

Enhancing Social Media Sentiment Classification Through Effective Text Preprocessing and Machine Learning

AnanSingh T
Assistant Professor
t.anansingh@gmail.com
Sapthagiri NPS University
Bangalore

Harish R
Assistant Professor
harishr@gmail.com
Sapthagiri NPS University
Bangalore

Ranganath TP
Student
Ranganathtp6468@gmail.com
Sapthagiri NPS University
Bangalore

Sathya Narayan
Student
Satyanarayana042001@gmail.com
Sapthagiri NPS University
Bangalore

Abstract— Sentiment analysis based on social media datasets is complicated since these texts usually contain informal speech, links, references to users, punctuation marks and other noisy components. In this paper, we explore the impact of text preprocessing on the process of sentiment classification via machine learning algorithms. The experiments have been performed using the sentiment subdataset of the TweetEval benchmark. This data collection includes texts with negative, neutral and positive sentiment. There were two types of text under investigation: minimally preprocessed texts and carefully preprocessed texts. For feature extraction, we applied Term Frequency-Inverse Document Frequency (TF-IDF). For classification, we utilized such algorithms as Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine (SVM). The evaluation of classifiers was performed through accuracy, precision, recall and F1-score metrics. We found that preprocessing did not positively affect all of the classifiers equally. The classifier Naive Bayes got some benefits from text preprocessing while Logistic Regression and SVM had better performance working with minimally preprocessed texts. The best performance was observed by Logistic Regression that gave accuracy of 58.95% and F1-score of 57.84%. The obtained results show that excessive text preprocessing can lead to losing valuable sentiment-related information and preprocessing methods need to be chosen according to the specificities of the dataset and classifier.

Keywords— Sentiment Analysis, Social Media, Text Preprocessing, Machine Learning, TF-IDF
Introduction (Heading 1)

I. INTRODUCTION

Social media has become a significant medium for giving voice to opinion, experience, preference and response to product, service, event and public issues. The amount of user-generated content they include is considerable and offers useful insights for organizations, researchers, and decision makers. But having to review so much information in text is tedious and time consuming. Sentiment analysis is a way of determining the emotional state of a text and, in general, classify it as positive, negative, or neutral in an automatic way. For sentiment classification, a number of techniques such as Natural Language Processing (NLP) and machine

learning (ML) are widely used. A typical sentiment analysis system consists of two phases: (1) converting the textual data into numerical features and (2) feeding the features to a classification algorithm. Computationally efficient, relatively easy to implement, and appropriate for high dimensional text features, traditional machine learning algorithms like Naive Bayes, Logistic Regression, and Support Vector Machine (SVM) still can be useful in this task. These algorithms can achieve good results when used in conjunction with Term Frequency-Inverse Document Frequency (TF-IDF), bypassing the need for complex deep learning architectures or significant computational power.

Although this is good, sentiment analysis of social media text is still difficult as the text is typically informal and unstructured in nature. Social media users often employ hyperlinks, user mentions, hash tags, punctuation, abbreviations, repeated characters, and inconsistent capitalization. A large portion of information is thus included in a single posting which may not necessarily be related to the sentiment. These noisy elements can make the feature space larger and can be detrimental to a machine learning model's ability to detect salient sentiment patterns.

Text preprocessing is, therefore, one of the crucial steps in the sentiment classification process. Typical pre-processing steps include converting text to lowercase, discarding hyperlinks and user mentions, filtering out superfluous punctuation, splitting sentences into tokens, discarding stop words and lemmatizing words. These operations are used in a variety of ways, but they are not always discussed by themselves in terms of their contribution to classification performance. Many sentiment analysis research work focus primarily on comparing sophisticated classification models, but the impact of the noisy social media text and its transformation to a clean representation is given relatively less practical consideration.

The present study aims to give light on this fact by studying the influence of effective text preprocessing on the performance of a machine learning-based sentiment classification. The study adopts a comparative experimental approach without the use of a complex model or a complicated algorithm. To investigate the impact of text

cleaning on classification results, the social media texts are processed in different ways. Social media texts are preprocessed in various ways to assess the effect of text cleaning on the classification results. The textual content is textually converted to numerical feature vectors using TFIDF method and classification models Naive Bayes, Logistic Regression, and SVM is used.

Performance of the classifiers selected is evaluated with the help of accuracy, precision, recall and F1 score. A comparative study is then carried out to find the best model and also identify the amount of improvement in sentiment classification that can be achieved by preprocessing. This allows one to understand how text preparation relates to model performance and yet keep a computationally efficient, repeatable model workflow.

