IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306147|

A Review of Advances in Text Classification through Image Processing Techniques

Pooja[1], Meenakshi Arora[2], Rohini Sharma[3]

P.G. Student, Department of CSE, Sat Kabir Institute of Technology and Management, Haryana, India[1]

Assistant Professor, Department of CSE, Sat Kabir Institute of Technology and Management, Haryana, India[2]

Assistant Professor, Department of CS, GPGCW, Rohtak, India[3]

ABSTRACT: Text classification has seen significant advancements through the integration of image processing techniques, leveraging the visual and spatial characteristics inherent in textual data representations. This review surveys recent developments in applying image processing methods to enhance the accuracy, efficiency, and interpretability of text classification systems. The integration of image processing techniques offers novel approaches to preprocess textual data, extract meaningful features, and improve classification performance. This review synthesizes key methodologies and evaluates their effectiveness across various domains of text classification, including sentiment analysis, document categorization, and information retrieval. Challenges such as data heterogeneity, model interpretability, and computational efficiency are discussed alongside promising avenues for future research. By exploring the convergence of image processing techniques with text classification, this review aims to provide insights into the evolving landscape of multimodal data analysis and its implications for advancing artificial intelligence applications.

KEYWORDS: Image processing, Text Classification

I. INTRODUCTION

Text classification, a fundamental task in natural language processing (NLP), has traditionally relied on textual features extracted from corpora to categorize documents, sentiment, or topics[1]. Recently, there has been a growing interest in enhancing text classification tasks by integrating image processing techniques. This integration leverages the visual information embedded within textual data representations, offering new avenues to improve classification accuracy, robustness, and interpretability[2]. The use of image-based approaches in text classification represents a departure from traditional methods that rely solely on textual features[3]. By treating textual data as images, researchers explore the application of Convolutional Neural Networks (CNNs) and other image processing techniques originally developed for computer vision tasks[4]. These techniques aim to capture spatial relationships, semantic structures, and visual patterns within textual documents, complementing traditional textual features[5].

This introduction provides a comprehensive overview of recent advancements in image-based text classification, highlighting the transformative impact of integrating visual information with textual analysis[6]. It explores key methodologies, including text-to-image conversion, feature extraction from visual representations of text, and hybrid models that combine textual and visual modalities[7].

The integration of image processing techniques in text classification not only expands the scope of NLP applications but also addresses inherent challenges such as data heterogeneity, model interpretability, and computational efficiency. By synthesizing current research findings and discussing future research directions, this paper aims to contribute to the evolving landscape of multimodal data analysis and its implications for advancing artificial intelligence applications in textual domains. Machines can now understand and communicate text meaning thanks to natural language processing, or NLP. The media, finance, healthcare, and other fields are all impacted by NLP.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306147|

II. TEXT CLASSIFICATION

In NLP, text classification is the process of classifying text documents, sentences, or phrases according to their content and giving them predetermined labels or subcategories. The goal of text categorization is to identify a text's class or category autonomously. It's a basic NLP problem with many real-world uses, such as language identification, sentiment analysis, spam detection, topic categorization, and more. In order to accurately forecast a text's category, text classification algorithms examine the text's properties and patterns. This allows machines to sort, filter, and comprehend vast amounts of textual data.

III. RULE BASED TEXT CLASSIFICATION SYSTEM

A rule-based text classification system relies on predefined rules or patterns to categorize text into predefined classes or categories. These systems are particularly useful when the classification task can be clearly defined with explicit rules that capture the characteristics of each category. One of the earliest NLP techniques is the rule-based strategy, which processes and analyzes textual data using pre-established linguistic rules. Using the rule-based method, certain structures are captured, information is extracted, or operations like text classification are carried out by implementing a particular collection of rules or patterns. Pattern matching and regular expressions are two popular rule-based approaches.

Rule Definition: Rule-based systems define classification rules based on domain knowledge, linguistic patterns, or specific features extracted from text. These rules are typically created manually by domain experts and can involve regular expressions, keywords, syntactic structures, or semantic rules.

