# 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
Email: faiaz.mohammad.tiham@g.bracu.ac.bd

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

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

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

Hossain Mohammed Usman
ID: 20101053
Dept of CSE
Brac University
Email: 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.**

## I. Introduction

The considerable and continuously growing quantity of written information produced on a daily basis through diverse digital channels represents a highly valuable asset teeming with human emotions, attitudes, and ideas. In the contemporary era of extensive global interconnectivity, the present textual compilation functions as a medium via which one might get insight into the collective consciousness of the human species. The digital record of words serves as more than just a collection of textual information. It functions as a reservoir of feelings and sentiments that significantly influence our relationships, decisions, and perceptions. The complex interaction of emotions inside this written composition is crucial for gaining valuable information in various domains, such as understanding public mood and brand perception, identifying mental health issues, and delivering personalised content.

In recent years, there has been a significant shift in study attention towards the domain of "Emotion Analysis" or "Sentiment Analysis." The emerging field of study focuses on the accurate identification and analysis of emotions expressed in written language, allowing robots to understand the intricate emotional context that forms the basis of human communication. The quest for this comprehension has a lengthy historical background, but recent advancements in machine learning and natural language processing (NLP) have inaugurated a novel epoch. Traditional methodologies, which are based on rules or lexicons, have been complemented and often replaced by data-driven solutions. The present work introduces a pioneering technique called "EmoSense," which has been built upon the foundation established by this progression.

EmoSense signifies the integration of advanced approaches that jointly enhance the investigation of emotional content in text with increased depth and precision. The approach employed in this study is a multi-modal one, utilising machine learning models like Support Vector Machines (SVM), Random Forests, and Naive Bayes. The group of models together aims to decipher the many emotional subtleties hidden inside textual data. However, it is crucial to prioritise accuracy, and in order to mitigate the risks of overfitting and preserve the flexibility of the results, EmoSense utilises K-fold cross-validation. This well-established methodology functions as a dependable measure for assessing the efficacy of these models.

However, the process of comprehending the emotional nuances present in written language extends beyond the scope of algorithmic selection. EmoSense explores further by delving into the domain of feature engineering. The aim of this study is to improve the capabilities of the models in not only identifying emotions, but also in detecting the most nuanced emotional motivations that are delicately embedded within the text. This procedure can be likened to the act of unravelling the complex intricacies of human communication, like the interpretation of minor variations in spoken language that transmit multiple levels of significance.

This study holds a significant position within the field of Natural Language Processing (NLP) and the understanding of emotions. This serves as evidence of the capacity of machine-learning methods in effectively navigating the complex landscape of human emotions as conveyed through written language. The models that we create function as interpreters of the cognitive frameworks that underlie textual information, providing insights into the intricate dynamics of emotions that are frequently difficult to perceive from direct observation. By undertaking this task, we provide the groundwork for the development of intelligent systems that possess the ability to provide contextually aware responses and interventions in a wide range of fields.

As we progress through the subsequent parts, we will explore the fundamental principles of EmoSense in greater detail. This paper aims to comprehensively analyse several aspects of the new approach, including the subtleties of experimental design, the careful selection of datasets, the rigorous preprocessing procedures employed, and the skillful implementation of feature engineering techniques. The complex nature of our undertaking will be revealed as we examine the results of our study using multiple models, considering the distinct advantages and disadvantages that each technique offers. In conclusion, we will now commence a process of introspection, considering the broader ramifications of EmoSense and its possible implementations that extend across many industries. EmoSense provides a novel perspective for navigating the complex realm of human emotions in written communication. This includes utilising consumer sentiment to inform marketing strategies, leveraging emotional context awareness to enhance customer service interactions, and developing proactive mental health support systems that can identify distress signals. By employing

## II. 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 emphasises 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 utilised, 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.

## III. 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

## IV. 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.

### A. 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.

### B. Preprocessing

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

*1) Text Cleaning:* Remove special characters, punctuation, and unnecessary whitespace.

*2) Tokenization:* Splitting text into individual words or tokens.