This work is significant for the development of systematic & practical preprocessing framework for noise reduction in social media text before the process of Sentiment classification. It compares the effect of text preprocessing on classification accuracy by conducting comparative experiments and reports the results of three widely used machine learning algorithms that are used with a common feature representation, TF-IDF. These models are evaluated for their performance using various metrics such as accuracy, precision, recall, and F1-score. Moreover, the study has presented a simple and repeatable framework for sentiment classification without special hardware and a high level of complexity deep learning architecture.

II. LITREATURE REVIEW

Rosenthal et al. [7] introduced the benchmark of the SemEval-2017 Task 4 for sentiment analysis in Twitter as an important benchmark for sentiment classification methods. The study focused on different sentiment-related tasks such as sentiment classification of tweets into positive, negative, and neutral categories. The authors emphasized that the context, expressions and irregularity of short and informal social media messages can have a negative impact on sentiment. The benchmark gave a common ground for which various sentiment analysis methods could be compared and was also an important precursor to social media sentiment datasets.

Barbieri et al. [9] presented TweetEval, a common benchmark for the evaluation of tweet classification techniques for various tasks in social media analysis. The benchmark features sentiment analysis and offers fixed partitions for training, validation, and testing to ensure consistent model evaluation. Their work highlighted the need to have standardized datasets for comparison of the classification approaches. The present study is also relevant to the sentiment component of TweetEval, as this component has negative, neutral, and positive social media posts and can serve as a good ground for investigating the impact of preprocessing on the performance of the classification model. Effrosynidis et al. [19] reviewed various pre-processing methods applied to tweets for sentiment analysis, and investigated the effects of each pre-processing step on the classification. They concluded that the preprocessing operations do not have the same effect on the model performance and the usefulness of any specific operation might vary from one data set to another and from one classification method to another. The study revealed that the

removal of noisy elements can enhance the textual consistency, however some cleansing may be more than necessary and useful sentiment information may be lost. It supports the necessity of an experimental assessment of preprocessing instead of the assumption that pre-processing a text extensively is always beneficial for sentiment classification.

Symeonidis et al. [18] have carried out a comprehensive comparative study of the preprocessing techniques for Twitter sentiment analysis and their interactions. The authors explored various strategies of text pre-processing and illustrated that text pre-processing could significantly impact classification accuracy. They also concluded that the preprocessing/classification relationship is non-linear in nature since some preprocessing is helpful in one experimental context and less useful in the other. The present study is closely related with this one as it compares Naive Bayes, Logistic Regression and SVM among different social media text representations with the help of TF-IDF, either with minimal processing or systematic pre-processing.

III. METHODOLOGY

The research uses a comparative experimental method to evaluate the influence of text preprocessing on social media sentiment classification. The research was performed based on an open-source social media dataset. The main steps of the methodology involve collecting data, preprocessing texts, extracting features, training the model, and evaluating its performance.

A. Dataset Description

The sentiment dataset used in this work was collected from the TweetEval official repository provided by the Cardiff NLP team. TweetEval is a benchmark dataset which is created to be applied in various tweet classifications. In the current research, only sentiment classification dataset was utilized.

The public link to the dataset is presented below:

Source of Dataset: <https://github.com/cardiffnlp/tweeteval>

The sentiment dataset consists of English language tweets categorized into 3 classes: negative, neutral, and positive. The tweets have real social media text that contains informal words, hyperlinks, usernames, hashtags, punctuation, and other features typical of online communication.

There are separate training, validation, and test sets available in the repository. This dataset structure was used in the experiment in order to get replicable results. The training set was used for model training, and test set – for evaluation of the performance.

B. Text Preprocessing

The central aim of this research is to find out if text preprocessing affects sentiment classification. That is why the experiments have been done on two variants of the same set of data. The first variant had minimally preprocessed tweets and served as the baseline for the experiments. The second variant was preprocessed according to the certain scheme.

This preprocessing scheme is as follows:

Tweet Before Preprocessing → Converting to Lowercase → Eliminating Hyperlinks and User Mentions → Deleting Unnecessary Punctuation and Special Characters → Removing Stopwords → Preprocessed Tweet

The performance of the models on the original data and preprocessed data has been compared to find out the impact of text cleanin.

C. Feature Extraction

The texts have been transformed into numbers using Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF places more emphasis on the terms which can help differentiate one document from another and less emphasis on terms which appear very often in the corpus. The TF-IDF Vectorizer has been fitted on training data only and later applied on test data. The process of feature extraction has been done in the same way for both minimally processed text and preprocessed text.

