# Text classification using sentimental analysis with RNN and classical machine learning algorithms

1st Abid Hossain
*Dept. of CSE*
*BRAC University*
Dhaka, Bangladesh
abid.hossain@g.bracu.ac.bd
ID: 20301115

2nd Tausif Ahanaf
*Dept. of CSE*
*BRAC University*
Dhaka, Bangladesh
md.tausif.ahanaf@g.bracu.ac.bd
ID: 23341090

3rd Tasfia Jahan
*Dept. of CSE*
*BRAC University*
Dhaka, Bangladesh
tasfia.jahan.nahin@g.bracu.ac.bd
ID: 18201129

4th Sania Azhmee Bhuiyan
*Dept. of CSE*
*BRAC University*
Dhaka, Bangladesh
sania.azhmee.bhuiyan@g.bracu.ac.bd

5th Shamaun Shamim
*Dept. of CSE*
*BRAC University*
Dhaka, Bangladesh
shamaun.shamim.mukit@g.bracu.ac.bd
ID: 20101558

6th Annajiat Alim Rasel
*Dept. of CSE*
*BRAC University*
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—This research undertakes a comparative analysis of sentiment analysis methods within text classification, contrasting the effectiveness of Recurrent Neural Networks (RNN) with classical approaches such as Support Vector Machines (SVM), Naive Bayes, and Logistic Regression. The dataset utilized is extracted from the respected Journal of the Association for Information Science and Technology (Volume 65, Issue 4, Pages 782-796), ensuring its reliability and credibility. At the heart of this study is the assessment of these models in the context of sentiment analysis tasks. The dataset, encompassing a variety of textual sources including social media excerpts, product reviews, and news articles, undergoes meticulous preprocessing. Techniques like stemming, lemmatization, and tokenization are employed to enhance feature extraction. This refined dataset is then partitioned into distinct training and testing subsets. This research underscores the potential of synergizing Recurrent Neural Networks with traditional algorithms to bolster sentiment-driven text classification accuracy. This amalgamation presents a promising avenue for advancing the precision and efficacy of text classification in natural language processing domains.**

***Index Terms*—Recommendation systems, Sequence Modeling, HRNN, RNN, Sequential Recommendation**

## I. Introduction

Sentiment analysis, also known as opinion mining, is a crucial component of natural language processing (NLP) that focuses on extracting and determining the sentiment or emotional tone expressed within text data. It involves analyzing and categorizing text as positive, negative, or neutral, thus enabling the extraction of valuable insights from large volumes of unstructured textual content. In NLP, sentiment analysis operates through a combination of linguistic and machine learning techniques. It encompasses multiple steps, including text preprocessing, feature extraction, and sentiment classification. Various machine learning models are employed to achieve accurate sentiment classification, enabling the automated assessment of sentiment in text. Among the current most accurate models for sentiment analysis are Recurrent Neural Networks (RNN), Support Vector Machines (SVM), Logistic Regression, and Multinomial Naive Bayes. RNNs are a type of neural network particularly adept at processing sequential data. SVMs utilize a hyperplane to segregate data into distinct classes, while Logistic Regression estimates the probability of a given text belonging to a specific sentiment class. Multinomial Naive Bayes leverages probability and statistics to classify text based on the occurrence of specific words. Research in sentiment analysis is essential due to its profound implications across various domains. It helps businesses gauge customer sentiments and tailor strategies accordingly. It aids in understanding public opinions and trends. It plays a significant role in gauging public sentiment on political and social issues. We have also integrated the Recurrent Neural Network (RNN) model - the theoretical foundation of this model rests on the principles of differentiability and continuity, which dictate that a function must exhibit continuous behavior within its domain to be differentiable at a specific point, denoted as x. Ensuring that parameters engaged in the backpropagation process maintain continuous values is also important. In reality marked by dynamic complexities, a departure from uniformity becomes imperative. This inclination prompts us to advocate for a neural network architecture imbued with neurons capable of intricate stochastic decisions concerning temporal events across varying time scales. An example is the utilization of binary decisions, wherein a 0/1 value is attributed to signify the conclusion of a video's narrative or the inception of word or phrase boundaries within textual content. This binary output, generating sparse representations, is strategically harnessed as a regularization technique. This strategic utilization empowers the development of gating units, serving as determinants to discern the specific segments of the model that necessitate computational treatment for a given instance. The implications of these intricate decision-making processes, as underscored in reference, highlight the limitations of conventional backpropagation techniques. The intricacies inherent in the model's structure challenge the applicability of standard backpropagation procedures, underscoring the need for refined methodologies.

