EmoSense: Exploring Textual Emotions using Multi-Model Analysis with K-fold Cross Validation and Feature Engineering

Faiaz Mohammad Tiham
ID: 20101026
Dept of CSE
Brac University
faiaz.mohammad.tiham@g.bracu.ac.bd

Ehsan Abdullah Khan Saad
ID: 20101512
Dept of CSE
Brac University
ehsan.abdullah.khan@g.bracu.ac.bd

Nihad Adnan Shah Tirtho
ID: 20101611
Dept of CSE
Brac University
nihad.adnan.shah@g.bracu.ac.bd

Gazi Md. Julcarnine
ID: 20101360
Dept of CSE
Brac University
gazi.md.julcarnine@g.bracu.ac.bd

Hossain Mohammed Usman
ID: 20101053
Dept of CSE
Brac University
hossain.mohammed.usman@g.bracu.ac.bd

***Abstract*- Emotion recognition from text is a major task in natural language processing. Its applications include sentiment analysis, user engagement prediction, and personalized content delivery. This paper introduces EmoSense, an approach that uses Random Forest, Naive Bayes, and Support Vector Machine models to recognize emotions from text. These models have various strengths as well as weaknesses, but EmoSense combines them to achieve better accuracy and depth of emotion classification. The dataset used in it is carefully curated to include a wide range of emotional expressions. A systematic feature engineering process is used to extract nuanced features that represent different emotional states. It also uses K-fold cross-validation to ensure the models are not overfitting the training data. EmoSense outperforms individual models in the accurate classification of textual emotions. The findings of this paper demonstrate the efficacy of EmoSense in emotion recognition. It is a promising new approach to emotion recognition from text. It has the potential to improve the accuracy and depth of emotion classification, and it can be used in a wide range of NLP applications.**

Chapter 1
Introduction

The massive volume of textual data produced every day on many platforms presents a rich resource of information representing human emotions, attitudes, and ideas in today's globally interconnected world. For a variety of applications, including sentiment analysis, brand monitoring, customer feedback analysis, and mental health support, it is essential to figure out the emotional context behind this textual data.

A lot of focus has been given to the capability of "Emotion Analysis" or "Sentiment Analysis," a field that studies the accurate detection and interpretation of emotions in text. Recent developments in machine learning and NLP have provided new avenues for studying emotions inside text in a data-driven manner, whereas conventional methods have mainly depended on rule-based or lexicon-based approaches. This study discusses "EmoSense," an innovative approach created to investigate textual emotions through multi-model analysis, K-fold cross-validation, and feature engineering. EmoSense uses machine learning models such as Support Vector Machines (SVM), Random Forests, and Naive Bayes to discover the complex emotional nuances contained in textual data. To guarantee the accuracy and adaptability of the findings, K-fold cross-validation, a reliable method for evaluating model performance, is used. In order to improve the models' capacity to detect subtle emotional motives in the text, feature engineering is also used. This study is noteworthy because it has the potential to increase our knowledge of how machine-learning techniques can be used to interpret the complex web of human emotions in text. We acquire insights into the mental structures of textual data by creating models that can effectively recognize and categorize emotions. We also set the ground for the creation of intelligent systems that can offer more context-aware responses and interventions.

In the sections that follow, we'll go into more detail about EmoSense's core ideas as well as the approaches and procedures used to investigate textual emotions. We will discuss the experimental design, datasets used, preprocessing procedures, and feature engineering techniques used. The outcomes of our multi-model analysis will then be discussed, along with the advantages and disadvantages of each strategy. Finally, we'll finish off with a discussion of EmoSense's larger implications and prospective uses in a range of industries, including marketing, customer service, and mental health support, among others.