Feature Extraction: Rule-based systems often extract features from text that are relevant to the classification task. This may include word frequencies, part-of-speech tags, named entities, or other linguistic features that help define the rules.

Rule Evaluation and Application: Text documents are processed through a series of rules that assign them to predefined categories based on the presence or absence of specific features or patterns. The classification decision is based on how well a document matches the criteria defined by the rules.

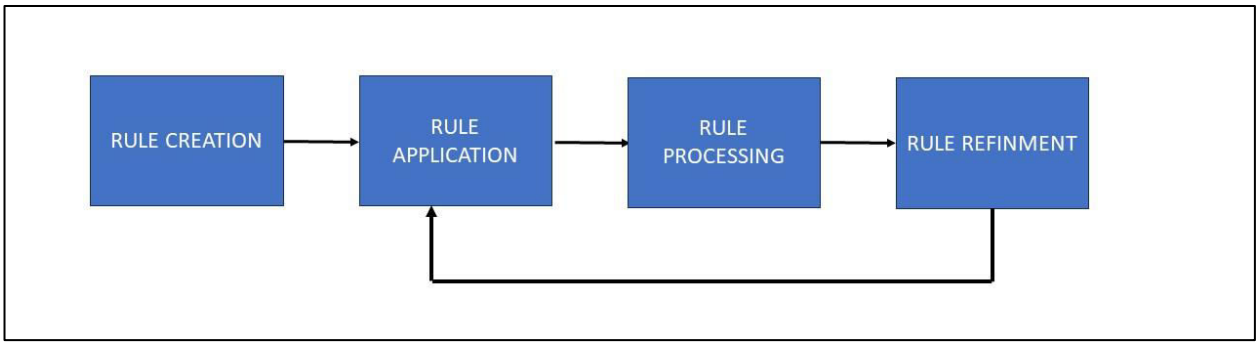

Figure 1: Rule Based NLP Approach

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306147|

Table 1: A comparative analysis of Rule Based NLP Approached for text classification

Aspect	**Resnik & Hearst** [8]	**Manning, Raghavan, & Schütze** [9]	**Mooney** [10]	**Hearst** [11]
Focus Area	Semantic rule learning from syntactic patterns	Information retrieval	Word sense disambiguation	Automatic acquisition of hyponyms
Methodology	Learning semantic rules using syntactic patterns	Comprehensive introduction to IR concepts and algorithms	Comparative experiments on different machine learning methods for WSD	Rule-based and statistical methods for hyponym acquisition
Key Contributions	Demonstrates how syntactic patterns can be used to infer semantics	Foundational text in IR, covering key algorithms and theories	Highlights the impact of bias in machine learning for WSD	Introduces methods to automatically extract hyponyms from text
Techniques	Rule learning, syntactic pattern analysis	Vector space models, search engines, relevance feedback	Machine learning algorithms, evaluation of bias	Rule-based extraction, statistical analysis
Applications	Semantic parsing, NLP	Information retrieval systems, search engines	NLP tasks requiring word sense disambiguation	Knowledge base construction, ontology building
Results	Effective extraction of semantic rules from syntactic structures	Widely adopted textbook, comprehensive coverage of IR topics	Demonstrates the significance of algorithm bias in WSD performance	Effective automatic extraction of hyponym relations from corpora
Strengths	Integration of syntactic and semantic information	Comprehensive and foundational resource	Insight into the role of bias in machine learning	Pioneering work in automatic hyponym extraction
Limitations	Dependence on quality of syntactic parsing	Broad coverage may lack depth in specific areas	Focused primarily on hyponyms, may not generalize to other relations	Dependence on quality of syntactic parsing

IV. MACHINE LEARNING-BASED TEXT CLASSIFICATION

Machine learning-based text classification in images is an innovative approach that combines text processing with image analysis. This method is used in scenarios where text embedded within images (such as memes, scanned documents, or screenshots) needs to be extracted and classified. Classifying a category from input text is significantly faster and easier with the machine learning model. Feature extraction is an essential stage in the application of machine learning. One method we use to convert text to numerical representation in the form of vectors is the use of bags of words, which is essentially counting each word in a text; another method is the application of tfidf (term frequency inverse document frequency).