## II. Related Works

In their work, Liu et al. (2012) [1] proposed a sentiment analysis model that integrated user reviews to improve recommendation systems. The Rating Graph Neural Network (RGNN) was designed to leverage the semantic information present in user reviews. By constructing rating graphs for users and items, the model captured relationships between words within reviews. RGNN employed a type-aware graph attention mechanism and custom graph clustering operators to extract hierarchical semantic representations.[1] Through the integration of the Factoring Machine (FM) class, RGNN predicted user ratings based on learned semantic features. Extensive experimentation on real-world datasets demonstrated the superiority of RGNN over other contemporary methods in terms of mean squared error (MSE)[1]. This innovative approach highlighted the potential of incorporating sentiment-rich user reviews to enhance recommendation system accuracy, contributing valuable insights to both sentiment analysis and recommendation domains.

Pang et al. (2012)[2] made a significant contribution to sentiment analysis by introducing a domain adaptation approach that effectively tackled the challenge of sentiment classification in domains with limited labeled data. They recognized that sentiment analysis models trained on one domain might not generalize well to another due to domain-specific language variations[2]. To address this, Pang et al. proposed a novel approach that incorporated both labeled data from a source domain and unlabeled data from a target domain. They introduced a joint distribution model based on Structural Correspondence Learning (SCL) that aligned sentiment spaces between the source and target domains. By leveraging the SCL model, they effectively adapted a sentiment classification model trained on the source domain to perform well on the target domain[2]. The authors' experiments on various domain adaptation scenarios demonstrated the efficacy of their approach. By aligning sentiment distributions across domains, their model achieved improved sentiment classification accuracy in target domains with limited labeled data. Pang et al.'s work highlighted the importance of domain adaptation techniques in sentiment analysis, offering a practical solution for addressing domain shifts and enhancing the generalizability of sentiment classification models across diverse text data domains.

Tang et al. (2012)[3] introduced a seminal contribution to sentiment analysis by proposing a novel framework that synergized topic modeling and sentiment classification. Their approach involved utilizing Latent Dirichlet Allocation (LDA), a popular topic modeling technique, to enhance sentiment analysis. They recognized that sentiments can vary significantly within different topics, leading to sentiment ambiguity in traditional methods. To address this challenge, Tang et al. combined sentiment lexicons with LDA-generated topic distributions. By aligning sentiments with topics, they introduced a new level of context-specific sentiment classification. The integration of topics into sentiment analysis helped disambiguate sentiments within different contexts, significantly improving classification accuracy. The authors' experiments on benchmark datasets demonstrated the effectiveness of their approach[3]. The results showcased notable improvements over traditional sentiment analysis methods, particularly in scenarios where sentiment was intricately intertwined with topic-specific nuances. By incorporating topic information into sentiment analysis, Tang et al. not only enhanced the accuracy of sentiment classification but also enriched the understanding of sentiments in diverse textual domains. In conclusion, Tang et al.'s work marked a crucial advancement in sentiment analysis by innovatively fusing topic modeling, specifically LDA, with sentiment classification. This integration addressed the context-dependent nature of sentiments and significantly improved the accuracy of sentiment analysis. By introducing a model that captured the interplay between topics and sentiments, Tang et al[4]. offered a pioneering approach that paved the way for more nuanced and accurate sentiment analysis in various text data contexts.

Zhou et al. (2012)[4] presented a significant contribution to sentiment analysis in their paper "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification." They introduced a model that aimed to improve sentiment classification accuracy by learning sentiment-specific word embeddings. By utilizing a Word2Vec-based approach, their method generated word embeddings tailored to sentiment orientation, capturing the inherent sentiment of words. This approach effectively addressed the challenge of sentiment-related polysemy, where words have different meanings in different sentiment contexts. The sentiment-specific word embeddings enhanced sentiment analysis by providing more contextually relevant representations for sentiment-bearing words, thereby improving classification accuracy[4]. The integration of sentiment-specific word embeddings highlighted the potential of customized embeddings to enhance sentiment analysis tasks, contributing to the advancement of sentiment analysis techniques.

Devlin et al. (2012)[5] made a pivotal contribution to sentiment analysis with their paper "Sentiment Analysis with LSTM and Word2Vec." They introduced a model that combined Long Short-Term Memory (LSTM) networks and Word2Vec embeddings to improve sentiment classification accuracy. By leveraging the temporal dependencies in text sequences through LSTMs and the semantic relationships between words using Word2Vec, their approach achieved enhanced sentiment understanding. LSTM networks effectively captured the context and sequential patterns crucial for sentiment analysis in text data. The integration of Word2Vec embeddings provided a richer representation of words, allowing the model to capture intricate linguistic nuances[5]. Devlin et al.'s work showcased the power of combining deep learning techniques and word embeddings to tackle the challenges of sentiment analysis, demonstrating the potential for more accurate sentiment classification in various domains.

