IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Deep knowledge-Guided Sentiment Prediction Using Content Contextual Learning

Dr.V. Gowri

Dept. of Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

Mulaka Sivarami Reddy

Dept. of Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

Karanam Manoj Kumar

Dept. of Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

Mannam Bhanuprakash

Dept. of Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

Abstract – Sentiment analysis on platforms like Twitter and Facebook is crucial for understanding user opinions and preferences. Despite its importance, the accuracy of sentiment analysis is often hindered by the complexities of natural language processing (NLP). Deep learning models, which outperform traditional statistical and lexical approaches, play a key role in advancing NLP tasks. An essential component of these models is word embedding, which helps in generating input features. There are various models for word embedding that cater to both classic and contextual text representations. In this paper, we conduct a comparative study of the most commonly used word embedding techniques, both in their trained and pre-trained forms, covering both classical and contextualized methods.

Current sentiment classification strategies, especially those for short texts, tend to increase the feature space by incorporating an external open knowledge base. Yet, many of these

strategies depend on extensive training data to build the model, leading to high data collection costs and suboptimal learning performance. We propose that there is a strong correlation between knowledge and text labels, and that leveraging this knowledge explicitly could enhance the efficiency of text classification.

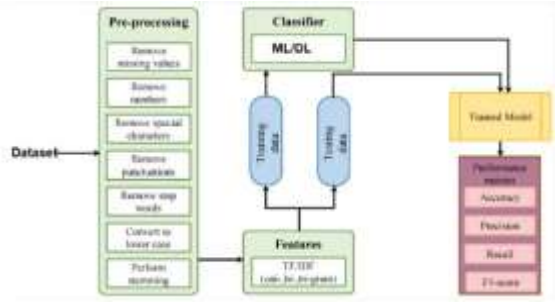
Key Words: **Sentiment Analysis, Convolution Neural Network, Long Short-Term Memory.**

I. INTRODUCTION

Social media platforms have become vital for businesses and organizations to gather feedback through user-generated content, including posts on social media and blogs, which may contain text, audio, visual, or a mix of these elements. Social media text, often characterized by short, informal, and semi-structured sentences, presents challenges in creating accurate vector representations and classifying sentiment. Sentiment analysis, a key technique in big data analytics, processes this text data to extract valuable business insights. It involves classifying texts into predetermined opinion categories and can be conducted at the document, sentence, or word level. Most social media sentiment analysis is performed at the sentence level due to the brevity of the posts, typically under forty words. Current tools for analyzing such brief social media texts are limited, mainly because these texts are unstructured, complicating feature extraction during the text representation phase. According to Zhiying Jiang et al., text representation, which follows data preprocessing, involves converting documents or sentences into numeric vectors using vector space models (VSM). Sentiment analysis is an established branch of text classification within the field of natural language processing (NLP). It automates the classification of texts based on predefined target polarities—positive, negative, or neutral—and examines opinions, sentiments, and emotions expressed in written text regarding entities and their attributes. This analysis is crucial for businesses, governments, and researchers to extract public sentiment, aiding in informed decision-making, especially in e-commerce where timely feedback can influence decisions.

The explosive growth of internet and mobile social networking technologies has led to a surge in textual data online, necessitating effective organization and management. Text classification serves this purpose by organizing text based on predefined standards, using deep learning algorithms like recurrent neural networks and one-dimensional convolutional neural networks to process text sequences efficiently. Transfer learning has notably enhanced NLP by pre-training language models like BERT and ELECTRA on extensive unsupervised data sets, which then undergo fine-tuning for specific tasks. However, most existing training corpora are imbalanced in terms of category quantity and features within categories, which can skew classifier performance and cause misjudgments. To address these challenges, a multi-stream neural network model incorporating background knowledge is proposed. This model uses keywords and frequently co-occurring words as inputs to supplement and reinforce the basal data stream, improving classification performance in both Chinese and English datasets significantly.

In conclusion, while deep learning-based word embeddings offer advanced vector representation for text classification, they are limited by their need for large corpora and their inability to consider word context or semantic orientation. Enhancements are possible by integrating conventional NLP techniques such as sentiment lexicons, POS tags, and word positioning to refine the performance of sentiment analysis models based on word embeddings.



