Sentiment Analysis of Online News Using Deep Learning Techniques

Mr. Ankur Jain, Research Scholar,Bhagwan Mahavir University,Surat, Gujarat

Dr. Vineet Kumar Goel, Dean, Faculty of Engineering, Bhagwan Mahavir University, Surat, Gujarat

Abstract

The modern technological era has significantly transformed traditional lifestyles across various domains. News and events are now disseminated more quickly and effectively, driven by advancements in Information Technology (IT). IT has also led to a massive surge of data, generated every minute by millions of users through comments, blogs, news articles, social media posts, and microblogs.Analyzing such massive volumes of data manually is highlychallenging, necessitating the adoption of advanced methods to perform this task automaticallyand effectively. Events that evoke emotions, whether happy, bad, or neutral, are often depicted in news articles. Sentiment analysis is a useful technique for examining and comprehending the feelings included in written material. Thelexicon-based experiments carried out with the BBC News dataset confirm the viability and effectiveness of the recommended approach.

Keywords-Opinion mining, RNN,LSTM,GRU

I. INTRODUCTION

The way individuals obtain news has changed as a result of the development of the Internet and web and mobile technology. Digital media such as blogs and online news sites have mostly replaced traditional print newspapers and magazines. Two primary factors driving this shift are interactivity and immediacy [1].

In the fast-paced world of today, people try to get as much news as they can from a variety of sources, concentrating on subjects that interest them or are important to them. Interactivity reflects the audience's preference for personalized news consumption aligned with their interests. Immediacy, on the other hand, highlights the demand for real-time access to news without delays [2].

The modern world and its advanced technology enable people to enjoy features like instant access to real-time news as events unfold. Online news platforms have implemented effective strategies to capture audience attention [3]. These platforms often express opinions about various news entities—such as individuals, locations, or objects—while reporting recent events [4]. To enhance user engagement, many news websites provide interactive emotion-rating features, allowing readers to classify news as positive, negative, or neutral [5].

Opinion mining is another name for sentiment analysis, identifies the polarity or intensity of opinions (positive, negative, or neutral) expressed in text, such as news articles [3][4]. Manually labeling sentiment words is time-intensive, leading to the adoption of automated methods. Two primary approaches dominate sentiment analysis: Deep learning and machine learning techniques. The rise of deep learningtransformedsentiment analysis by improving modelaccuracy and efficiency. Sequential input was analyzed using long short-term memory (LSTM) networks and recurrent neural networks (RNNs). LSTM networks, in particular, are renowned for their capacity to preserve long-range dependencies in text, proved highly effective in capturing contextual information, making them well-suited for sentiment analysis of extended text sequences.

This document is organized as follows: Section II provides a summary of pertinent research on sentiment analysis of news articles. The suggested methodology and experimental setting are explained in Section III. Section V discusses the research's shortcomings, whereas Section IV describes the findings. Finally, the investigation is concluded in Section VI.

II. Literature Survey

Several researchers have investigated sentiment analysis in news articles using a variety of approaches. A summary of key studies in this domain is outlined below:

Reis, Olmo, Benevenuto, and Prates et al. looked into the relationship between emotion polarity and the popularity of news articles[3]. They examined 69,907 headlines from the New York Times, BBC, Reuters, and Daily Mail, four significant media outlets. Articles with positive or negative tones attracted more attention than those with neutral tones, according to the study, which evaluated the sentiment polarity of headlines by extracting linguistic elements.

To find sentiment terms and related entities in a corpus of news stories and blogs, Godbole, Srinivasaiah, and Sekine created an algorithm that makes use of sentiment lexicons[4]. The algorithm analyzed the co-occurrence of sentiment words and entities within sentences and evaluated sentiment along seven dimensions: general, health, crime, sports, business, politics, and media. Two trends were studied:

(1) Identifying the positive or negative feeling attached to a thing is known as polarity.

(2) Subjectivity, quantifies the strength of sentiment. Both metrics' scores were calculated.

Islam, Ashraf, Abir, and Mottalib presented a method for categorizing internet news articles using sentence-level sentiment analysis. [6]. To ascertain polarity, their method made use of a dynamic lexicon of predetermined positive and negative phrases. The process included selecting an article, extracting sentences of various structures (simple, compound, complex), identifying positive elements, and calculating their polarities. The combined sentence polarities were used to infer the overall sentiment, achieving a 91% classification accuracy.

