Identifying Correlation between Sentiment Analysis and Septic News Sentences Classification Tasks

Soma Das[1,2] Sagarika Ghosh[1,3] Sanjay Chatterji[1]
soma_phd_2018july@iiitkalyani.ac.in sagarika.ghosh304@gmail.com sanjayc@iiitkalyani.ac.in[1]

[1]Indian Institute of Information Technology Kalyani, West Bengal, India
[2]Institute of Engineering & Management, India, , University of Engineering & Management, India
[3]University of Engineering and Management Jaipur, India

Abstract

This research investigates the correlation between Sentiment and SEPSIS(SpEculation, oPinion, biaS, and twISt) characteristics in news sentences through an ablation study. Various Sentiment analysis models, including TextBlob, Vader, and RoBERTa, are examined to discern their impact on news sentences. Additionally, we explore the Logistic Regression(LR), Decision Trees(DT), Support Vector Machines(SVM) and Convolutional Neural Network (CNN) models for Septic sentence classification.

Keywords: Septic Sentence, Sentiment Analysis, Textblob, RoBERTa

1 Introduction

In an era dominated by information and rapid communication, news serves as an invaluable source of knowledge, shaping public opinion and influencing decision-making. The term 'Septic' encapsulates four distinctive linguistic characteristics i.e., 'SEPSIS': SpEculation, oPinion, biaS, and twISt (Das and Chatterji, 2023). When sentences exhibit one or more of these SEPSIS properties, they are labelled as Septic sentences. We can interpret a sentence as Pure if it is not having any of these properties. Purity is irrelevant to legal and correctness qualities, in contrast to dictionary meaning. Purity, therefore, denotes goodness in general.

We believe that the news articles in the e-papers we have chosen accurately present a certain truth. However, a newspaper article may contain one or more Septic sentences. Although the context is intimately connected to Purity, each sentence's authenticity will be determined separately in this work. We regard a sentence as Pure if it cannot be resolved without context.

This research explores the fusion of machine learning methods namely Logistic Regression(LR), Decision Trees(DT), and Support Vector Machines(SVM) and deep learning techniques, specifically Convolutional Neural Networks (CNN), with advanced Sentiment analysis pretrained models to improve Septic-Pure sentence classification. We present a dataset of 6000 news sentences annotated with Septic and Pure labels, derived from reputable e-newspapers, such as The Hindu and The Telegraph. Three annotators evaluated the dataset, indicating substantial inter-annotator agreement.

Our study focuses on how Sentiment analysis tools, including TextBlob, the VADER Sentiment analyzer, and the RoBERTa model, can be used to enhance Septic-Pure sentence classification. We demonstrate that RoBERTa, a state-of-the-art transformer model, surpasses other Sentiment analysis methods and is particularly well-suited to this task. Our analysis also examines the impact of opinionated and non-opinionated sentences on Septic and Pure classification. We transform Sentiment labels into two categories: 'Opinionated(negative and positive Sentiment)' and 'Non-Opinionated(neutral Sentiment),'. It analyses the relationship between the Sentiment expressed in sentences and their classification.

2 Related Work

Numerous forms of misinformation and deceptive content are prevalent on the internet. These encompass fabricated news and deduced information that contributes to the identification of false reports (Rampersad and Althiyabi, 2020; Pulido et al., 2020; Naeem et al., 2021; Zhang et al., 2020), as well as the recognition of rumors (DiFonzo and Bordia, 2007; Zubiaga et al., 2015), and the propagation of false information (Torabi Asr and Taboada, 2019).

In their study, Das and Chatterji (2019) conducted research into automatically recognizing and labeling phrases within distorted news articles.

They utilized Machine Learning techniques and rule-based post-processing to establish the concept of legitimacy in news sentences. These sentences were categorized into two groups: Pure and Septic. Additionally, Das et al. (2022) introduced guidelines for annotating non-legitimacy in Bengali, addressing both Shallow and deep levels, while considering the semantic attributes of Bengali words.