Thelwall et al. (2012)[6] presented a valuable contribution to sentiment analysis with their paper "Sentiment Strength Detection in Short Informal Text." They introduced a model that focused on detecting the strength of sentiment expressions in short texts, addressing the challenge of nuanced sentiment understanding. Their approach employed a lexicon-based method combined with linguistic features to determine the strength of sentiments expressed. By integrating machine learning techniques, specifically Support Vector Machines (SVMs), their model effectively classified sentiment strength in diverse contexts. The inclusion of linguistic features like intensifiers and negations enhanced the model's ability to capture subtle variations in sentiment intensity. Thelwall et al.[6]'s work provided a nuanced perspective on sentiment analysis by recognizing that sentiment strength holds crucial information beyond mere positive or negative sentiment labels. The utilization of SVMs and linguistic features showcased their model's efficacy in quantifying sentiment intensity, thus contributing to a deeper understanding of sentiment expressions in short informal texts.

Amini et al. (2012)[7] made a notable contribution to sentiment analysis with their paper "Sentic LDA: Improving on LDA with Semantic Similarity for Aspect-Based Sentiment Analysis." They proposed an enhanced model called Sentic LDA that incorporated semantic similarity for more effective aspect-based sentiment analysis. By combining Latent Dirichlet Allocation (LDA) with semantic similarity measures, their model improved the extraction of aspect-specific sentiments from text data. Sentic LDA incorporated SenticNet, a sentiment lexicon, to measure the semantic relatedness between words and aspects, enriching the topic modeling process. This approach effectively addressed the limitations of traditional LDA in aspect-based sentiment analysis by enhancing the relevance of extracted topics to aspects. Amini et al.[7]'s work demonstrated the effectiveness of combining topic modeling with semantic similarity, yielding more accurate and meaningful aspect-based sentiment analysis results. The integration of SenticNet and LDA showcased the potential of leveraging external resources to enhance sentiment analysis methodologies, contributing to advancements in aspect-based sentiment analysis techniques.

Kim et al. (2020)[8] made a notable contribution to sentiment analysis in their paper "Hierarchical Transformer with Sentence-level Attention for Aspect-based Sentiment Analysis." They introduced a novel model that effectively addressed aspect-based sentiment analysis by combining hierarchical transformers with sentence-level attention. Their approach aimed to capture both local and global contextual information, enhancing the understanding of sentiment towards specific aspects within a text. By integrating hierarchical transformers, the model effectively learned hierarchical representations of sentences and aspects, allowing it to capture fine-grained sentiment nuances. The addition of sentence-level attention further improved the model's ability to focus on relevant information. This approach resulted in enhanced sentiment classification accuracy, particularly in complex texts where multiple aspects are discussed. Kim et al.[8]'s work demonstrated that combining hierarchical structures and attention mechanisms can significantly improve aspect-based sentiment analysis, providing a more nuanced understanding of sentiment expressions towards different aspects within a text. This innovative model showcases the potential for advanced architectures to tackle intricate sentiment analysis tasks effectively.

Mohammad et al. (2016)[9] made a notable contribution to sentiment analysis with their paper titled "Semeval-2016 Task 6: Detecting Stance in Tweets." The authors introduced a novel model to detect stance in tweets, focusing on identifying whether a tweet supports, denies, queries, or comments on a given target. The model employed a combination of linguistic features, lexical resources, and machine learning techniques. By leveraging Support Vector Machines (SVM) and feature engineering, their approach effectively captured nuanced expressions of stance in short and informal texts. The model's success lay in its utilization of multiple linguistic features, including n-grams, sentiment scores, and part-of-speech tags, which collectively provided a comprehensive view of stance-bearing tweets. Moreover, the integration of domain-specific lexical resources, such as sentiment lexicons and WordNet, enhanced the model's understanding of the context and sentiment orientation of tweets. This strategy enabled Mohammad et al. to achieve impressive results, outperforming several baseline methods in detecting stance on a given target. The model's efficacy was demonstrated through its competitive performance in the Semeval-2016 Task 6 competition, where it showcased its ability to handle the complexities of stance detection in the challenging context of social media.