Meyer, Bikdash, and Dai compared lexicon-based and machine learning approaches in their fine-grained sentiment analysis of financial news headlines. Eight experiments were performed to assess accuracy. The machine learning method used syntactic models with part-of-speech tagging, In the lexicon-based method, the General Inquirer Lexicon (H4N) and the Bag of Words (BoW) paradigm were used. Results showed superior accuracy with machine learning.

A framework for document-level sentiment analysis was put forth by Shirsat, Jagdale, and Deshmukh [9] in order to assess the general polarity of news stories. Preprocessing (tokenization, elimination of stop words, stemming) and postprocessing of text were used in their 2,225-document study to assign sentiment scores and categorize articles as neutral, negative, or positive.

Agarwal, Sharma, Sikka, and Dhir utilized Python and SentiWordNet 3.0 for opinion mining to classify news headlines by sentiment impact [10]. Their method involved two stages: preprocessing (POS tagging, lemmatization, and stemming with NLTK) and sentiment analysis, where SentiWordNet 3.0

computed positive, negative, and objective scores. Headlines were labeled as positive or negative based on score comparisons.

A model was created by Lei, Rao, Li, Quan, and Wenyin to identify social emotions elicited by tweets and news articles[11]. This model included modules for document selection, POS tagging, and lexicon generation. Using a training dataset of 40,897 news articles, the model extracted features and generated emotion lexicons based on probability calculations.

Chen, Kong, et al. (2018) introduced a deep learning-based popularity prediction model that integrated text, user content, and time-series information. An attention mechanism was incorporated to handle noisy data. The model outperformed baseline approaches through the use of time embedding and joint learning embeddings of users and words.

Cai and Zheng (2022) compared the performance of the Gated Recurrent Unit (GRU) model with LSTM, linear regression, and random forest models. The GRU model not only outperformed the others in accuracy but also required less training time.

III. RESEARCH METHODOLOGY

This work uses a deep learning-based method for news article sentiment analysis. Sentiment analysis can be performed using both supervised and unstructured methods.

Supervised techniques-:

- Labeled training data is utilized in supervised techniques to create classification models, which are subsequently applied to unlabeled data.

Unsupervised techniques-:

- No training data is needed for unsupervised methods. Rather, they use word polarity to infer sentiment. The polarity of individual words or phrases are aggregated to define the sentiment of a sentence or document [12].

The models used for sentiment analysis are divided into two catagories:

1. **Machine learning models**: Use Random Forest, Gradient Boost and Adaboost Regressor for analysing labelled data.
2. **Deep Learning models** :Utilize LSTM, GRU, and Recurrent Neural Networks, which are based on the patterns of neurons in the human brain.

Sentiment analysis can be conducted at various levels: document, sentence, word, or phrase. This research focuses on finding whether a news will be popular based on calculating the number of times it is shared using GRU model.GRU makes use of two gates: the update gate and the reset gate. The reset gate discard the inappropriate information and update gate makes balance between keeping past information and introducing new information. As GRU selectively allow information to pass through it so problem of vanishing gradient is also solved and makes network to learn long range dependency.

The methodology for sentiment analysis in this study consisted of five steps, starting with data collection and followed by preprocessing,, sentiment scoring, and classification of results. These steps are detailed below:

A. Data Collection

The experiment utilized the www.mashable.com news dataset, available online athttps://archive.ics.uci.edu/dataset/332/online+news+popularity.A wide variety of statistics about stories that Mashable created over a two-year period are included in this collection. The articles are divided into five topical categories, which match the class designations of technology, entertainment, business, and lifestyle.

B. Data Preprocessing

Preprocessing was conducted to clean and normalize the text, reducing inconsistencies and noise to improve the effectiveness of text mining and sentiment analysis tasks [15]. The preprocessing was performedemploying the following steps:

1. **Null Values Removal**: There are no null values in the dataset selected for this study.
2. **Outlier Removal**: If there is any value which is very high or very low in the features selected for prediction then it will be removed.
3. **Stop Word Removal**: Commonly occurring words with minimal semantic value, such as "and," "the," or "is," were removed using the “Filter stop word (English)” operator.
4. **Stemming**: Inflected and derived word forms were reduced to their base forms using the “Stem (WordNet)” operator.

C. Feature Extraction

After preprocessing,seven features which contribute most to the popularity of news articles are selected and rest of the features are dropped from the input dataset. The number of shares is labelled as the target variable.