Methodologies

1. **Text Extraction**:
    - **Optical Character Recognition (OCR)**: The first step involves extracting text from images using OCR technology. OCR tools, such as Tesseract, Google Cloud Vision, and Adobe OCR, convert images of text into machine-readable text.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306147|

- **Preprocessing**: This includes noise reduction, binarization, and text segmentation to improve OCR accuracy.

2. **Text Classification**:

   - **Feature Extraction**: Extract features from the recognized text. This could include traditional methods like TF-IDF (Term Frequency-Inverse Document Frequency) or modern embeddings like Word2Vec, GloVe, or BERT.

   - **Machine Learning Models**: Apply machine learning algorithms for classification. Common models include:

Machine learning-based text classification in images leverages OCR for text extraction and various machine learning techniques for text classification. This hybrid approach is powerful for applications requiring automated text understanding from visual data. The ongoing advancements in OCR accuracy and deep learning models continue to enhance the effectiveness and applicability of these systems.

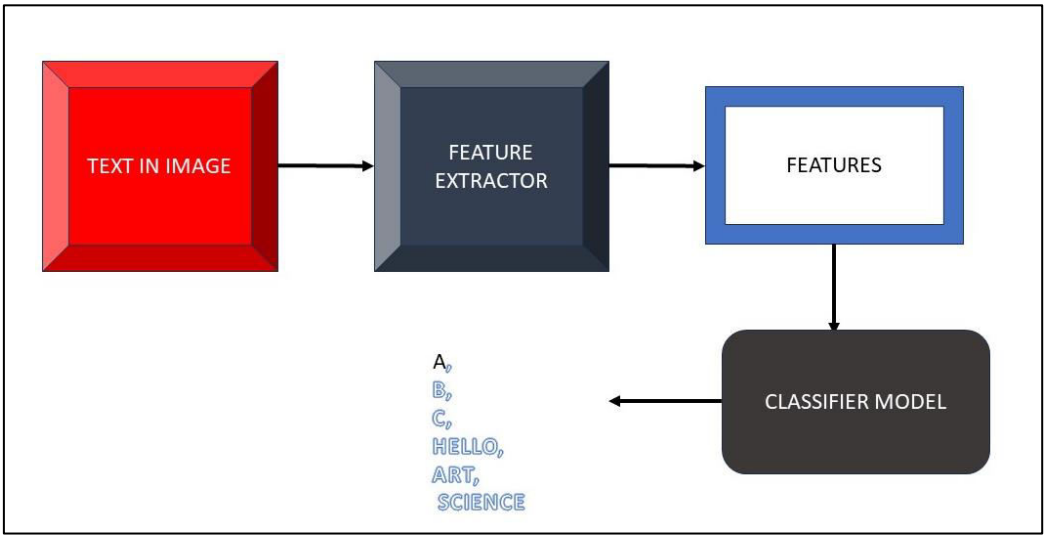

Figure 2: Machine Learning Based Text Classification

Table 2: A comparative analysis of Machine Learning Based NLP Approached for text classification

Aspect	**Smith, R.** [12]	**Patel, A., Dahyot, R.** [13]	**Mikolov, T. et al.** [14]	**Devlin, J. et al.** [15]	**Joachims, T.** [16]	**Yoon Kim** [17]	**Gatt, A., & Krahmer, E.** [18]
Focus Area	Optical Character Recognition (OCR)	Text detection and recognition using deep learning	Word embeddings and compositionality	Pre-training deep bidirectional transformers for language understanding	Support Vector Machines (SVM) for text categorization	Convolutional Neural Networks (CNNs) for text classification	Natural Language Generation (NLG)
Methodolo	OCR	Deep	Development	Pre-training	SVM with	Applicatio	Comprehens

