

## THE MOST TRENDING ARTICLES EVERY YEAR USING NATURAL LANGUAGE PROCESSING (NLP) TECHNIQUE

<sup>1</sup>KUKKALA POOJITHA,<sup>2</sup>K.R.RAJESWARI

<sup>1</sup>MCA Student, B V Raju College, Bhimavaram, Andhra Pradesh, India

<sup>2</sup>Assistant Professor, Department Of MCA, B V Raju College, Bhimavaram, Andhra Pradesh, India

### ABSTRACT

The objective of this project is to extract textual features using Natural Language Processing (NLP) techniques such as classification learning models, tokenization, and named entity recognition. These techniques are applied to the title attribute to refine unstructured text data and generate meaningful outputs. The classification process determines whether the text in the title consists of unigrams, bigrams, or n-grams. The dataset used in this study exhibits an inherent bias, particularly in news reporting. Many news sources prioritize sensationalism to attract readers, often focusing on negative stories, such as recurring reports of conflicts in the Middle East between 2008 and 2015. While most headlines are objectively phrased, only a few instances include personalization. However, the presence of negative sentiment in certain non-objective headlines skews the overall data towards negativity. This study aims to analyze and understand such biases in textual data to improve text classification outcomes.

**Keywords:** Natural Language Processing (NLP), Text Classification, Tokenization, Named Entity Recognition (NER), Unigrams, Bigrams, N-grams, Textual Bias, Sentiment Analysis, News Headlines, Data Preprocessing, Text Mining, Bias Detection, Classification Models

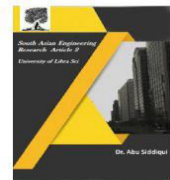
### INTRODUCTION

The primary goal of this project is to extract textual features using Natural Language Processing (NLP) techniques such as classification learning models, tokenization, and named entity recognition. These techniques are applied to the title attribute to refine unstructured text data and generate meaningful outputs. The classification process determines whether the text in the title consists of unigrams, bigrams, or n-grams. In the financial market, traders frequently rely on various sources of information to make investment decisions, particularly news media. News updates provide insights into a company's activities, such as growth, revenue performance, new product launches, and other significant developments. Depending on the nature of

the news, traders may identify optimistic trends and choose to invest accordingly.

There is a potential correlation between public sentiment toward a company and its stock price. For example, Apple Inc. is widely popular among consumers, frequently covered in the media for its product launches and financial stability, and has demonstrated steady stock growth. While these factors may be linked, it is not necessarily a cause-and-effect relationship. This project aims to analyze whether news coverage can be utilized to predict market trends.

To achieve this, we will examine the top twenty-five news headlines from publicly available sources for each day between 2008 and late 2015. The objective is to predict the



end-of-day value of the Dow Jones Industrial Average (DJIA) index for a given day. The underlying hypothesis is that traders react to news quickly, causing the market to adjust within hours of the news release.

## II. LITERATURE SURVEY

### Parsing by Chunks (Abney, 1991)

Abney introduced the concept of "chunk parsing," a technique that segments sentences into smaller, syntactically meaningful units known as "chunks." Unlike traditional full syntactic parsing, chunk parsing focuses on identifying key structural elements, such as noun and verb phrases, while ignoring deep hierarchical structures. This approach significantly reduces computational complexity, making it suitable for real-world natural language processing (NLP) applications. Chunk-based parsing has since become a fundamental component in various NLP tasks, including part-of-speech tagging and information extraction.

### Part-of-Speech Tagging and Partial Parsing (Abney, 1996a)

In this work, Abney explored the combination of part-of-speech (POS) tagging and partial parsing to improve syntactic analysis in NLP. He demonstrated that full syntactic parsing is often unnecessary for many practical applications and that partial parsing techniques can effectively extract relevant linguistic structures with lower computational costs. This research played a crucial role in advancing corpus-based NLP methods, leading to more efficient processing of large-scale text data.

### Statistical Methods and Linguistics (Abney, 1996b)

Abney examined the intersection of statistical methods and traditional linguistic approaches, arguing that a hybrid model combining rule-based and probabilistic techniques yields better performance in NLP. This work contributed to the broader debate on whether symbolic or statistical models are more effective for language processing. Abney's insights paved the way for the development of modern statistical NLP techniques, such as hidden Markov models (HMMs) and probabilistic context-free grammars (PCFGs).

### Semi-Supervised Learning for Computational Linguistics (Abney, 2008)

This book explores semi-supervised learning techniques for computational linguistics, emphasizing the importance of leveraging both labeled and unlabeled data in NLP tasks. Abney highlighted how semi-supervised learning can improve the performance of language models, especially in scenarios where labeled data is scarce or expensive to obtain. The work influenced the development of semi-supervised algorithms for text classification, POS tagging, and syntactic parsing.