Chapter 2

Related works

In order to deal with high dimensionality, this literature [1] emphasizes feature selection as it examines machine learning for sentiment analysis. It also examines filter and wrapper feature selection strategies, emphasizing the computationally effective filter approach. The strategy that performs best is identified as Information Gain. As the best option for sentiment analysis, the Multinomial Naive Bayes with Term Frequency classifier is recognized with a brief mention. It describes how SVM works to transform data and identify the best-separating hyperplanes. The introduction of a feature vector creation technique with semantic clustering utilizing PMI and binary weighting. mRMR-based composite features outperform IG. It is proposed that BMNB outperforms SVM. Trigrams are discovered to perform badly, however, combining unigrams with multi-word feature vectors is successful. Without providing specific outcomes, the essay [1] emphasizes the significance of feature selection and classifier selection in sentiment analysis.

This paper [2], talks about how sentiment analysis and expression have changed in the internet era. For automated analysis, knowledge-based and machine-learning techniques are used. Sentiment analysis is graded on a scale from coarse to fine, with sentence-level analysis acting as the middle level. Due to Twitter's shortness, sentiment analysis is difficult, and feature extraction is required for effective sentiment analysis using gathered tweet data. Lexical resources are frequently used in symbolic techniques, such as Turney's bag-of-words method. It addresses how to evaluate emotional word content using a variety of methods, such as WordNet and Finite State Automata. Term presence, term frequency, negation, n-grams, and part-of-speech features are used to categorize reviews using machine learning techniques like Naive Bayes and novel models. It [2] highlights the use of noisy training sets, clustering techniques, and ensemble frameworks while mentioning various studies and classifiers. Using SVM, Naive Bayes, Maximum Entropy, and ensemble classifiers, the method performs sentence-level sentiment

analysis on a fresh dataset of tweets about electronic products. By obtaining accuracy through feature extraction and testing with several classifiers, it highlights the effectiveness of machine learning over symbolic techniques in sentiment analysis.

Francisca Adoma Acheampong, Chen Wenyu and Henry Nunoo-Mensah (2020) mentioned in-text emotion detection and challenges in their paper named “Text-based emotion detection: Advances, challenges,and opportunities”. The discussed paper [3] addresses the challenges in detecting emotions from text when compared to multimodal methods. Texts often lack clear emotional cues, and this difficulty is compounded by short texts, emojis, and grammatical intricacies. The evolving lexicon adds to the complexity. Despite growing interest, text-based emotion detection is still in its early stages, lacking robust research and effective techniques. The paper [3] highlights two main approaches: emotion dictionaries, limited by the need for distinct emotion categories and keyword ambiguity, and the lexical affinity method, which oversimplifies emotions. Research has shown promise in detecting emotions across languages, involving data preprocessing, feature extraction, training, and SVM classifiers. While progress has improved human-computer interaction, challenges and research gaps remain in this field.

Kush Shrivastava, Shishir Kumar & Deepak Kumar Jain (17 july,2019) published a paper titled as “An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network” which discussed some effective approaches for detection emotion in multimedia text. The document [4] highlights the role of emotions in text communication and the challenges of annotating emotions in text. The objective of the research work mentioned in the document [4] is to automatically recognize emotions from multimedia textual content. The primary aim is to enable the use of a Convolutional Neural Network (CNN) methodology for emotion detection in text. To address the challenges in emotion detection, researchers can consider building a large corpus of text that is rich in emotions, incorporating contextual information, developing advanced algorithms, expanding emotion databases, and considering linguistic features. These approaches can help improve the accuracy and effectiveness of emotion detection in text.

Nourah Alswaidan and Mohamed El Bachir Menai (18 march, 2020) discuss the given text in their paper titled "A survey of state-of-the-art approaches for emotion recognition in text"
This document [5] is a comprehensive exploration of emotion recognition in text, highlighting challenges related to implicit emotions and the significance of contextual comprehension. The paper further delves into four primary approaches addressing emotion recognition in text: rule-based, classical learning-based, deep learning-based (particularly LSTM), and hybrid methodologies. Rule-based techniques involve predefined patterns to detect explicit emotions, while facing difficulties in capturing implicit emotional nuances. Classical learning relies on emotion-annotated data to train models, showcasing effectiveness at the cost of data intensity.