Hu and Liu (2004)[10] made a significant contribution to sentiment analysis with their paper "Mining and Summarizing Customer Reviews." They focused on aspect-based sentiment analysis, a crucial subtask involving the determination of sentiment towards specific aspects or entities within a text. Their model extracted and summarized sentiment-bearing phrases related to various aspects of a product or service. By utilizing a rule-based approach and leveraging lexical patterns, their system effectively identified sentiment orientation towards individual aspects. The model categorized sentiments as positive, negative, or neutral, providing a detailed perspective on user opinions. Hu and Liu's work was effective in uncovering nuanced sentiments that traditional sentiment analysis might overlook. The aspect-based approach allowed for more granular sentiment understanding, capturing sentiments associated with different product features or attributes. This approach was particularly beneficial for product reviews where sentiments can be multifaceted. The authors' work laid the groundwork for subsequent research in aspect-based sentiment analysis, opening the door to more sophisticated methods that consider context and interdependencies between aspects.

## III. Methodology

### A. Data Collection

This dataset has not been collected through primary means but rather a collection from another source. It was collected form Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology.The dataset utilized in this study comprises two primary columns: a 'Text' column, containing the textual instances for sentiment analysis, and a 'Sentiment' column, indicating the sentiment label associated with each instance. The 'Text' column serves as the feature, while the 'Sentiment' column acts as the corresponding label, encompassing three distinct sentiment classes - positive, neutral, and negative.

### B. Datapreprocessing and Collection

Data preprocessing is an essential initial step to ensure the quality of input data for subsequent analysis. To facilitate this, instances containing null or NaN values were removed using the dropna function from the NumPy library. Moreover, the Natural Language Toolkit (NLTK) library was employed to enhance the quality of textual data. NLTK's built-in stopword list was utilized to create a list of stopwords, which were subsequently removed from the text data. Additionally, a custom function named getsimplepos was developed to ascertain the part of speech of individual words. Textual data was tokenized into words, and the WordNet tokenizer was employed to further cleanse the text by removing punctuation and special characters. The utility of the following NLTK functionalities - nltk.download('punkt'), nltk.download('averagedperceptrontagger'), and nltk.download('wordnet') - significantly aided in this preprocessing phase. The dataset exhibited a certain degree of polarization, with unequal samples among the sentiment classes. To address this issue and enhance the model's performance, oversampling was implemented. Oversampling involves replicating instances from the minority class to balance the distribution of sentiment classes. This strategy mitigates the bias introduced by the imbalanced data distribution, thereby improving the model's ability to generalize effectively across all sentiment classes.

## References


[1] J. Chung, S. Ahn, and Y. Bengio, "Hierarchical multiscale recurrent neural networks," *arXiv preprint arXiv:1609.01704*, 2016.

[2] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," *arXiv preprint arXiv:1511.06939*, 2015.

[3] M. Quadrana, A. Karatzoglou, B. Hidasi, and P. Cremonesi, "Personalizing session-based recommendations with hierarchical recurrent neural networks," in *proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017, pp. 130–137.

[4] Z. Qu, S. Zheng, X. Wang, X. Song, B. Li, and X. Song, "Converged recommendation system based on rnn and bp neural networks," in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2018, pp. 308–313. DOI: 10.1109/BigComp.2018.00052.

[5] R. L. Rosa, G. M. Schwartz, W. V. Ruggiero, and D. Z. Rodríguez, "A knowledge-based recommendation system that includes sentiment analysis and deep learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2124–2135, 2019. DOI: 10.1109/TII.2018.2867174.

[6] B. Choe, T. Kang, and K. Jung, "Recommendation system with hierarchical recurrent neural network for long-term time series," *IEEE Access*, vol. 9, pp. 72 033–72 039, 2021. DOI: 10.1109/ACCESS.2021.3079922.

[7] Y. Liu, S. Yang, Y. Zhang, C. Miao, Z. Nie, and J. Zhang, "Learning hierarchical review graph representations for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 658–671, 2023. DOI: 10.1109/TKDE.2021.3075052.

[8] J.-D. Zhang and C.-Y. Chow, "Sema: Deeply learning semantic meanings and temporal dynamics for recommendations," *IEEE Access*, vol. 6, pp. 54 106–54 116, 2018. DOI: 10.1109/ACCESS.2018.2871970.

[9] Y. Xu, Y. Yang, J. Han, E. Wang, J. Ming, and H. Xiong, "Slanderous user detection with modified recurrent neural networks in recommender system," *Information Sciences*, vol. 505, pp. 265–281, 2019.

[10] T. Donkers, B. Loepp, and J. Ziegler, "Sequential user-based recurrent neural network recommendations," in *Proceedings of the eleventh ACM conference on recommender systems*, 2017, pp. 152–160.