According to Rhanoui et al. (2019), Opinion refers to an assessment, an evaluation, or an individual perspective. Dave et al. (2003) describe an opinion mining system as one capable of analyzing search results for an item. Kim and Hovy (2004) define opinion as a quadruple (Topic, Holder, Claim, Sentiment).

3 Our Task

In the landscape of natural language processing and text analysis, the task at hand is an intersection of Sentiment analysis and Septic sentence classification within news text data. We examine how the Sentiment of a sentence influences its classification as Septic or Pure, and conversely, how the presence of Septic properties might affect Sentiment labelling. To analysis this we have done the experiments with and without adding these extra feature for each cases. This ablation study proves that there is an influence of sentiment in the Septic sentence, however the opposite is not correct. The importance of the Septic sentence in the analysis of Sentiment is not important.

3.1 Data Collection and Annotation

Our analysis focuses on an English language dataset, which consists of 6000 news sentences sourced from a diverse range of genres and publications namely The Hindu [1] and The Telegraph [2]. The dataset encompasses various domains, including politics, sports, crime, business, education, celebrity news, and weather reports from English newspapers namely The Hindu and The Telegraph. It's important to note that the initial dataset exhibited an imbalance between *Pure* and *Septic* sentences. To address this, we augmented the dataset by incorporating additional Septic sentences, ensuring a more equitable representation of both categories.

The annotation task involved the collaboration of three annotators, each responsible for labelling sentences as either Septic or Pure. To assess the agreement level among annotators, we calculated inter-annotator agreement on 100 randomly selected sentences using the kappa score, resulting in a substantial agreement score of 0.75. This score underscores the strong consensus among annotators (McHugh, 2012).

3.2 Sentiment Analysis

In this experiment, we have used Twitter-roBERTa-base for the Sentiment Analysis model from huggingface[3]. This model is based on RoBERTa and was trained on approximately 124 million tweets. It was subsequently fine-tuned for Sentiment analysis using the TweetEval benchmark(Camacho-Collados et al., 2022) and (Loureiro et al., 2022).

The experimental results are presented in Table 3, highlighting the performance of various Sentiment analysis models on our dataset. RoBERTa achieved the highest accuracy, outperforming other models. Specifically, TextBlob yielded an accuracy of 70%, Vader Sentiment achieved an accuracy of 75%, while RoBERTa demonstrated superior performance with an accuracy of 89%. These results emphasize RoBERTa as the top-performing model for Sentiment analysis on our dataset.

3.3 Septic Pure Classification

Now, we apply Machine Learning and Deep Learning techniques, specifically Logistic Regression(LR), Decision Trees(DT), and Support Vector Machines(SVM) and Convolutional Neural Networks (CNN), for the classification of text data into 'Septic' and 'Pure' categories. We delve into the architecture and training process of the CNN model. The architecture of our model is shown in fig 1.

The training process involves iterating over the training data in batches of size 128 for a total of 10 epochs. 10% of the training data is set aside for validation.

The experimental results revealed that the CNN model achieved an accuracy of 85.6%. Further detailed analyses, including precision, recall, and F1-score, were conducted to comprehensively evaluate the model's performance. The results of the conducted experiments are presented in Table 1.

[1] https://www.thehindu.com/
[2] https://www.telegraphindia.com/
[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-Sentiment-latest

Input → Embedding → Dropout → Convolutional → Maxpooling → Dense → Activation → Output

Figure 1: CNN Model Architecture

Model	Accuracy	Precision	Recall
LR	80.2	82.5	81.3
DT	78.6	79.8	78.1
SVM	82.6	83.8	82.8
CNN	**85.6**	**86.2**	**84.7**

Table 1: Experimental Results of Septic Sentence Classifications (Percentage)

4 Correlation between Septic Sentence and Sentiment Analysis

Understanding the correlation between Septic sentences and Sentiment analysis is crucial. We examine how the Sentiment of a sentence (positive, negative and neutral) influences its classification as Septic or Pure, and conversely, how the presence of Septic properties might affect Sentiment labelling. This section describes the relationships and interactions between Sentiment analysis and the identification of Septic and pure sentences. To simplify the results and make them more interpretable, we categorized negative and positive Sentiments as “Opinionated" and neutral Sentiments as “Non-Opinionated."

