IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105329 |

Emotion Classification on Textual Data

B. Anjali Reddy[1], K. Damodhar Rao[2], M. Sathyanarayana[3]

B. Tech Student, Dept. of Computer Science and Engineering, SNIST, Hyderabad, India[1]

Assistant Professor, Dept. of Computer Science and Engineering, SNIST, Hyderabad, India[2,3]

ABSTRACT: The classification of textual data into emotions is a crucial aspect of understanding human communication, as emotions play a significant role in shaping behavior, decision-making, and perception. This paper highlights the significance of emotion classification, emphasizing its relevance in providing insights into human psychology, identifying patterns in emotional expression, aiding communication research, and facilitating social media analysis. Emotion classification involves the process of identifying and categorizing emotions expressed in any form of text or speech, and its importance is increasing with the widespread use of digital platforms and the internet. This paper explores the various applications of emotion classification in textual data, including predicting consumer behavior, detecting mental health issues, improving customer service, and coordinating disaster response efforts. Overall, emotion classification is a vital tool that can aid businesses, organizations, and researchers in understanding and analyzing emotions expressed in textual data, leading to improved decision-making and strategies.

KEYWORDS: -Textual Data Analysis; Predictive Modelling, Sentiment Analysis; Emotion Classification; Pattern Recognition

I. INTRODUCTION

Emotions are a fundamental aspect of human behavior and play a crucial role in shaping our thoughts, actions, and communication. As humans, we express our emotions in various ways, including through spoken and written language. With the increasing popularity of social media platforms and the internet, the volume of text data containing emotions has surged, making automated emotion classification an essential task in various fields such as psychology, social science, and marketing.

Automated emotion classification involves the use of machine learning algorithms to analyze and classify emotions from text data accurately. The technology has numerous potential applications, including sentiment analysis in social media, identifying customer feedback in marketing, and understanding emotional responses in healthcare.
To develop a machine learning model that can accurately classify emotions from text data, various techniques and methods are employed. The first step involves text pre-processing, which includes cleaning and transforming the raw text data into a structured format suitable for analysis. Common text pre-processing techniques include tokenization, stemming, and stop word removal.

After pre-processing the text data, feature engineering techniques are employed to extract meaningful information from the text. These techniques involve transforming the text data into a numerical format suitable for machine learning algorithms. Common feature engineering methods include bag-of-words, n-grams, and word embeddings.
Once the text data has been pre-processed and features extracted, various machine learning algorithms can be employed to classify emotions accurately. These algorithms include logistic regression, and support vector machines. The performance of each algorithm is evaluated using various performance metrics.

The potential applications of automated emotion classification in text data are numerous. In marketing, the technology can be used to understand customer feedback and identify areas for improvement. In healthcare, it can be used to understand the emotional responses of patients, leading to improved patient outcomes. In social media, it can be used to understand public sentiment and identify trends and topics of discussion.

In conclusion, automated emotion classification from text data is a crucial task with numerous potential applications. Developing accurate machine learning models involves various techniques and methods, including text pre-processing, feature engineering, and machine learning algorithms. As the volume of text data containing emotions continues to grow, the need for accurate and automated emotion classification will become increasingly important in various fields, leading to improved insights and decision-making.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105329 |

II. RELATED WORK

Related work in the field of automated emotion classification from text data has been widely explored in the literature. The research can be categorized into two main approaches: rule-based and machine learning-based methods. Rule- based approaches involve manually defining rules to identify emotions in text data. These methods rely on predefined dictionaries or lexicons of emotion-related words and phrases, such as the Affective Norms for English Words (ANEW) and the Linguistic Inquiry and Word Count (LIWC). Rule-based approaches have been found to have limitations in accurately capturing the complexity and context specificity of emotional expression. In contrast, machine learning-based approaches involve training a model using annotated data to learn patterns and relationships between features and emotions. These approaches have shown to outperform rule-based methods in accuracy and flexibility. Various machine learning algorithms have been employed, including Naïve Bayes, Support Vector Machines (SVM), Decision Trees, Random Forests, and Artificial Neural Networks (ANN). Several studies have compared the performance of different machine learning algorithms for automated emotion classification. For example, a study by Mohammad et al. (2013) compared the performance of SVM, Naïve Bayes, and Decision Trees on classifying emotions in tweets, with SVM achieving the highest accuracy. Another study by Poria et al. (2017) compared the performance of various deep learning models, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), on classifying emotions in movie reviews, with LSTM-RNN achieving the highest accuracy. In addition to comparing the performance of different machine learning algorithms, research has also explored the impact of different features and pre-processing techniques on emotion classification accuracy. For example, a study by Bajaj et al. (2016) evaluated the impact of using different n gram sizes and different types of text pre-processing techniques on emotion classification accuracy, with stemming and bigrams achieving the highest accuracy. Overall, the related work in automated emotion classification from text data has shown the potential of machine learning based approaches in accurately identifying emotions in text data. However, the research also highlights the importance of carefully selecting and evaluating pre-processing techniques and machine learning algorithms to achieve optimal performance.

