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

A Semantic Significantly Improved Vector Space Model for Text Classification using Progressive Learning Network Algorithm

Mr. K. Rajesh, M.Tech
Assistant Professor, CSE
SRM IST, Ramapuram
Chennai, India

Rachagolla Likhith
Dept. Of Computer Science and Engineering
SRMIST, Ramapuram
Chennai, India

Dinesh Bala S
Dept. Of Computer Science and Engineering
SRMIST, Ramapuram
Chennai, India

Gopesh M
Dept. Of Computer Science and Engineering
SRMIST, Ramapuram
Chennai, India

***Abstract*—With digital data being increased on the internet, the text classification has become a very important thing. In Natural Language processing, text classification is being considered as a fundamental task with several applications across domains starting from sentiment analysis, document categorization, and spam detection. Even though the hierarchical classification stands best in classifying heterogenous data, Alternative Relative Discrimination (ARDC) has been proven as a better approach, which focuses on identifying terms frequently occurring in positive class. This paper presents a novel neural network-based technique for text classification which is optimized to mitigate existing text classification issues. We introduce improved Vector Space Model with Progressive Learning Network Algorithm (PLN). The current approach enhances traditional Vector Space Model by incorporating semantic information through advanced techniques such as word embeddings, contextual embeddings, and semantic similarity measures. This enriched representation enables the model to capture complex relationships and context dependencies within the text, leading to more accurate and nuanced classification results. Furthermore, we introduce the Progressive Learning Network algorithm, which facilitates continual learning and adaptation to new data without forgetting previously learned knowledge. By dynamically updating model parameters and representations over time, PLN ensures the adaptability and scalability of the SSIVSM in evolving text classification tasks.**

***Keywords*—Alternative Relative Discrimination Criterion, Progressive Learning Network Algorithm, Vector-Space-Model**

I. INTRODUCTION

Text classification is essential for organizing and accessing vast amounts of electronic data, utilizing AI, NLP, and machine learning techniques. Hierarchical text classification organizes classes into categories and subcategories, resembling real-world systems like web directories and organizational structures using Progressive Learning Algorithm. Multilingual text classification frameworks are crucial for disaster mitigation, enabling the identification and categorization of actionable information across various languages. Rule-based, data-driven (machine learning/deep learning), and hybrid approaches are among the automated text classification techniques available; machine learning is particularly successful in this regard.

Traditional text classification methods often rely on representations such as Alternative Relative Discrimination (ARDC), which focuses on classifying the data into the positive classes, the Vector Space Model (VSM), which represent documents as high-dimensional vectors in a semantic space. While these methods have been widely used and studied, they often struggle to capture the semantic nuances and contextual information present in textual data. This limitation can lead to suboptimal performance, particularly when dealing with complex and diverse text corpora.

To address these challenges, there has been growing interest in developing more advanced and sophisticated approaches to text classification that leverage semantic information and context-aware representations. In this context, this research paper proposes a novel approach to text classification, namely the **Semantic Significantly Improved Vector Space Model (SSIVSM)**, augmented with the Progressive Learning Network (PLN) algorithm.

The SSIVSM extends traditional VSM by incorporating advanced techniques such as word embeddings, contextual embeddings, and semantic similarity measures. These enhancements enable the model to capture complex relationships and context dependencies within textual data, leading to more accurate and nuanced classification results. Additionally, the PLN algorithm facilitates continual learning and adaptation to new data without forgetting previously learned knowledge, ensuring the adaptability and scalability of the SSIVSM in evolving text classification tasks.

In this paper, we present the methodology, implementation details, and experimental results of the proposed approach. We evaluate the performance of the SSIVSM with PLN on benchmark datasets across different text classification domains and compare it with traditional methods and state-of-the-art approaches. Our findings demonstrate the effectiveness and potential of the proposed approach in addressing the challenges of text classification and advancing the state-of-the-art in NLP research.