Deep learning, featuring LSTM models, excels in intricate emotion recognition by capturing long-term dependencies. Hybrid approaches amalgamate methods to optimize results by leveraging strengths and mitigating limitations. The survey [5] offers a comparative analysis of these diverse approaches, identifying the most effective ones and scrutinizing their respective advantages and limitations.

Adil Majeed, Hasan Mujtab and Mirza Omer Beg published a paper on 22 January 2021 titled “Emotion detection in Roman Urdu text using machine learning” where they discussed the challenges of extracting emotions from text and the importance of sentiment analysis. It [6] highlights the limitations of detecting emotions in text, such as subtle expressions, ambiguous words, and sarcastic or slang language. To address Roman Urdu emotion detection, the paper [6] employs diverse methods. Data collection involves sourcing from platforms like hamariweb, YouTube, and social media, using Selenium and Twint. Manual annotation by expert annotators assigns emotions like happy, sad, anger, fear, love, and neutral. Preprocessing follows, encompassing noise, punctuation, and URL removal to prepare the corpus. In order to enhance classification, Word2Vec features are used to train machine learning algorithms such as KNN, Random Forest, Decision Tree, and SVM on the processed data. Overall, the document focuses on the difficulties and significance of extracting emotions from text and the need for automated systems to analyze sentiment accurately.

Alieh Hajizadeh Saffar, Tiffany Katharine Mann and Bahadorreza Ofoghi distributed a paper named "Text based feeling identification in wellbeing: Advances and applications," in which they mention that implementing TED in healthcare settings faces difficulties and limitations.[7] The difficulty of accurately detecting emotions in text, particularly in clinical settings where the language used can be complex and nuanced, is one of the obstacles. The lack of standardization in the field is another obstacle, making it challenging to compare results from different studies. Moreover, there are worries about protection and security while involving touchy wellbeing information for feeling recognition. Lastly, more research is required to establish standardized evaluation metrics and validate the efficacy of TED in healthcare settings. Implementing and evaluating dimensional emotional models, overcoming annotation difficulties with the help of health-related lexicons, and utilizing deep learning techniques for multifaceted and real-time applications are some of the solutions. Also, creators propose the requirement for more cooperation between specialists, clinicians, and industry accomplices to create and carry out TED applications in medical services settings.

The publication [8] compares the Support Vector Machines (SVM) method to the Enhanced Text Emotion Prediction (ETEP) algorithm for text emotion prediction. The study's methodology, experimental design, and findings are discussed in the publication [8], which also emphasizes how well the ETEP algorithm predicts emotions from textual data. The performance of the Support Vector Machines (SVM) method and the Enhanced Text Emotion Prediction (ETEP) algorithm for text emotion prediction is compared in the publication [8]. The ETEP algorithm uses feature extraction, data preprocessing, and machine learning methods to forecast emotions from textual input.The benchmark for comparison is the SVM algorithm.The dataset utilized, the experimental setting, and the assessment measures used are all covered in the document [8].The results reveal that the ETEP algorithm outperforms SVM in terms of performance, proving its usefulness for text emotion prediction.The study's conclusions provide insightful information for the creation of precise and effective text emotion prediction systems, advancing sentiment analysis and natural language processing.

This document [9] discusses the methodology for depression diagnosis using machine learning (ML) algorithms. It covers various feature extraction methods and supervised learning classifiers used in the diagnosis process. The aim is to provide clinicians and healthcare professionals with a better understanding of ML approaches for depression diagnosis.The document [9] focuses on the methodology for depression diagnosis using ML algorithms.It discusses feature extraction methods such as SelectKBest, Particle Swarm Optimization (PSO), Maximum Relevance Minimum Redundancy (mRMR), Boruta, and RELIEFF.The document [9] also mentions supervised learning classifiers used in the diagnosis process.It highlights the limitations of existing studies in the depression diagnosis domain.Future research possibilities in the field of depression diagnosis are listed.The document [9] also highlights the limitations of existing studies in the domain and suggests future research possibilities.