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306147|

gy	technology overview and improvements	learning techniques for text detection and recognition	of word embeddings like Word2Vec	and fine-tuning BERT for various NLP tasks	high-dimensional feature spaces	n of CNNs to sentence classification	ive survey of NLG tasks, methods, and applications
Key Contributions	Detailed explanation of Tesseract OCR engine	Applying CNNs and RNNs for text recognition in images	Introduced Word2Vec, capturing semantic relationships	BERT model that significantly improves performance on various NLP benchmarks	Demonstrated effectiveness of SVM for text classification	Showed that CNNs can effectively classify sentences	Overview of NLG advancements, challenges, and evaluation techniques
Techniques	OCR, character recognition	Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs)	Word embeddings, neural networks	Transformers, masked language modeling	Support Vector Machines, feature selection	Convolutional Neural Networks	Various NLG techniques including rule-based and statistical methods
Applications	Digitization of printed text, document scanning	Text recognition in natural images, scene text detection	NLP tasks requiring semantic understanding	Wide range of NLP tasks including text classification, Q&A, and more	Text categorization for information retrieval	Text classification tasks such as sentiment analysis and topic categorization	NLG in chatbots, automated report generation, and more
Notable Results	Significant improvements in OCR accuracy and speed	Improved accuracy in text detection and recognition in images	High-quality word embeddings capturing semantic nuances	State-of-the-art performance on multiple NLP benchmarks	High accuracy and robustness in text categorization	Demonstrated CNNs as a powerful tool for text classification	Comprehensive review and categorization of NLG research
Strengths	Robust and accurate OCR engine	Advanced deep learning methods improving	Efficient and scalable method for creating word	Highly versatile model for various NLP	Strong theoretical foundation and empirical	Effective feature learning and high classificati	In-depth analysis and wide coverage of NLG

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306147|

		text recognition accuracy	embeddings	applications	performance	on accuracy	techniques
Limitations	Primarily focused on OCR, limited to printed text	Focused on text within images, requires high-quality image data	Requires substantial computational resources for training	Computationally intensive, large pre-training data required	SVMs can be computationally expensive for very large datasets	Requires substantial labeled data for training	Primarily a survey, does not introduce new algorithms or methods

V. HYBRID APPRAOCH OF TEXT CLASSIFICATION

Vector Semantic

Word and sequence analysis can also be done using vector semantics. In order to explain properties like similar and opposing terms, it defines semantics and evaluates the meaning of words. Vector semantics' central tenet is that two words are similar if they have been employed in comparable contexts. Within a multidimensional vector space, words are divided by vector semantics. Sentiment analysis benefits from vector semantics.

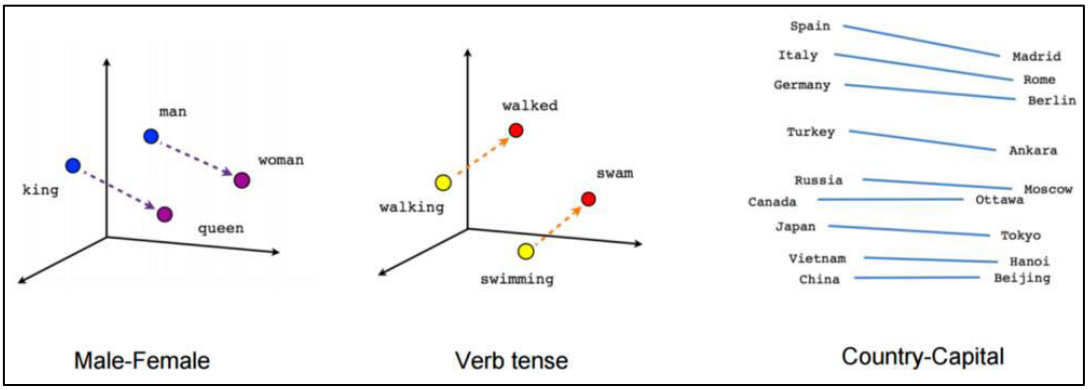

Figure 3: Vector Semantic Approach for text Classification[19]