II. Literature Survey

In order to improve sentiment classification by obtaining specific local characteristics from text, a convolutional neural network with various convolutions and pooling layers is introduced in the paper Variable Convolution and Pooling Convolutional Neural Network for Text Sentiment Classification (VCPCNN). [1]

A multi-Kernel CNN model with n-gram word embedding is presented in the publication A Superior Arabic Text Categorization Deep Model (SATCDM) for Arabic text categorization. Its accuracy, which ranges from 97.58% to 99.90%, is extremely high, exceeding earlier studies. With its effective design and skip-gram word embedding enhanced with sub-word information, SATCDM outperforms other Arabic datasets in terms of performance. [2]

The study focused on automating Arabic Bloom's taxonomy classification in the paper named Arabic Questions Classification Using Modified TF-IDF, creating a dataset and proposing an enhanced TF-IDF method for feature extraction. Results showed the proposed method significantly outperformed traditional and modified TF-IDF methods for English questions, indicating its potential for accurately classifying Arabic assessment questions. [3]

Paper titled, 'A review in feature extraction approach in question classification using Support Vector Machine' published in 2014 states that, the text highlights the challenge of question classification within Bloom's taxonomy due to overlapping verb keywords. It proposes an integrated approach combining semantic features with Support Vector Machine classification to improve accuracy, aiming to better categorize questions based on cognitive levels. [4]

In an effort to overcome the drawbacks of conventional vector space models, scholars have looked at the application of semantic representations like word and contextual embeddings. Word embeddings are created by utilizing methods such as Word2Vec, GloVe, or FastText to represent words as dense, connected strings. Vectors with modest dimensions in a continuous vector space. Models like ELMo and BERT are examples of contextual embeddings, which produce word representations that are sensitive to the context in which they appear and can thus capture more complex semantic information. [5]

Progressive learning algorithms aim to enable models to continually learn and adapt to new data without forgetting previously learned knowledge. These algorithms are particularly relevant in scenarios where the data distribution is non-stationary or evolves over time. One such approach is the Progressive Learning Network (PLN), which incrementally updates model parameters and representations based on new data while preserving previously learned knowledge. PLN has been applied successfully in various machine learning tasks, including classification, regression, and anomaly detection. [6]

The paper titled 'Text Classification Based on Conditional Reflection' introduces RCNNA, a text classification model inspired by human conditional reflexes, combining BLSTM, attention, and CNNs. It outperforms baseline methods on Chinese and English datasets, emphasizing future neural network structure exploration. [7]

III. Existing Methodology

These days, electronic information extraction tools are everywhere—from automatic archives with millions of data to instant messaging apps on mobile devices. The abundance of data has resulted in a considerable number of issues. But one initiative is to automatically classify a portion of this text material in order to facilitate user access, interpretation, and modification of data for the purpose of creating patterns and knowledge.

A growing number of people and businesses are interested in the problem of categorizing vast amounts of electronic data. Text classification is the only way to solve this issue. The process of classifying documents based on their categories is called text classification. It makes use of several different specialties, such as AI, NLP, and machine learning.

It makes use of a supervised learning-based methodology in which a lot of data is provided to train a model. A number of methods, including SVM and k-Nearest Neighbor, have been proposed recently. Bayes without learning. The outcomes demonstrate how successful these methods are in traditional text classification settings.

Text categorization has numerous applications, including spam detection, topic modelling, sentiment analysis, and intent detection.

Tasks involving text classification have few classes. Text classification problems are extended when there are several classes in the classification task because they present unique problems that call for unique solutions. There are several important problems with the hierarchical classification system, which consists of related categories. Hierarchical classification therefore emerges in this situation. When we apply hierarchical classification to textual data, we get hierarchical text classification. In hierarchical classification, classes are arranged into categories and subcategories, or in other words, classes are arranged into a class hierarchy.

These days, a variety of apps use a hierarchical framework for document organization. If we use an actual case study The process of classifying texts hierarchically is similar to that of a librarian organizing books on a shelf. Numerous Businesses in the IT, legal, and medical fields are among those that benefit from automatic document classification. Consequently, it demonstrates that hierarchical classification has a big influence on a lot of applications and organizations.