II. LITERATURE SURVEY :

The 2022 study "Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods" by Iqra Ameer et al. introduces a benchmark corpus for multi-label emotion classification on code-mixed SMS messages. It includes 11,914 messages in English and Roman Urdu, annotated with 12 emotions, facilitating research in under-resourced languages.

In the 2022 study "OGSSL: A Semi-Supervised Classification Model Coupled With Optimal Graph Learning for EEG Emotion Recognition" by Yong Peng et al., the OGSSL model unifies graph learning and emotion recognition, showcasing improved performance and feature selection in EEG data analysis.

In the 2021 study "Semi-Skipping Layered Gated Unit and Efficient Network: Hybrid Deep Feature Selection Method for Edge Computing in EEG-Based Emotion Classification" by Muhammad Adeel Asghar et al., a novel feature selection approach enhances efficiency in EEG-based emotion classification for Edge Computing.

In the 2021 study "Autoencoder With Emotion Embedding for Speech Emotion Recognition" by Chenghao Zhang and Lei Xue, a novel algorithm is proposed, combining autoencoder and emotion embedding techniques to extract deep emotion features, enhancing speech emotion recognition performance.

In the 2021 study "Photogram Classification-Based Emotion Recognition" by Juan Miguel López-Gil and Néstor Garay-Vitoria, a method for facial emotion recognition based on parameterized photograms and machine learning is proposed, achieving high emotional photogram classification rates in facial emotion recognition tasks.

In the 2021 study "Deep learning-based classification of the polar emotions of moe-style cartoon pictures" by Qinchen Cao, Weilin Zhang, and Yonghua Zhu, a deep learning-based method is proposed to classify polar emotions of moe-style cartoon pictures, addressing challenges in the cartoon animation industry by achieving competitive experimental accuracy.

III. PROBLEM STATEMENT

Sentiment classification, a fundamental task in NLP, involves categorizing text into predefined sentiment categories. Each input sentence, denoted as x = (w1, w2, w3, ..., wn), comprises a sequence of words, where wi ∈ V and V represents the vocabulary. Our objective is to classify the sentiment into pre-defined categories Y. However, the availability of annotated data is limited in the model setting, with only a few training samples $(xi, yi)^N j = 1$. Our aim is to learn a sentiment classification model from this dataset D, enabling us to predict the label of any unseen sentence. Alternatively, in scenarios where no training samples are available, our goal is to infer sentiment labels directly from the input sentences.

IV. EXISTING SYSTEM

ArEmotive, or Arabic Emotive, is an advanced sentiment analysis system that operates without human input and continually grows its dataset and precision. It expands its knowledge by integrating new sources through ontology augmentation and classification. By utilizing various data sources, it maintains a central repository accessible on mobile devices. It's particularly significant in automated ontology alignment and mapping, using semi-automated methods to integrate new data seamlessly.

ArEmotive stands out in identifying nuanced emotions in text through a dynamic ontology continuously enriched by machine learning. It goes beyond simple positive or negative categorizations, recognizing complex emotional states and employing zero-shot classification. The system adapts to different data structures and properties using semantic and syntactic similarity calculations.

Future plans for ArEmotive include incorporating ensemble learning, extending multilingual capabilities, tuning hyper-parameters for specific applications, and expanding ontology resources. The base ontology and source code are openly available, and there are no conflicts of interest or external funding involved. Three researchers contributed to data annotation during development.

V. PROPOSED SYSTEM

The study investigates the application of word embedding techniques for sentiment classification by first tokenizing textual reviews, followed by generating vector space representations of each word to form an embedding layer for deep neural networks. For the experiments, both advanced pre-trained word embeddings and custom embeddings trained on specific datasets were utilized. Datasets of various lengths and characteristics, including hotel reviews and topics ranging from sports to politics and the arts, were selected to study the impact of word embeddings across different domains.

Preprocessing is crucial in sentiment analysis to standardize input text for effective processing, involving cleaning the text to remove irrelevant characters and symbols to enhance model accuracy. Word embedding, which transforms text into dense vector representations, has proven superior to traditional methods like bag-of-words, particularly in deep learning applications. Parameters in embedding can vary in vector length and include trainable weights that adapt during the training phase.