III. METHODOLOGY

This study aims to compare the performance of two machine learning algorithms, logistic regression, and support vector machines (SVM), for automated emotion classification on textual data using embeddings stored by means of TF- IDF. The methodology consists of the following steps:

1. Data Collection: The first step in developing a machine learning model for emotion classification is to collect a relevant dataset of textual data containing emotions. This dataset can be obtained from various sources such as social media, customer feedback, or healthcare records. The dataset should be sufficiently large and diverse to ensure that the machine learning model is robust and generalizable.
2. Text Pre-processing: Once the dataset has been collected, the next step is to pre-process the text data. Text pre-processing involves cleaning and transforming the raw text data into a structured format suitable for analysis. This includes removing stop words, stemming, and converting all text to lowercase. In addition, any irrelevant data such as hyperlinks, usernames, or hashtags should be removed.
3. Feature Engineering: After pre-processing the text data, the next step is to extract meaningful features from the text. Feature engineering techniques involve transforming the text data into a numerical format suitable for machine learning algorithms. Common feature engineering methods include bag of-words, n-grams, and word embeddings. In this study, we will use word embeddings stored by means of TF-IDF.
4. Model Training: With the pre-processed and feature engineered data, we will train two machine learning models, namely logistic regression and SVM. These models will be trained on a portion of the dataset, known as the training set. The objective of model training is to learn a mapping between the input data and the output labels.
5. Model Evaluation: Once the models have been trained, we will evaluate their performance on a separate portion of the dataset, known as the test set. The test set is used to measure the accuracy of the model on new, unseen data. Common evaluation metrics for emotion classification include accuracy, precision, recall, and F1-score.
6. Model Comparison: After evaluating the performance of each model, we will compare the results of logistic regression and SVM. We will consider factors such as accuracy, speed, and ease of implementation in our

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105329 |

comparison. Additionally, we will perform a statistical analysis to determine if one model significantly outperforms the other. Interpretation of Results: Finally, we will interpret the results of our study and draw conclusions about the effectiveness of logistic regression and SVM for emotion classification on textual data with embeddings stored by means of TF-IDF. We will discuss the strengths and limitations of each model and suggest areas for future research.

IV. IMPLEMENTATION

1. Import the necessary libraries: pandas, numpy, os, random, re, nltk, seaborn, matplotlib.pyplot, sklearn.
2. Download the required NLTK packages: 'punkt', 'wordnet', and 'stopwords'.
3. Load the train and test datasets as pandas dataframes, with columns 'Sentences' and 'Emotion'. The data should be separated by a semicolon and the columns should be named accordingly.
4. Check the shape of the training and testing dataframes using the shape() function to ensure the data is loaded correctly.
5. Drop any duplicate rows and missing values in the training data using the drop_duplicates() and dropna() functions, respectively.
6. Visualize the distribution of emotions in the training and testing datasets using the countplot() function from seaborn library.
7. Define the list of stop words using the stopwords.words() function from the NLTK library.
8. Define the WordNetLemmatizer object using the WordNetLemmatizer() function from the NLTK library.
9. Define the function 'expand' to expand the contractions in the sentences.
10. Define the function 'process' to preprocess the text data by removing non-alphabetic characters, expanding contractions, converting all characters to lowercase, tokenizing the sentences into words, lemmatizing the words, and joining the words back into sentences.
11. Apply the 'process' function to the 'Sentences' column of both the train and test dataframes to preprocess the text data.
12. Define the TfidfVectorizer object using the TfidfVectorizer() function from the sklearn.feature_extraction.text library with the maximum number of features set to 8000.
13. Transform the preprocessed 'Sentences' column of the train dataframe into a sparse matrix of TF-IDF features using the fit_transform() function of the TfidfVectorizer object.
14. Transform the preprocessed 'Sentences' column of the test dataframe into a sparse matrix of TF-IDF features using the transform() function of the TfidfVectorizer object.
15. Extract the emotion from the 'Emotion' column of both the train and test dataframes.
16. Define the logistic regression model object using the LogisticRegression() function from the sklearn.linear_model library with the maximum number of iterations set to 100000.
17. Train the logistic regression model on the training data using the fit() function of the logistic regression object.
18. Use the trained logistic regression model to predict the emotion labels of the test data using the predict() function of the logistic regression object.
19. Evaluate the performance of the logistic regression model using accuracy score, confusion matrix, and classification report using the appropriate functions from the sklearn.metrics library.
20. Define the SVM model object using the SVC() function from the sklearn.svm library with the kernel set to 'linear' and the regularization parameter set to 1.
21. Train the SVM model on the training data using the fit() function of the SVM object.
22. Use the trained SVM model to predict the emotion labels of the test data using the predict() function of the SVM object.
23. Evaluate the performance of the SVM model using accuracy score, confusion matrix, and classification report using the appropriate functions from the sklearn.metrics library.
24. Compare the performance of both models based on the evaluation metrics and draw conclusions.

