditional data mining classifiers and neural network classifiers, the study highlights the superior performance of the KINN algorithm and the proposed ensemble method. These approaches significantly enhance sentiment classification accuracy, addressing the challenges posed by noisy and subjective textual data. The findings suggest that integrating sophisticated algorithms like KINN with ensemble methods can lead to more precise sentiment analysis, ultimately benefiting businesses and organizations by providing deeper insights into customer opinions. Future work could explore the application of these techniques to larger and more diverse datasets, as well as the integration of additional contextual and semantic analysis methods to further refine sentiment classification.

REFERENCES

[1] Chen.L, Qi.L (2011) Social opinion mining for supporting buyers' complex decision making: exploratory user study and algorithm comparison. Soc. Netw. Anal. Min. Page No. 301–320. doi:10.1007/s13278-011-0023-y

[2] Ganu.G, Kakodkar.Y, Marian.A (2012) Improving the quality of predictions using textual information in online user reviews. Information Systems, vol 38, Page No. 1-15. doi:10.1016/j.is.2012.03.001

[3] Yang.S, Ko.Y (2011) Finding relevant features for Korean comparative sentence extraction. Pattern Recognition Letters. vol 32, Page No. 293–296. doi:10.1016/j.patrec.2010.09.003

[4] Carenini.G, Ng.R, and Pauls.A (2006) Multi-document summarization of evaluative text. In Proceedings of European of the Association for Computational Linguistics (EACL 2006) Page No. 305-312

[5] Allan J., "Introduction to topic detection and tracking," in Topic Detection and Tracking:Event-based Information Organization, (J. Allan, ed.), Page No. 1–16, Norwell, MA, USA: Kluwer Academic Publishers, ISBN 0-7923-7664-1, 2002.

[6] Schumaker.RP, Zhang.Y, Huang.CN, Chen.H (2012) Evaluating Sentiment in Financial News Chapters. Decision Support Systems (2012), Page No. 458-463. doi: 10.1016/j.dss.012.03.001

[7] Wiebe.J, Riloff.E (2011) Finding Mutual Benefit between Subjectivity Analysis and Information Extraction. IEEE Transactions on Affective Computing. vol 2, Page No. 175 – 191. doi:10.1109/T-AFFC.2011.19

[8] Wang.H, Zhu.J, Zhu.M, Tsou.BK, Ma.M (2011) Aspect-Based Opinion Polling from Customer Reviews. IEEE Transactions on Affective Computing, vol 2, Page No. 37-49. doi.ieeecomputersociety.org/10.1109/T-AFFC.2011.2

[9] Wu.Q, Tan.S (2011) A two-stage framework for cross-domain sentiment classification. Expert Systems with Applications. vol 38, Page No. 14269–14275. doi:10.1016/j.eswa.2011.04.240

[10] Sun.A, Ling. E.-P., Lui.Y (2010) On strategies for imbalanced text classification using SVM: a comparative study. Decision Support Systems. vol 48, Page No. 191–201. doi:10.1016/j.dss.2009.07.011

[11] Tan.S, Wang.Y (2011) Weighted SCL model for adaptation of sentiment classification.Expert Systems with Applications. vol 38, Page No. 10524–10531. doi:10.1016/j.eswa.2011.02.106

[12] Zhang.R, Tran.T, Mao.Y (2011) Opinion helpfulness prediction in the presence of "words of few mouths",WorldWide Web. Vol 15. Page No. 117-138. doi:10.1007/s11280-011-0127-3,2011

[13] Melville.P, Gryc.W, Lawrence.RD (2009) Sentiment analysis of blogs by combining lexical knowledge with text classification. In: Proceedings of the Fifteenth ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '09, ACM, 2009, Page No. 1275–1284. doi:10.1145/1557019.1557156

[14] Miao.Q,, Li.Q,, Dai.R (2009) Amazing: a sentiment mining and retrieval system. Expert SystAppl 2009; vol 36, Page No. 7192–7198. doi:10.1016/j.eswa.2008.09.035

[15] Liu.B, GeliFei, Mukherjee.A, Meichun Hsu, MaluCastellanos, and RiddhimanGhosh (2013) Exploiting Burstiness in Reviews for Review Spammer Detection. To appear in Proceedings of The International AAAI Conference on Weblogs and Social Media (ICWSM-2013), Boston, USA.