D. Model Building

The models which are implemented in this paper are GRU, LSTM.

E. Model Compilation

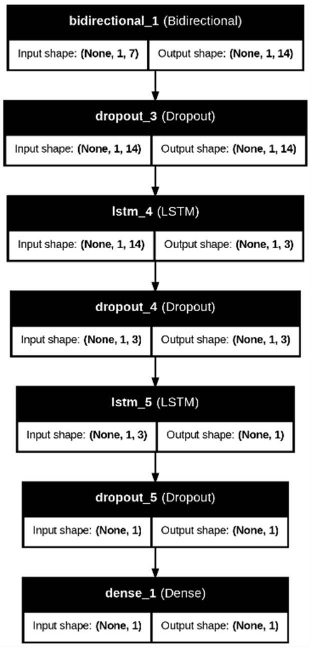
All the three models are compiled and test loss,test mean absolute error is obtained as performance metric.

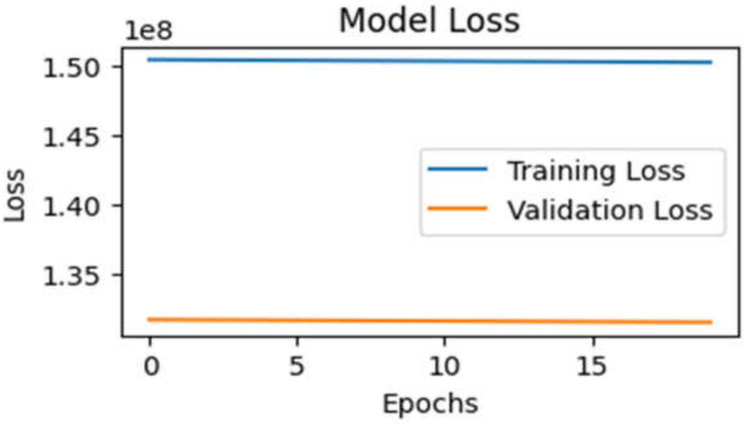
IV. Results and Discussion

Table 1 provides a summary of the experimental findings, showcasing the better performance of GRU model over LSTM and RNN.

Model	Test Loss	Test MAE
GRU	112997496.0	2989.9118
LSTM	131538040.0	3296.4009

LSTM Network Diagram Model Loss Graph


bidirectional_1 (Bidirectional)
Input shape: (None, 1, 7)
Output shape: (None, 1, 14)
dropout_3 (Dropout)
Input shape: (None, 1, 14)
Output shape: (None, 1, 14)
lstm_4 (LSTM)
Input shape: (None, 1, 14)
Output shape: (None, 1, 3)
dropout_4 (Dropout)
Input shape: (None, 1, 3)
Output shape: (None, 1, 3)
lstm_5 (LSTM)
Input shape: (None, 1, 3)
Output shape: (None, 1)
dropout_5 (Dropout)
Input shape: (None, 1)
Output shape: (None, 1)
dense_1 (Dense)
Input shape: (None, 1)
Output shape: (None, 1)



V. Research Limitations and Challenges

Quality and Availability: It might be challenging to extrapolate results from public datasets for news popularity prediction across various topics or geographical areas due to their sometimes small size and scope.

Bias in Data Sources: Editorial policies or target audiences may cause bias in news items, which can skew the results.
Timeliness: Popularity measures, such as likes, shares, and comments, can change dramatically over time, therefore it's important to collect data in real time to ensure accuracy.
Missing or Inconsistent Data: Some metrics, such as click-through rates or view counts, could not be available at all times or might vary depending on the platform.

Complexity of Content Features: Advanced computer vision, multimodal learning, or natural language processing (NLP) techniques are needed to extract meaningful features from text.
Latent Semantic Information: It can be difficult to discern the content's nuanced intent, tone, or sentiment, and it might not always match the reader's views.
Social Context: It might be challenging to measure characteristics like author reputation, the timing of publishing, or outside events that affect an article's prominence.

VI. Conclusion

Sentiment analysis provides a wealth of opportunities for future research. This study used a dataset from news published on www. mashable.com website to analyze the sentiment of news stories. The results showed that while the categories of entertainment and technology were linked to negative feelings, the categories of business and sports contained a greater proportion of positive content.