V. RESULTS

The proposed system for emotion classification of textual data using logistic regression and support vector machine algorithms with TF-IDF embedding has shown promising results. The system was trained and evaluated on a dataset of 10,000 text documents with labelled emotions, and the performance was measured using accuracy as the evaluation metric. The logistic regression algorithm achieved an accuracy of 87%, while the support vector machine

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105329 |

algorithm achieved an accuracy of 89%. These results demonstrate the potential of using machine learning algorithms for automated emotion classification in text data. The results also indicate that the support vector machine algorithm with TF-IDF embedding outperforms the logistic regression algorithm. Therefore, the proposed system has the potential to be applied in various applications such as sentiment analysis, customer feedback analysis, and other text-based emotion recognition

VI. ADVANTAGES OF SYSTEM

The proposed system of automated emotion classification of textual data using SVM and Logistic Regression algorithms with TF-IDF embeddings has several advantages: Firstly, the system is highly accurate, achieving an accuracy of 89% with SVM and 87% with Logistic Regression algorithms. This accuracy is a significant improvement compared to traditional methods of manual emotion classification, which can be time-consuming and subjective. Secondly, the system is efficient and scalable. As the volume of textual data containing emotions continues to grow, the system can be easily scaled up to handle large datasets. This makes it an ideal solution for various applications, including sentiment analysis in social media, identifying customer feedback in marketing, and understanding emotional responses in healthcare. Thirdly, the system is automated, which reduces the need for human intervention, leading to faster processing times and improved efficiency. This means that the system can be used in real-time applications such as social media monitoring, where timely analysis of emotions is crucial for making informed decisions

VII. CONCLUSION AND FUTURE WORK

In conclusion, the proposed system for emotion classification of textual data using logistic regression and support vector machine algorithms with TF-IDF embedding has demonstrated promising results. The system achieved an accuracy of 87% with logistic regression and 89% with support vector machine. These results demonstrate the potential of using machine learning algorithms for automated emotion classification in text data. The study also highlighted the significance of text pre-processing and feature engineering techniques in achieving accurate results. TF-IDF embedding was found to be an effective feature engineering technique for representing text data in a numerical format suitable for machine learning algorithms. The system's accuracy and efficiency can lead to improved insights and decision-making, ultimately resulting in better outcomes.

In the future, further improvements could be made to enhance the accuracy and effectiveness of the emotion classification system. One potential area for enhancement could be to incorporate deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to better capture the complex and nuanced relationships between words and emotions. Another possible improvement could be to incorporate additional features, such as syntactic and semantic features, to improve the accuracy of emotion classification.

REFERENCES

1. Peng, X., & Xue, G. (2019). Emotion classification of Chinese microblogs using a combination of machine learning models. Information Processing & Management,56(1),1-14. https://doi.org/10.1016/j.ipm.2018.08.006
2. Zhang, L., & Zhou, L. (2019). Emotion classification of Chinese microblogs based on convolutional neural networks and support vector machines. IEEE Access, 7, 13347-13356.
3. https://doi.org/10.1109/access.2019.2890913
4. Kaur, H., & Gupta, A. (2020). Sentiment Analysis of Movie Reviews Using Machine Learning Algorithms. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(5), 4415-4421. https://doi.org/10.35940/ijitee.J9876.099520
5. Du, Z., Zhang, Y., & Hu, Q. (2020). An Improved Method for Emotion Classification in Textual Data Based on SVM Algorithm. In 2020 International Conference on Advanced Communication Technologies and Networking (CommNet) (pp. 249- 254). IEEE
6. "Emotion and Sentiment Analysis: A Survey" by Fabrizio Sebastiani. (2018) Link: https://dl.acm.org/doi/10.1145/3159652

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105329 |

7. "Emotion Recognition and Classification: A Review" by T. Huang, Z. Zeng, and M. Li. (2020) Link: https://www.sciencedirect.com/science/article/abs/pii/S1568494619310469
8. "Emotion Classification of Online Textual Conversations" by Chongyang Wang, Chao Wang, Fang Chen, and Guozhen Zhao. (2020)
9. Link: https://www.mdpi.com/2073-8994/12/5/707
10. "Emotion Recognition from Textual Conversations: A Review" by Ahmad Anwar, Hafeez Anwar, and Tahseen Jilani. (2020)
11. Link: https://www.sciencedirect.com/science/article/pii/S2212017320305595
12. "Emotion Recognition from Text: A Survey" by Jia Chen, Mengting Wan, Julian McAuley, and Irwin King. (2020)
13. Link: https://arxiv.org/abs/2005.00091

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details