The distribution of Opinionated (both positive and negative Sentiment) and Non-Opinionated(neutral Sentiment) sentences is visualized in Figure 2. This figure illustrates the significant impact of Opinionated sentences on Septic sentences and the prevalence of Non-Opinionated sentences in Pure sentences. Opinionated sentences tend to exhibit a higher likelihood of being associated with Septic content, while non-opinionated sentences are more likely to be indicative of Pure content. This alignment is considered favourable.

The correlation analysis reveals a notable influence of ‘Septic-Pure’ on Sentiment classification. Sentences containing SEPSIS characteristics exhibit distinct Sentiment patterns, emphasizing the significance of integrating SEPSIS-aware Sentiment analysis for more precise classification and information retrieval in news articles.

We extended our analysis by incorporating the Sentiment information as an additional feature in a Convolutional Neural Network (CNN) model. We

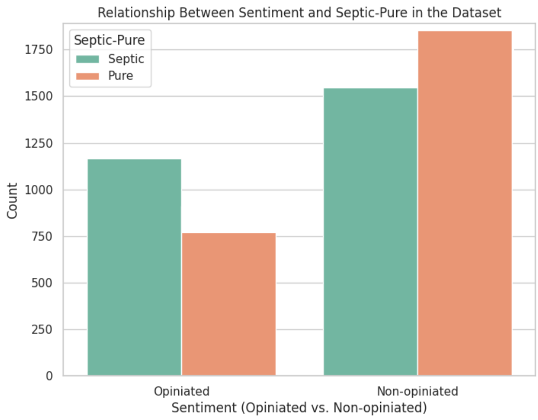

Figure 2: Distribution of Opinionated and Non-Opinionated Sentences on our Dataset

conducted experiments both with and without the Sentiment feature to evaluate its impact on classification performance.

The results, presented in Table 2, highlight the effectiveness of utilizing Sentiment information as an additional feature in the CNN model.

These results underscore the valuable role of Sentiment analysis as an additional feature for improving the accuracy of the CNN model in Septic-Pure sentence classification.

Now, we describe the impact of incorporating the ‘Septic-Pure’ feature into the Sentiment analysis task. we utilize the RoBERTa pretrained Sentiment analysis model’s output on our dataset as the Sentiment labels, followed by a manual correction process. We evaluate two different configurations for Sentiment labels, one using opinionated and non-opinionated sentences and the other using neutral, positive, and negative labels. We examine the impact of incorporating Septic-Pure classification as an additional feature for both cases in CNN model.

We observe a 2% increase in accuracy when incorporating ‘Septic-Pure’ as a feature. The results are summarized in the table 3. We further conducted experiments where we incorporated Sentiment labels as negative(0), neutral(1) and positive(2). The results are summarized in the table 4.

The reason behind this adverse effect of the Septic-Pure feature in the ternary sentiment clas-

Model Configuration	Accuracy	Precision	Recall	F1 Score
CNN without Sentiment Feature	0.85	0.84	0.86	0.86
CNN with Sentiment Feature	0.92	0.90	0.92	0.91

Table 2: CNN Model Performance of Septic Sentence Classification task with and without Sentiment Feature

Model Configuration	Accuracy	Precision	Recall	F1 Score
CNN without Septic-Pure Feature	0.76	0.75	0.76	0.75
CNN with Septic-Pure Feature	0.78	0.77	0.78	0.78

Table 3: Accuracy of the CNN model of Sentiment Analysis with and without the 'Septic-Pure' feature with Opinionated and Non-opinionated Sentiment Labels

Model Configuration	Accuracy	Precision	Recall	F1 Score
CNN without Septic-Pure Feature	0.74	0.74	0.74	0.75
CNN with Septic-Pure Feature	0.70	0.70	0.71	0.70