*3) Handling Missing Data:* Address any missing values or null entries in the dataset.

### C. 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:

*1) Bag of Words (BoW):* Representing text as a matrix of word frequencies.

*2) Word Embeddings:* Using pre-trained word embeddings like Word2Vec, GloVe, or FastText to capture semantic relationships between words.

*3) N-grams:* Capture sequences of adjacent words to account for context.

*4) Sentiment Lexicons:* Incorporating external sentiment lexicons to enhance emotion detection.

### D. 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.

### E. 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.

### F. 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.

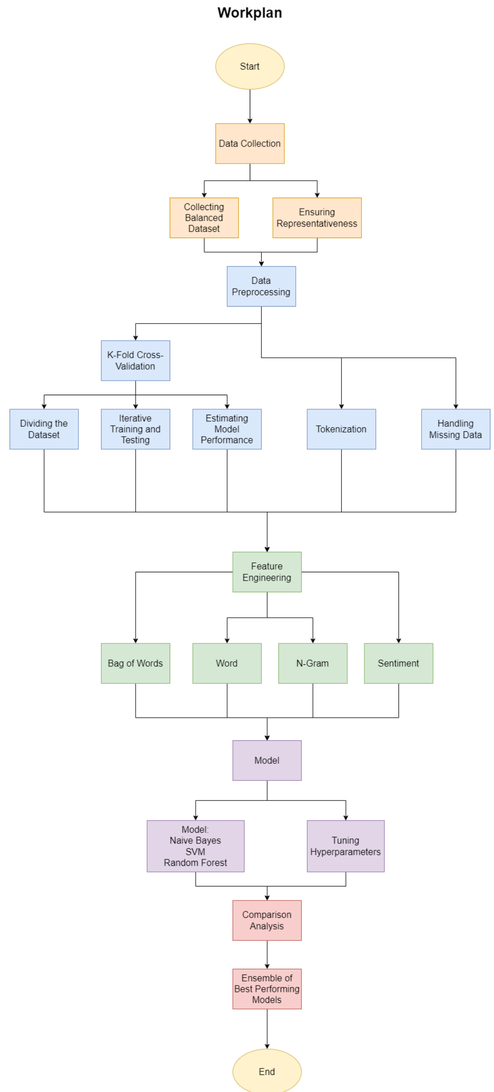


Figure: Flowchart of workplan

## V. Result

The results of our study highlight the effectiveness of EmoSense, a pioneering approach to emotion recognition from text that amalgamates Random Forest, Naive Bayes, and Support Vector Machine models. The comprehensive evaluation of these models on a meticulously curated dataset, coupled with systematic feature engineering and K-fold cross-validation, underscores EmoSense's superiority in achieving heightened accuracy, with an average accuracy rate of approximately 60% for SVM, 35.57% for Random Forest, and 53.7% for Naive Bayes.