Word Embedding

Word embeddings are a type of word representation that allows words to be represented as vectors in a continuous vector space. These embeddings capture semantic meanings and relationships between words based on their usage in large text corpora. They are fundamental for word and sequence analysis in various natural language processing (NLP) tasks.

1. **Word2Vec** [20]:
    - **Methodology**: Word2Vec uses two main architectures for learning word embeddings: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicts a target word from its context words, while Skip-gram predicts context words from a target word.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306147|

2. **GloVe (Global Vectors for Word Representation)** [3]:
    - **Methodology**: GloVe is based on the global co-occurrence matrix of words. It captures the global statistical information by factorizing the matrix.
3. **FastText** [21]:
    - **Methodology**: Extends Word2Vec by representing words as bags of character n-grams. This helps capture sub-word information and morphology.
    - **Advantages**: Handles out-of-vocabulary words better by composing embeddings from n-grams.
    - **Limitations**: Still context-independent.

Contextualized Word Embeddings

1. **ELMo (Embeddings from Language Models)** [22]:
    - **Methodology**: ELMo generates word embeddings from a bidirectional LSTM trained on a language model. It captures context-dependent meanings of words by considering the entire sentence.
2. **BERT (Bidirectional Encoder Representations from Transformers)** [15]:
    - **Methodology**: BERT uses a transformer architecture to pre-train a deep bidirectional language model. It considers both left and right context simultaneously.

Word embeddings, both static and contextual, have significantly advanced the field of NLP by providing effective representations of text data. These embeddings are crucial for various word and sequence analysis tasks, enabling models to understand and process natural language more effectively. As research progresses, we can expect further improvements in embedding techniques and their applications in diverse NLP tasks.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306147|