Table 4: Accuracy of the CNN model of Sentiment Analysis with and without the 'Septic-Pure' feature with Neutral, Positive, and Negative Sentiment Labels

sification task is understandable. Septic sentences encompass four distinctive characteristics: Speculation, Opinion, Bias, and Twist. Thus, 'Septic' represents opinionated sentences and 'Pure' represents non-opinionated sentences. Thus, Septic feature proves to be one of the contributing factors to identifying Opinionated Sentences, leading to an increase in accuracy in binary classification (Opinionated vs. Non-opinionated). However, as Opinionated sentences encompassing both positive and negative Sentiments are likely to possess SEPSIS properties, the Septic feature is unable to distinguish between positive and negative sentiment. This feature creates confusion in the ternary classification (Negative, Neutral, Positive) task leading to a decrease in accuracy.

5 Limitations of the work

While our research contributes to improving the classification of Septic and Pure sentences, there are certain constraints and challenges that warrant consideration. Our experiments shed light on intriguing aspects of the relationship between Septic-Pure features and Sentiment analysis. However, these findings come with uncertainties that impact the overall scope of our work.

In the opinionated and non-opinionated Sentiment classification task, the inclusion of the Septic-Pure feature leads to a increase(7%) in accuracy. Conversely, applying the Septic-Pure feature results on opinionated and non-opinionated Sentiment classification task leads to a decrease(4%) in accuracy. Unfortunately, the specific reasons behind these variations remain unclear.

The complexity arises from the intricate nature of news sentences, where Sentiment and Septic properties may interact in unpredictable ways. The observed increase and decrease in accuracy highlight the need for a more comprehensive understanding of the relationship between Sentiment and Septic-Pure features. Further investigation into the subtleties of these interactions is necessary to unravel the underlying factors contributing to the observed performance variations.

6 Conclusion and Future Scope

In this paper, we conducted an extensive analysis of Septic-Pure sentence classification within the context of news articles, employing a diverse set of Sentiment analysis models and a Convolutional Neural Network (CNN) architecture. We introduced Sentiment information as an additional feature in the CNN model and observed a marked improvement in its classification performance, highlighting the value of Sentiment analysis in this task. In the future, we plan to further refine our approach by fine-tuning the model with our news sentence Sentiment-annotated dataset, aiming to enhance its performance on Septic-Pure sentence classification.

References

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. 2022. Tweetnlp: Cutting-edge natural language processing for social media. *arXiv preprint arXiv:2206.14774*.

Soma Das and Sanjay Chatterji. 2019. Identification of synthetic sentence in bengali news using hybrid approach. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 193–200.

Soma Das and Sanjay Chatterji. 2023. Sanitization of sepsis news sentences with the help of paraphrasing. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '22, page 210–215, New York, NY, USA. Association for Computing Machinery.

Soma Das, Pooja Rai, and Sanjay Chatterji. 2022. Deep level analysis of legitimacy in bengali news sentences. *Transactions on Asian and Low-Resource Language Information Processing*, 21:1 – 18.

Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528.

Nicholas DiFonzo and Prashant Bordia. 2007. Rumor, gossip and urban legends. *Diogenes*, 54(1):19–35.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

Salman Bin Naeem, Rubina Bhatti, and Aqsa Khan. 2021. An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal*, 38(2):143–149.

Cristina M Pulido, Laura Ruiz-Eugenio, Gisela Redondo-Sama, and Beatriz Villarejo-Carballido. 2020. A new application of social impact in social media for overcoming fake news in health. *International journal of environmental research and public health*, 17(7):2430.

Giselle Rampersad and Turki Althiyabi. 2020. Fake news: Acceptance by demographics and culture on social media. *Journal of Information Technology & Politics*, 17(1):1–11.

Maryem Rhanoui, Mounia Mikram, Siham Yousfi, and Soukaina Barzali. 2019. A cnn-bilstm model for document-level sentiment analysis. *Machine Learning and Knowledge Extraction*, 1(3):832–847.

Fatemeh Torabi Asr and Maite Taboada. 2019. Big data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1):2053951719843310.

Jiawei Zhang, Bowen Dong, and S Yu Philip. 2020. Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1826–1829. IEEE.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Towards detecting rumours in social media. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.