The study used embeddings like GloVe, which constructs vectors using a co-occurrence probability matrix and contextual windows to capture word relations, showing superior performance in tasks like word analogies and named entity recognition. However, methods like Word2Vec and GloVe face limitations in addressing word ambiguity and only capture context within a specific window, affecting their effectiveness.

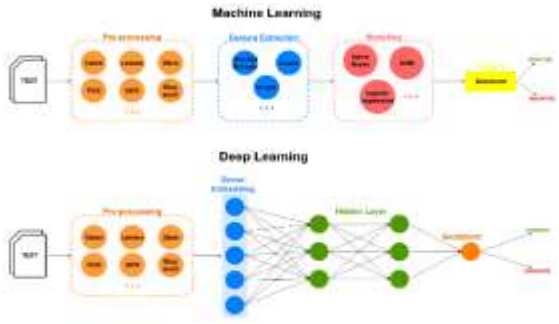
To enhance sentiment analysis at the sentence and document levels, the architecture employs bidirectional LSTM layers, with the first layer deriving sentence sentiment from GloVe embeddings and the second layer aggregating these into document-level sentiment representations. A two-layer attention mechanism is used to emphasize relevant words and sentences, filtering out less important elements.

Key contributions of the research include:

- The introduction of a sentiment-enhanced word embedding method that maps words to their sentiment orientations through vector addition.

- Evaluation of performance across various benchmark datasets, including those with diverse review lengths, Arabic language formats, and different domains, addressing the challenge of dataset diversity.

VI. MODULE DESCRIPTION



Module 1 : Module 1: Text Preprocessing

Data preprocessing is crucial for enhancing model performance, especially when dealing with noisy data from platforms like Twitter, which may have null values and diverse media types like images and videos. While BERT-based models utilize all information in a sentence, including punctuation and stop words, other models require preprocessing.

In this research project, Python libraries and functions were used for preprocessing, including steps like converting text to lowercase, removing punctuation, usernames, special characters, and emojis, deleting hashtags' hash symbols, eliminating English stop words, and tokenization. Additionally, rows with missing labels were dropped.

Data cleaning offers numerous benefits, including improved data quality by removing incorrect or inconsistent information. It addresses issues like typos in feature names, mislabeled classes, and extra spaces, ensuring the dataset's integrity before model training.

Module 2: Module 2: Global Vectors for Word Representation

Once the data is cleaned and preprocessed, the next step is to convert it into a format that the model can understand, known as feature extraction or vectorization. This process involves transforming all variables into numerical form, which helps reduce dimensionality and improve model accuracy by focusing on relevant features. Feature extraction methods assist in assessing the significance of words in the dataset and removing redundant data. Moreover, new features can be generated from existing ones, enhancing the overall quality of the dataset.

In this investigation, the method of Count Vectorizer was applied for text vectorization, transforming text data into numerical vectors. Furthermore, GloVe (Global Vectors for Word Representation) embeddings were utilized. GloVe embeddings rely on a co-occurrence matrix, which captures word associations within a given text corpus. Subsequently, the vectors derived from these embeddings underwent a dimensionality reduction process using singular value decomposition (SVD). This transformation into a lower-dimensional space aims to retain essential information while reducing computational complexity and noise within the data. By employing these techniques, the study sought to effectively represent textual information in a numerical format suitable for subsequent analysis and modeling tasks.

Module 3 : Module 3: Content Contextual Learning Network

The Content Contextual Learning Network, a sequential processing model, consists of two layers that propagate data both forward and backward simultaneously. LSTM, a component within this model, leverages past and future context information to capture long-term dependencies and contextual features in text, making it highly effective for tasks like text classification.

Within this architecture, a CNN deep learning model functions as the classifier, taking LeBERT embedding vectors as input and generating

sentiment class outputs. CNNs excel in supervised learning tasks by detecting and learning characteristic patterns through convolution layers, followed by pooling operations to select optimal features. The dense output layer then categorizes these features into positive or negative classes using softmax activation.

Word embedding plays a vital role in reducing the dimensionality of text data and providing dense vectors for subsequent neural network layers. Unlike one-hot encoding, which is computationally expensive due to its many zeros, word embedding efficiently represents text data by assigning learned weights.