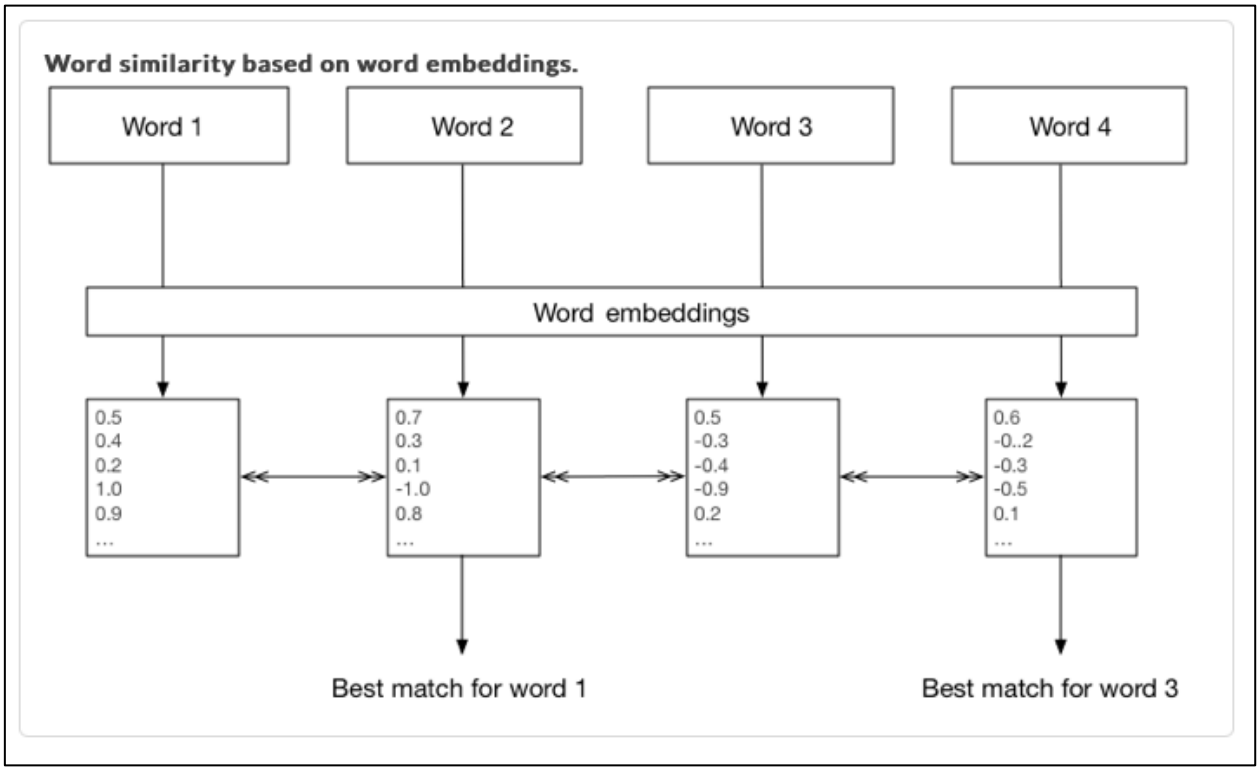

Figure 4: Embedding based word Similarity [19]

Sequence Labeling

Sequence labeling is a crucial technique in natural language processing (NLP) used for tasks where each element in a sequence (such as words in a sentence) needs to be assigned a label. This technique is commonly applied in text classification tasks that involve understanding the context and structure of the input text, such as named entity recognition (NER), part-of-speech (POS) tagging, and chunking. Here's a detailed look at sequence labeling for text classification, including methodologies, applications, and key references.

Methodologies

1. **Traditional Machine Learning Approaches**:
    - **Hidden Markov Models (HMMs)**: Uses statistical properties of sequences for labeling. HMMs are particularly effective for POS tagging.
    - **Conditional Random Fields (CRFs)**: A probabilistic model used for structured prediction. CRFs consider the entire sequence context for predicting labels, making them suitable for NER and similar tasks.

2. **Neural Network Approaches**:
    - **Recurrent Neural Networks (RNNs)**: These models, including their variants Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), are designed to handle sequential data. They can maintain information over time, making them useful for sequence labeling tasks.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306147|

- **Bidirectional RNNs (Bi-RNNs)**: Enhance RNNs by processing the sequence in both forward and backward directions, capturing context from both ends.
- **Attention Mechanisms**: Applied to RNNs or Transformer models to focus on relevant parts of the sequence, improving performance in tasks like NER.
- **Transformers**: Transformer models, such as BERT (Bidirectional Encoder Representations from Transformers), are highly effective for sequence labeling due to their self-attention mechanisms that capture dependencies across the entire sequence.
- **Task**: Grouping words into chunks that represent syntactic units, like noun phrases or verb phrases.

Sequence labeling is essential for many NLP tasks, particularly those requiring detailed text classification. Techniques like HMMs, CRFs, RNNs, and transformers (e.g., BERT) have advanced the field, enabling accurate and efficient labeling of text sequences. These methods help in understanding the structure and meaning of text, contributing to various applications such as NER, POS tagging, and chunking. As research progresses, we can expect further improvements and innovations in sequence labeling techniques.