Supervised machine learning or manually defined rules can be used for text classification. The latter automatically learns the mapping from an input to the correct output, whereas the former is susceptible to changes in data or circumstances. In the watchful machine learning methodology Typically, a bag-of-words model is used to represent text; characteristics are then extracted and supplied into a classifier.

After a tragedy, people use online social networks to relay relevant information and ask for assistance. Effective rescue and relief operation planning can be aided by the identification and classification of this useful data. Nevertheless, the majority of these user-generated materials, like tweets, are written in the local tongue of the catastrophe. However, the vast bulk of literary works only concentrate on the English language. Comprehending and analyzing documents written in various languages is crucial for efficient disaster relief efforts.

The potential of multilingual systems to function in multiple languages will increase their usefulness for relief and rescue efforts. This makes the need for a multilingual text categorization system that can recognize and classify practical and actionable data produced in the wake of tragedies. A further obstacle in developing such an automated system is the deficiency of adequate labeled data in the field of catastrophe mitigation. During a crisis, it may not be possible or cost-effective to label instances.

IV. Proposed Methodology

In the preprocessing data phase, the text undergoes filtering to remove punctuation symbols, which can otherwise hinder the accuracy of the classification process. This step is crucial for enhancing the algorithm's performance. Leveraging the Natural Language ToolKit (NLTK) package facilitates efficient removal of punctuation marks, followed by word count calculation. Feature selection, crucial for constructing an efficient model, involves identifying relevant variables.

Moving on to the vectorizing textual inputs phase, the focus lies on word-level neural models. This approach divides text into words, with each word converted into a word vector. Unlike one-hot encoding, word embeddings generate low-dimensional, distributed representations of words. These embeddings reveal semantic similarities based on the distributional hypothesis, which posits that words appearing in similar contexts share similar meanings. Integration of word embeddings into the proposed system excels the performance in various NLP duties.

Our approach to developing a classification model for multi-class tasks involves a systematic methodology integrating data preprocessing, feature extraction, normalization, and evaluation metrics. The methodology unfolds as follows:

A. Text preprocessing

Initially, we clean the pertinent text content to ensure quality and consistency in our dataset. This step is crucial for subsequent analysis and feature extraction.

B. Feature Extraction

Leveraging various data sources, extracting the content features from the reliable dataset, user information will be extracted from user features, by constructing a spreading tree, we extract the propagation features. Furthermore, semantic features extracted from the original dataset are concatenated to form a comprehensive high-dimensional feature vector.

C. Feature Selection

To mitigate overfitting and reduce redundancy in our feature space, we employ feature selection techniques to identify the optimal subset of features. This not only prevents overfitting but also enhances the convergence speed of our algorithm by avoiding local optima.

D. Normalization

Given the disparate weight ranges of the different types of statistical features, normalization becomes

imperative. Through normalization, we standardize the feature values, ensuring uniformity across the feature space and facilitating effective classification. The normalization process also considers negative sentiment feature values for comprehensive coverage.

E. *Evaluation Metrics*

To gauge the effectiveness of our model, we employ accuracy rate (Acc) and F-score as evaluation metrics. These metrics provide insights into the model's performance across various classification tasks.

V. RESULTS

New system efficiently processes text data by cleaning and extracting features, incorporating user information and propagation patterns. Through feature selection and normalization, it minimizes overfitting and ensures uniformity across disparate feature ranges. Evaluation metrics like Accuracy Rate and F-score validate its effectiveness in multi-class classification tasks.

The new model with networking of neurals, with the support of phrases being connected performs well than earlier Term-frequency based methods, providing context-aware representations and facilitating faster decision-making. The system excels in distributed optimization, delivering reliable outcomes with explainability and guiding label assignment confidently. Its strong generalization capabilities reduce human effort significantly, making it an asset across various applications.