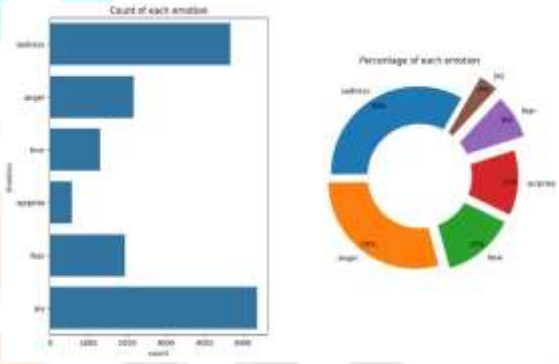
Before building the models, hyperparameters need to be selected and fine-tuned. Crucial parameters include batch size, epochs, optimizer, dropout rate, and classifier type. Grid search was used to assess different options, with optimal performances achieved with a batch size of 128, Adam optimizer with a learning rate of 0.001, 50 training epochs, a dropout rate of 0.2, and Softmax function for output logits.

Module 4 : Module 4: Support Vector Machine

The Support Vector Machine (SVM), pioneered by Cortes and Vapnik primarily for binary classification tasks, operates as a linear classifier devoid of probability estimations. It delineates hyperplanes or a series of hyperplanes within a high-dimensional domain to delineate data points. Renowned for its versatility, SVMs find utility across classification, regression, and diverse computational tasks. Leveraging kernel mapping techniques, SVMs facilitate the transformation of data from the input space into a higher-dimensional feature space, wherein the inherent problem achieves linear separability. This process allows SVMs to effectively handle datasets that may not be linearly separable in their original form, thereby enhancing their applicability to a wide array of real-world problems.

In sentiment classification, SVMs seek to locate a hyperplane that effectively divides documents or datasets based on sentiment, with a maximized margin between sentiment classes (e.g., positive, negative, neutral). Data points situated on the separation boundaries are referred to as support vectors.

VII. RESULT AND DISCUSSION:



The project would present the performance of the proposed model compared to existing methods. This could include baseline models, traditional machine learning approaches, and possibly other deep learning architectures. The results might show improvements in sentiment prediction accuracy, especially in cases where context plays a crucial role in determining sentiment.

VIII. CONCLUSION

The paper introduces the sentiment-enhanced word embedding method, aimed at improving sentiment classification tasks. This technique establishes connections between words in sentiment lexicons and their corresponding sentiment orientations. By training a sentiment mapping matrix and integrating it with word embeddings, sentiment-enhanced word embeddings are generated. These enhanced embeddings are then utilized within sentiment classification models to determine the sentiment orientations of sentences in datasets. Experimental findings demonstrate that models leveraging Word2Vec and GloVe embeddings enhanced by the sentiment-enhanced word embeddings exhibit superior performance compared to those using original embeddings,with satisfactory convergence times. Notably, the method solely enhances the embedding layer without modifying the underlying model architecture, ensuring compatibility with existing frameworks. However, it's important to acknowledge that while the method improves the performance of standard word embeddings, it doesn't significantly enhance contextualized word embeddings. Overall, the sentiment-enhanced word embedding method presents a promising approach to enhancing sentiment classification tasks, providing researchers and practitioners with an effective tool for capturing nuanced sentiment orientations in text data.

IX. REFERENCES

[1] Amer Jaradeh,Mohamad-Bassam Kurdy ArEmotive Bridging the Gap: Automatic Ontology Augmentation Using Zero-Shot Classification for Fine-Grained Sentiment Analysis of Arabic Text IEEE Access, 2023

[2] Iqra Ameer,Grigori Sidorov,Helena GÃ³mez-Adorno,Rao Muhammad Adeel Nawab Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods IEEE Access, 2022

[3] Yong Peng,Fengzhe Jin,Wanzeng Kong,Feiping Nie,Bao-Liang Lu,Andrzej Cichocki OGSSL: A Semi-Supervised Classification Model Coupled With Optimal Graph Learning for EEG Emotion Recognition IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2022

[4] Xueqiang Zeng,Qifan Chen,Xuefeng Fu,Jiali Zuo Emotion Wheel Attention-Based Emotion Distribution Learning IEEE Access, 2021

[5] Fuji Ren,Qian Zhang An Emotion Expression Extraction Method for Chinese Microblog Sentences IEEE Access,2020

[6] Chenghao Zhang,Lei Xue Autoencoder With Emotion Embedding for Speech Emotion Recognition IEEE Access, 2021