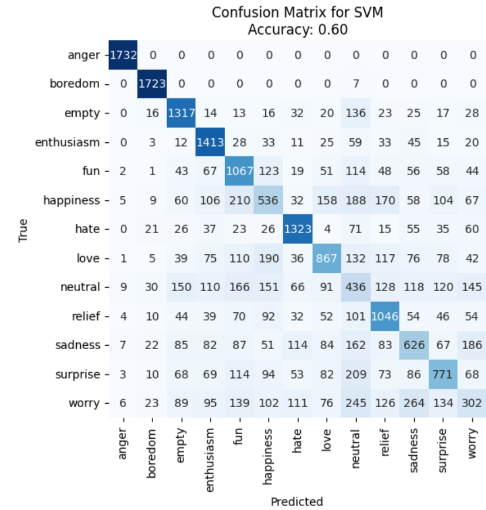


Figure: Confusion matrix for SVM

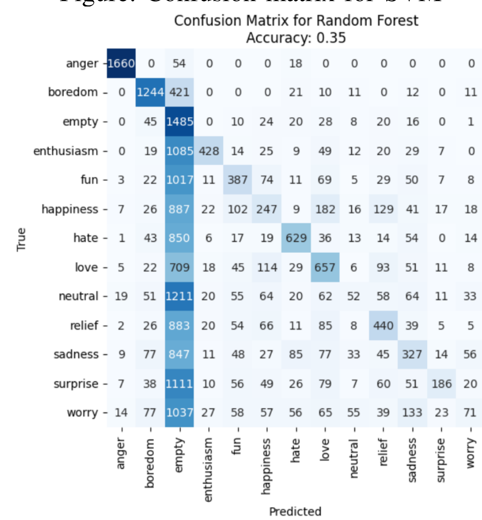


Figure: Confusion matrix for Random Forest

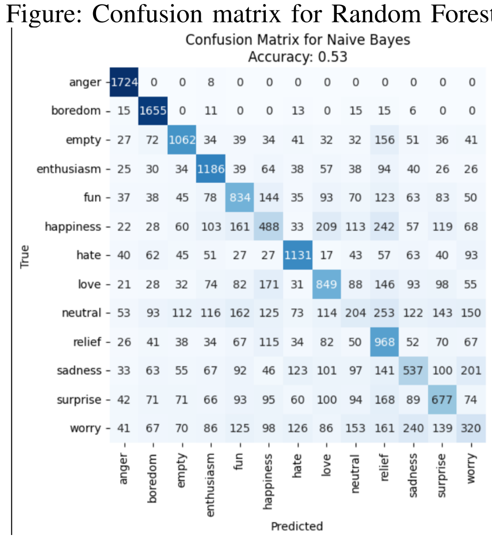


Figure: Confusion matrix for Naive Bayes

Moreover, EmoSense showcases nuanced emotion classification, consistently outperforming individual models as evidenced by improved accuracy scores. Notably, each model demonstrates varied strengths across emotion categories, with SVM excelling in generalization, Naive Bayes showcasing probabilistic simplicity, and Random Forest adapting well to noisy data. EmoSense successfully capitalizes on these strengths, paving the way for more refined sentiment analysis, user engagement prediction, and personalized content delivery in diverse applications of natural language processing.

## VI. Conclusion

In conclusion, EmoSense presents a novel approach to emotion recognition from text, harnessing the power of machine learning models like Support Vector Machines, Random Forests, and Naive Bayes. Through careful dataset curation, feature engineering, and K-fold cross-validation, EmoSense demonstrates improved accuracy in classifying emotions compared to individual models. Its potential applications span diverse industries, promising context-aware responses and personalized interventions in various contexts, from marketing to mental health support. This approach stands as a significant contribution to natural language processing, offering insights into the complex realm of emotions within textual data.

## References


[1] B. Agarwal and N. Mittal, "Machine Learning Approach for Sentiment Analysis," in *Prominent Feature Extraction for Sentiment Analysis*, Socio-Affective Computing, Springer, Cham, 2016.

[2] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, Tiruchengode, India, 2013, pp. 1-5. doi: 10.1109/ICCCNT.2013.6726818.

[3] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports*, vol. 2, p. e12189, 2020.

[4] K. Shrivastava, S. Kumar, and D. K. Jain, "An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network," *Multimedia Tools and Applications*, vol. 78, pp. 29607-29639, 2019.

[5] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, vol. 62, pp. 2937-2987, 2020.

[6] A. Majeed, H. Mujtab, and M. O. Beg, "Emotion detection in Roman Urdu text using machine learning: Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering," 2020, pp. 125-130.

[7] S. Aleem, N. ul Huda, R. Amin, S. Khalid, S. S. Alshamrani, and A. Alshehri, "Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions," *Electronics*, vol. 11, no. 7, p. 1111, 2022.

[8] A. H. Saffar, T. K. Mann, and B. Ofoghi, "Textual emotion detection in health: Advances and applications," *Journal of Biomedical Informatics*, vol. 137, p. 104258, 2023.

[9] B. Desmet and V. Hoste, "Emotion detection in suicide notes," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6351-6358, 2013. doi: 10.1016/j.eswa.2013.05.050.