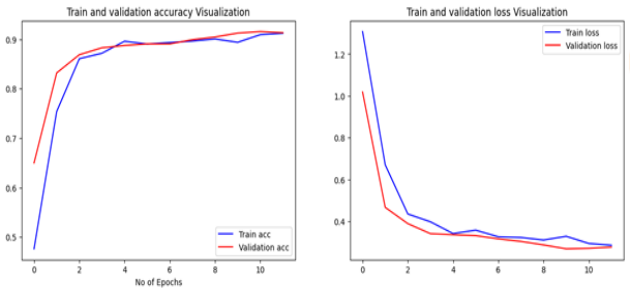

Fig: Accuracy and validation loss in the proposed model with progressive learning network algorithm.

VI. CONCLUSION

Covered the difficulties and fixes related to the text classification problem in this study. In order to improve classification performance, pre-processing steps are essential, even though the task could be quite simple in some languages. We displayed a word embedding model that records semantic parallels at the sub-word level amongst words. Spelling problems resulting from the previous segmentation stage can be handled by the word embedding model because it employs sub-words. We suggested neural network models for text classification that make use of the word embedding model. The word embedding model neural network models consistently beat the baseline TF-IDF model, as demonstrated by the experimental findings using multiclass classification datasets.

Explored the complexities of the text classification assignment, looking at the problems it poses and how to solve them. Although the classification procedure may appear simple in some languages, we stress how important pre-treatment steps are to improving classification results. In particular, we emphasize how important it is to use efficient pre-processing methods in order to maximize the classification task's accuracy and efficiency.

The outcomes of our experiments highlight the neural network models that are outfitted with the word embedding model consistently outperforming baseline models that depend on TF-IDF representations. This actual data supports our method's effectiveness in obtaining better classification accuracy and robustness across a range of datasets and text classification scenarios. All things considered, our research advances the field of text classification with insightful discoveries and useful fixes, opening the door to improved effectiveness and efficiency in practical settings.

REFERENCES

[1] A. Adeleke, N. A. Samsudin, Z. A. Othman, and S. K. A. Khalid, ‘‘A two step feature selection method for quranic text classification,’’ Indonesian J. Electr. Eng. Comput. Sci., vol. 16, no. 2, pp. 730–736, 2019.

[2] L. Jin and L. Zhang, ‘‘De-redundancy relative discrimination criterion based feature selection for text data,’’ in Proc. Int. Joint Conf. Neural Netw. (IJCNN), May 2023.

[3] J. Piri, P. Mohapatra, M. R. Pradhan, B. Acharya, and T. K. Patra, ‘‘A binary multi-objective chimp optimizer with dual archive for feature selection in the healthcare domain,’’ IEEE Access, vol. 10, pp. 1756–1774, 2022.

[4] H. Banka, H. Mahmood, K. Fatih, S. Mehmet, and V. Üniversitesi, A Hybrid Feature Selection Approach Based on LSI for Classification of Urdu Text, vol. 907. Cham, Switzerland: Springer, 2021.

[5] A. Kumar and A. Jaiswal, ‘‘Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on Twitter,’’ Multi media Tools Appl., vol. 78, no. 20, pp. 29529–29553, Oct. 2019.

[6] S. Lefkovits and L. Lefkovits, ‘‘Gabor feature selection based on information gain,’’ Proc. Eng., vol. 181, pp. 892–898, 2017.

[7] A. K. Uysal and S. Gunal, ‘‘A novel probabilistic feature selection method for text classification,’’ Knowl.-Based Syst., vol. 36, pp. 226–235, Dec. 2012.

[8] S. Z. Mishu and S. M. Rafiuddin, ‘‘Performance analysis of supervised machine learning algorithms for text

classification,'' in Proc. 19th Int. Conf. Comput. Inf. Technol. (ICCIT), Dec. 2016.

[9] A. Rehman, K. Javed, H. A. Babri, and M. Saeed, ''Relative discrimination criterion—A novel feature ranking method for text data,'' Exp. Syst. Appl., vol. 42, no. 7, pp. 3670–3681, May 2015.

.