This paper [10] explores the use of sentiment mining methods for the purpose of identifying sentimental element within suicide notes. This statement explains that these systems use language analysis or language processing to evaluate the overall mood or expression of a given text. The paper [10] also discusses the historical background of language analysis in relation to the different levels of emotion that can be analyzed with suicidal notes. The proposal is to enhance the performance of sentiment identification algorithms by integrating lexical text features with semantic information and other contextual factors. The article concludes by discussing the complexities associated with accuracy rate of predicting emotions and the potential benefits gained by having a large set of training datasets.

Chapter 3
Dataset description :

The dataset employed in this study encompasses a total of 40,000 instances, each characterized by three essential attributes: the unique tweet ID, the associated sentiment label, and the corresponding textual content. The sentiment labels encapsulate a spectrum of emotional expressions, ranging from "empty" to sentiments like "sadness" and "enthusiasm." The data is structured in a tabular format with 40,000 rows and 3 columns. Among these, the "sentiment" attribute exhibits some missing values, with 549 instances lacking sentiment labels, while the "content" attribute is complete with no null entries. The dataset is divided into training and testing subsets, constituting 80% and 20% of the data, respectively, to facilitate robust evaluation of the models. The dataset's unique characteristics and comprehensive design make it an invaluable resource for exploring and enhancing emotion recognition through natural language processing techniques.

Chapter 4
Methodology

This section of this research paper discusses the steps and procedures followed to construct the "EmoSense" framework, which aims to explore textual emotions using multi-model analysis with K-fold cross-validation and feature engineering. This section is divided into several key stages: data collection, preprocessing, feature engineering, model selection, K-fold cross-validation, and performance evaluation.

4.1 Data collection:

The foundation of EmoSense is a carefully organized dataset comprising textual data with associated emotional labels. The choice of a dataset is critical to the success of emotion analysis tasks. We made sure that our dataset was balanced and representative of the emotions of interest.

4.2 Preprocessing:

Our textual data contained noise and inconsistencies that could have hindered machine learning model performance. The preprocessing steps include:

4.2.1 Text Cleaning: Remove special characters, punctuation, and unnecessary whitespace.
4.2.2 Tokenization: Splitting text into individual words or tokens.
4.2.3 Handling Missing Data:** Address any missing values or null entries in the dataset.

4.3 Feature Engineering:

Feature engineering is a crucial step in transforming raw text into numerical features that machine learning models can process effectively. EmoSense utilizes some techniques, including:

4.3.1 Bag of Words (BoW): Representing text as a matrix of word frequencies.
4.3.2 Word Embeddings: Using pre-trained word embeddings like Word2Vec, GloVe, or FastText to capture semantic relationships between words.
4.3.3 N-grams: Capture sequences of adjacent words to account for context.
4.3.4 Sentiment Lexicons: Incorporating external sentiment lexicons to enhance emotion detection.

4.4 Model Selection:

EmoSense employs a suite of machine learning models, including Support Vector Machines (SVM), Random Forests, and Naive Bayes. The choice of models Depends on the dataset and the specific requirements of the emotion analysis task. Model hyperparameters are tuned to optimize performance.

4.5 K-fold Cross-Validation:

In order to make sure the models are reliable and generalizable, K-fold cross-validation is employed. The dataset is divided into K equally sized folds and K iterations of training and testing are performed, with each fold serving as the testing set exactly once. This helps estimate model performance on unseen data and mitigates overfitting.

4.6 Performance Evaluation:

Model performance is evaluated using several metrics tailored to emotion analysis tasks, including accuracy, precision, recall, F1-score, and confusion matrices. The choice of evaluation metrics aligns with the research objectives and the nature of the dataset.

Throughout the methodology, transparency and reproducibility are emphasized. Detailed documentation of dataset sources, preprocessing steps, feature engineering techniques, and model configurations is essential for future researchers to replicate and build upon the EmoSense framework.

The ensuing sections of this research paper will delve into the experimental results and analysis, shedding light on the performance of the selected models, the impact of feature engineering, and the insights gained from the exploration of textual emotions.