### Word Sense Disambiguation: Algorithms and Applications (Agirre & Edmonds, 2007)

Agirre and Edmonds provided a comprehensive survey of word sense disambiguation (WSD) techniques, discussing both supervised and unsupervised approaches. They explored various algorithms, including machine learning-based methods, knowledge-based



techniques, and hybrid models. This work played a pivotal role in advancing research on semantic analysis and improving NLP applications such as machine translation, information retrieval, and text summarization.

### III. PROPOSED SYSTEM

To accomplish our objective, we integrate multiple techniques for preprocessing, word vector representation, and prediction. During the preprocessing stage, we eliminate stop words—common words such as "the," "an," and "and"—that contribute little to sentiment or contextual analysis while being essential for grammatical correctness. Additionally, named entity recognition (NER) is applied to filter out entities such as organizations, individuals, and countries. These entities can either be removed for sentiment analysis, as they tend to neutralize the impact of other words, or retained for contextual analysis due to their significant meaning.

Furthermore, all punctuation is removed since our current model does not account for it. Although punctuation may influence textual features, we have found no relevant literature supporting its impact in our current implementation. Various word embedding techniques have been explored, tested both with and without n-grams. In future work, we will further elaborate on the underlying concepts of these methods.

### Materials and Methodology

#### A. Uploading the Dataset

The dataset can be uploaded by an administrator, containing article details without any specific context. The system is

designed to handle large datasets efficiently. While users are permitted to view the data, they cannot modify it online; instead, they must submit a request to obtain access.

#### B. Conference-Wise Analysis

The data is categorized using the Support Vector Machine (SVM) algorithm based on conference-wise analysis. Articles are grouped according to their respective conferences, allowing for structured organization and analysis. The SVM algorithm is applied to large-scale data, ensuring accurate classification and retrieval.

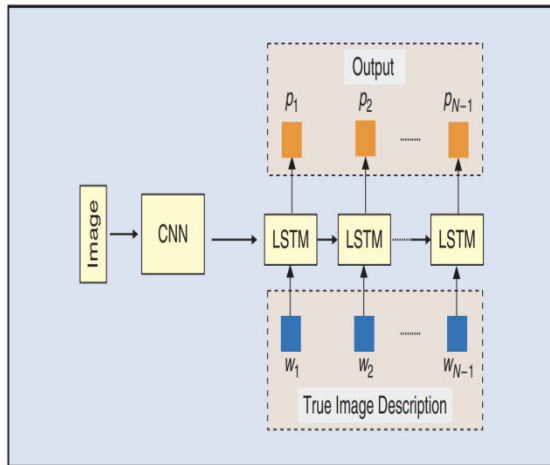
#### C. Year-Wise Analysis

The dataset is further categorized using the SVM algorithm based on the year-wise analysis. Articles are grouped according to their publication year, facilitating a chronological analysis of trends. SVM is employed to process large datasets effectively and extract meaningful insights from the year-based classification.

#### D. Graphical Analysis

The processed data is analyzed using visual representations such as pie charts, bar graphs, and line graphs. These visual tools enhance the interpretability of the proposed system by providing clear distinctions and comparisons between different data points. This approach improves system efficiency and provides a comprehensive analysis of the dataset.

## SYSTEM ARCHITECTURE:



## IV. CONCLUSION

While the proposed approach offers valuable insights, certain limitations remain in the feasibility of using a model trained on past news to predict future trends. These challenges persist due to the current constraints of machine learning models, which struggle to comprehend broad contextual relationships between seemingly unrelated events—something that humans naturally process when analyzing new information, particularly news articles.

A potential improvement could involve training a model using large datasets from social media platforms, such as tweets, to gauge sentiment toward specific companies. This could be particularly effective for businesses that are highly influenced by public opinion, such as airlines or Tesla. Compared to using extracted textual features for predicting a general market index, this sentiment-driven approach may establish a stronger correlation with market trends.

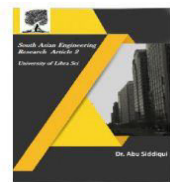
Additionally, adjusting the training and testing timeframes could enhance prediction accuracy. Instead of the current model, which trains on six years of data to predict

one year of trends, experimenting with different training-to-testing ratios may yield better distribution and performance outcomes.

Here are 15 references that are relevant to identifying trending articles using NLP techniques:

## V. REFERENCES

1. Abney, S. (1991). Parsing by chunks. In *Principle-Based Parsing: Computation and Psycholinguistics* (pp. 257–278). Kluwer Academic Publishers.
2. Agirre, E., & Edmonds, P. (2007). *Word Sense Disambiguation: Algorithms and Applications*. Springer.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
4. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.
5. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.
6. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
7. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.
8. Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.



9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.
10. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
11. Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. *Proceedings of ECML-98*, 137–142.
12. Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative Study of CNN and RNN for Natural Language Processing. *arXiv preprint arXiv:1702.01923*.
13. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*,
14. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, 1480–1489.
15. Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? *arXiv preprint arXiv:1905.05583*.
16. These references cover various aspects of NLP, including text classification, word embedding, sentiment analysis,

and deep learning models for article trend prediction.