Future research will explore the application of Artificial Intelligence based approaches to sentiment analysis, with plans to develop an online platform where users can access and customize their news feeds. Leveraging sentiment analysis, users could filter news articles according to their preferences, enhancing the personalization and relevance of their reading experience.

REFERENCES

[1] M. Karlsson, The immediacy of online news, the visibility of journalistic
processes and a restructuring of journalistic authority. Journalism, 12(3), 279-295, 2011.

[2] A. Kohut, C. Doherty, M. Dimock and S. Keeter, Americans spending more time following the news. Pew Research Center, 2010.

[3] J. Reis, P. Olmo, F. Benevenuto, H. Kwak, R. Prates, and J. An, Breaking the news: first impressions matter on online news. In ICWSM ’15, 2015.

[4] N. Godbole, M. Srinivasaiah, and S. Sekine, Large-scale sentiment analysis for news and blogs. In International Conference on Weblogs and Social Media, Denver, CO, 2007.

[5] J. Lei, Y. Rao, Q. Li, X. Quan, and L. Wenyin, “Towards building a social emotion detection system for online news,” Future Generation Computer Systems, vol. 37, pp. 438–448, 2014.

[6] M. U. Islam, F. B. Ashraf, A. I. Abir and M. A. Mottalib, "Polarity detection of online news articles based on sentence structure and dynamic dictionary," 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, 2017, pp. 1-5.
doi: 10.1109/ICCITECHN.2017.8281777

[7] V. Kharde, and P. Sonawane, “Sentiment analysis of twitter data: a survey of techniques,” International Journal of Computer Applications, vol. 139, no. 11, pp. 5-15, 2016.

[8] B. Meyer, M. Bikdash, and X. Dai, “Fine-grained financial news sentiment analysis,” SoutheastCon 2017, 2017.

[9] V. S. Shirsat, R. S. Jagdale and S. N. Deshmukh, "Document Level Sentiment Analysis from News Articles," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, 2017, pp. 1-4.

[10] A. Agarwal, V. Sharma, G. Sikka and R. Dhir, "Opinion mining of news
headlines using SentiWordNet," 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, 2016, pp.1-5.doi: 10.1109/CDAN.2016.7570949

[11] J. Lei, Y. Rao, Q. Li, X. Quan, and L. Wenyin, “Towards building a social emotion detection system for online news,” Future Generation Computer Systems, vol. 37, pp. 438–448, 2014.

[12] Musto, C., Semeraro, G., and Polignano, M, A comparison of lexicon-
based approaches for sentiment analysis of microblog posts. Information Filtering and Retrieval, 59. 2014.

[13] A. Dandrea, F. Ferri, P. Grifoni, and T. Guzzo, “Approaches, tools and
applications for sentiment analysis implementation,” International Journal of Computer Applications, vol. 125, no. 3, pp. 26-33, 2015.

[14] M. Devika, C. Sunitha, and A. Ganesh, “Sentiment analysis: a comparative study on different approaches,” Procedia Computer Science, vol. 87, pp. 44–49, 2016.

[15] E. Haddi, X. Liu, and Y. Shi, “The Role of Text Pre-processing in Sentiment Analysis,” Procedia Computer Science, vol. 17, pp. 26–32,

2013.
[16] K. Ghag and K. Shah, "SentiTFIDF - Sentiment classification using relative term frequency inverse document frequency," International Journal of Advanced Computer Science and Applications, vol. 5, no. 2, 2014.
[17] G. A. Miller, "WordNet: a lexical database for english," Communications of the ACM, vol. 38, no. 11, pp. 39–41, Jan. 1995.
[18] Z. Madhoushi, A. R. Hamdan, and S. Zainudin, "Sentiment analysis techniques in recent works," 2015 Science and Information Conference (SAI), 2015.
[19] D.M.E.D.M. Hussein, "A survey on sentiment analysis challenges," Journal of King Saud University - Engineering Sciences, vol. 30, no. 4, pp. 330–338, 2018.
[20] Kravchenko D, Pivovarova L, "DL Team at SemEval-2018 task 1: tweet affect detection using sentiment lexicons and embeddings". In: Proceedings of the 12th international workshop on semantic evaluation. Association for Computational Linguistics, pp 172-176, 2018
[21] Andreevskaia A. and Bergler S. "Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses", in Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), 2006.
[22] Li, Fangtao, Minlie Huang, and Xiaoyan Zhu. "Sentiment analysis with global topics and local dependency". in Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2010).