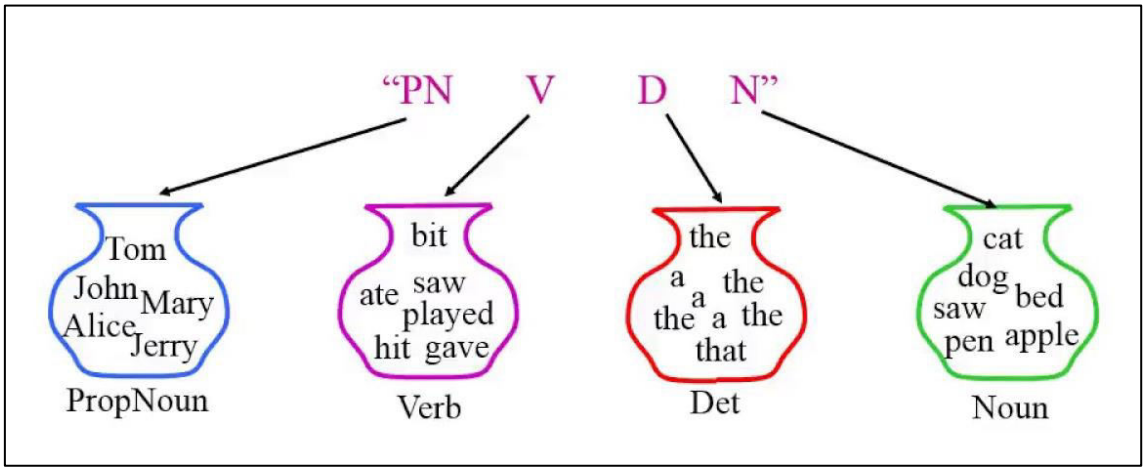

Figure 5: Sequence Labelling[19]

VI. CONCLUSION

Each text classification method has its unique advantages and trade-offs, making them suitable for different scenarios. Rule-based methods are valuable for small-scale, domain-specific applications where interpretability is crucial. Machine learning-based methods and sequence labeling models are ideal for tasks requiring high accuracy and the ability to generalize from large datasets. Word embeddings and vector semantics provide robust and scalable representations that enhance the performance of various NLP applications. Understanding these strengths and limitations is essential for selecting the appropriate method for a given text classification task, ensuring the optimal balance between performance, interpretability, and resource requirements.

REFERENCES

[1] R. Kiros *et al.*, "Skip-thought vectors," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.
[2] R. S. Shailja Sharma, Princy, Kirti Bhatia, "A Study on Image Categorization Techniques," *Int. J. Multidiscip. Res. Sci. Eng. Technol.*, vol. 6, no. 5, pp. 1147–1152, 2023.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306147|

[3] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[4] Q. Wang *et al.*, "LCM-Captioner: A lightweight text-based image captioning method with collaborative mechanism between vision and text," *Neural Networks*, vol. 162, pp. 318–329, 2023.

[5] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[6] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[7] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[8] M. A. Resnik, P., & Hearst, "Syntactic clues to semantics: Learning semantic rules from syntactic patterns," *Proc. ACL.*, 1993.

[9] C. D. Manning, *Introduction to information retrieval*. Syngress Publishing, 2008.

[10] R. J. Mooney, "Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning," *arXiv Prepr. C.*, 1996.

[11] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *COLING 1992 volume 2: The 14th international conference on computational linguistics*, 1992.

[12] R. Smith, "An overview of the Tesseract OCR engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, 2007, vol. 2, pp. 629–633.

[13] A. Beheshti, S. Ghodratnama, M. Elahi, and H. Farhood, *Social data analytics*. CRC press, 2022.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Adv. Neural Inf. Process. Syst.*, vol. 26, 2013.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv Prepr. arXiv1810.04805*, 2018.

[16] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*, 1998, pp. 137–142.

[17] Y. Kim, "Convolutional neural networks for sentence classification. arXiv [J]," *preprint*, 2014.

[18] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *J. Artif. Intell. Res.*, vol. 61, pp. 65–170, 2018.

[19] P. Neupane, "Understanding Text Classification in NLP." https://www.analyticsvidhya.com/blog/2020/12/understanding-text-classification-in-nlp-with-movie-review-example-example/

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv Prepr. arXiv1301.3781*, 2013.

[21] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017.

[22] M. E. Peters *et al.*, "Knowledge enhanced contextual word representations," *arXiv Prepr. arXiv1909.04164*, 2019.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details