Do Stopwords and Punctuation Matter? A Systematic Evaluation for Sentiment Classification

Drishya Dinesh
Trinity College Dublin
Dublin, Ireland
drishyad@tcd.ie

Parth Deshmukh
Trinity College Dublin
Dublin, Ireland
deshmukhp@tcd.ie

Mukul Ghare
Trinity College Dublin
Dublin, Ireland
gharem@tcd.ie

Tanmay Ghosh
Trinity College Dublin
Dublin, Ireland
ghoshta@tcd.ie

Raghav Manish Gupta
Trinity College Dublin
Dublin, Ireland
guptar3@tcd.ie

Yilin Wen
Trinity College Dublin
Dublin, Ireland
wenyi@tcd.ie

1 Abstract

High frequency function words (stop words) and punctuation are routinely removed in natural language processing, largely for the reason of reducing feature dimensionality and the assumption that these items contribute little discriminative information. These closed class items although carry stylistic, structural and sometimes affective information that may be relevant for tasks such as sentiment classification.

This paper aims to analyze the impact of stopword and punctuation removal on sentiment classification performance using the IMDB62 movie review dataset containing 61,987 reviews from different users about various movies, in which ratings from 1–10 are mapped to positive, negative and neutral sentiment categories. We evaluate the performance of a Support Vector Classifier (SVC) in five different scenarios: (1) all stopwords and punctuation retained, (2) only stopwords removed, (3) only punctuation removed, (4) both stopwords and punctuation removed and (5) frequency discounted features using a square root weighting scheme to reduce the influence of high frequency items without completely eliminating them. By comparing classification accuracy and precision across these conditions, we assess whether traditional deletion practices improve or degrade performance and whether frequency discounting offers a more principled alternative. The results aim to clarify the informational contribution of stopwords and punctuation to sentiment classification.

2 Introduction

Yilin Wen

Sentiment analysis extracts emotions and opinions from texts which is crucial for understanding public perception in domains like social media and online reviews (Kumar et al. (2025)).

....

3 Research Question

Raghav Manish Gupta

4 Literature Review

4.1 The Origins and Assumptions of Stopword Removal

The concept of stopwords originates in early information retrieval work. Luhn (1958) proposed removing both the most and least frequent words from document collections as a part of its process for identifying significant words to improve retrieval efficiency. It is specifically mentioned that common, frequently used words such as pronouns, propositions etc carry low discriminating value. This heuristic was subsequently formalized into exclusion lists that have been incorporated into virtually every major NLP toolkit. The assumption behind their use is that high frequency words correspond to low information value and add noise, but this idea has been challenged by both theoretical and empirical work, even though the practice remains largely unchanged in many NLP pipelines.

4.2 Stopword and Punctutaion Removal and Sentiment Classification

Symeonidis et al. (2018) provide one of the most comprehensive evaluations of text pre-processing for sentiment analysis, comparing sixteen techniques across two Twitter datasets and multiple classifiers. Their findings are particularly relevant to the role of stopwords and punctuation. The study shows that removing punctuation consistently reduced accuracy across all models, as punctuation especially repeated marks often encode sentiment intensity in microblog text. Conversely, collapsing repeated punctuation (e.g: "!!!" $\rightarrow$ "!") improves performance, highlighting its affective value. Stopword removal also failed to yield consistent gains. On one dataset it reduced accuracy for all algorithms and on the other, it offered only modest improvements. The authors argue that many stopwords carry sentiment or modify affect. These findings show that punctuation and stopwords can carry useful sentiment information and removing them without care can lower classification accuracy, suggesting that traditional deletion-based preprocessing should be reconsidered.

4.3 Punctuation as an Information Channel

Darmon et al. (2018) provide a detailed analysis of punctuation sequences as a textual signal independent of lexical content. Using a range of corpora and analysis methods, they demonstrate that punctuation patterns encode genre, syntactic rhythm and authorial habit, and that punctuation alone carries substantial discriminating information between text categories. For sentiment analysis, this is relevant because punctuation marks such as exclamation marks and ellipses carry affective load. Their reflexive removal, treated as equivalent to stopword removal in most NLP pipelines may constitute a measurable information loss, which we will investigate explicitly in our experimental design.

4.4 TF-IDF and Frequency-Sensitive Feature Weighting

4.5 The Square Root Rule as a Frequency Discounting Mechanism

4.6 Sentiment Analysis: Broader Context

Kumar et al. (2025) perform sentiment classification of movie reviews using several machine learning algorithms. The authors apply standard cleaning steps including stopword and punctuation removal before extracting TF-IDF features, which are used to train models such as Support Vector Classifier(SVC), Random Forest, Decision Tree, Naive Bayes and K-Nearest Neighbours(KNN). Their results show that SVC achieves the highest accuracy showcasing the effectiveness of SVC for text-based sentiment analysis and the suitability of TF-IDF for weighting informative words while down-weighting frequent, less meaningful terms. Although the paper removes stopwords and punctuation as part of its pipeline, it does not evaluate the impact of these steps individually, leaving the open question of whether such deletions help or harm classification performance. This gap directly motivates the present research which isolates and tests the contribution of stopwords, punctuation and frequency-discounting to better understand how these preprocessing choices influence sentiment classification accuracy.

4.7 Synthesis and Research Gap

Overall, the literature reviewed shows that stopwords and punctuation often carry useful sentiment information, so removing them can hurt accuracy, especially when the corpus contains many affective function words or important negation markers. TF-IDF helps by reducing the weight of very frequent words, but it does not solve issues like negation unless handled separately. Punctuation also encodes sentiment strength, meaning that deleting it removes valuable cues. However, no prior work has examined the combined effects of stopword removal, punctuation removal and frequency-discounting within a single study. The present research fills this gap by comparing five preprocessing conditions with specific attention to stopwords, punctuation and square-root frequency weighting.

5 Research Methods

5.1 Dataset Description

Mukul Ghare

5.2 Data Preprocessing

Parth Deshmukh

5.3 Feature Extraction

5.4 Model Implementation

6 Result Discussion

7 Conclusion

References

Darmon, A. N., M. Bazzi, S. D. Howison, and M. A. Porter (2018). Pull out all the stops: Textual analysis via punctuation sequences. *arXiv*.

Kumar, N., M. F. Alam, and N. Kumari (2025). Sentiment analysis of movie reviews using tf-idf. In *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)*, pp. 1399–1402.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development 2*(2), 159–165.

Symeonidis, S., D. Effrosynidis, and A. Arampatzis (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications 110*, 298–310.