D. Machine Learning Models

Three popular algorithms for machine learning were chosen for sentiment classification: Multinomial Naive Bayes, Logistic Regression, and SVM. The choice was made due to the simplicity and high efficiency of these models in text classification. Naive Bayes algorithm uses probability calculation for classifying texts. Logistic Regression recognizes the relationship between textual data and sentiments. SVM determines the boundary separating the three classes of sentiments. All of the algorithms were trained individually with the use of TF-IDF features.

E. Experimental Procedure

These experiments were carried out under two different text scenarios. The first experiment involved the training of the models with minimally processed text. The second experiment involved the training of the models with preprocessed text.

The general process was as follows:

TweetEval Dataset → Two Text Scenarios → TF-IDF Features Extraction → Naïve Bayes, Logistic Regression, and SVM → Performance Evaluation

The same training and testing dataset was used for all models.

F. Performance Evaluation

The performance of the three models was measured by means of accuracy, precision, recall, and F1 score. While the accuracy represented the total correctness of the models' prediction, the precision and recall showed how well the models predicted the sentiment classes. Additionally, the confusion matrix was applied to assess both correct and wrong classification of negative, neutral, and positive tweets. The performances of the three models were compared both before and after the preprocessing to find out which model performed better and to estimate the influence of text preprocessing on sentiment classification.

IV. RESULTS AND DISCUSSION

The experiments were performed on the dataset of sentiments which were extracted from the official TweetEval repository. The dataset included 45,615 training tweets, 2,000 validation tweets, and 12,284 test tweets. There were 3,972 negative, 5,937 neutral, and 2,375 positive tweets in the test set. Three models, including Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine (SVM), were tested under two types of text condition: the one that had undergone minimal processing and the systematic preprocessing of text.

The TF-IDF approach was utilized to convert the textual information into numerical features. In order to have an experiment framework as simple as possible, only 5,000 TF-IDF features were used. The same training and test sets were applied to all six experimental scenarios.

A. Overall Classification Results

The results of all the three ML algorithms for the two preprocessing scenarios are shown in Table I. Accuracy, weighted precision, weighted recall, and weighted F1-score have been considered for comparison.

TABLE I: PERFORMANCE COMPARISON OF SENTIMENT CLASSIFICATION MODELS

Text Condition	Model	Accuracy	Precision	Recall	F1-Score
Minimally Processed	Multinomial Naive Bayes	55.33%	59.23%	55.33%	51.89%
Minimally Processed	Logistic Regression	58.95%	59.89%	58.95%	57.84%
Minimally Processed	SVM	57.83%	58.30%	57.83%	57.14%
Systematically Preprocessed	Multinomial Naive Bayes	55.36%	58.36%	55.36%	52.88%
Systematically Preprocessed	Logistic Regression	58.03%	58.81%	58.03%	56.97%

Systematically Preprocessed	SVM	56.71 %	57.17 %	56.71 %	56.07 %
-----------------------------	-----	---------	---------	---------	---------

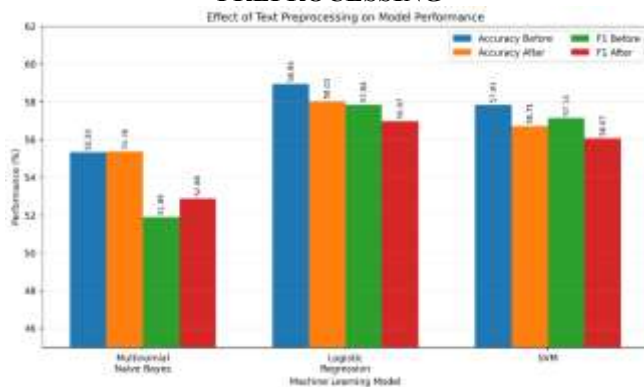
The findings indicate that Logistic Regression on minimally processed texts performed better compared to the other algorithms. It gave 58.95% accuracy, 59.89% precision, 58.95% recall, and 57.84% F1-score. SVM gave the second-best results after applying it to the minimally processed text with 57.83% accuracy and 57.14% F1-score.

Multinomial Naive Bayes had worse overall performance in terms of accuracy than Logistic Regression and SVM. Nevertheless, Multinomial Naive Bayes is the only algorithm that had slight improvement after the systematic preprocessing. Accuracy and F1-Score of Multinomial Naive Bayes changed slightly from 55.33% to 55.36% and from 51.89% to 52.88% correspondingly.

B. Effect of Text Preprocessing

The main objective of the study was to determine whether systematic preprocessing improves sentiment classification. Table II presents the change in accuracy and F1-score after preprocessing.

TABLE II: EFFECT OF SYSTEMATIC TEXT PREPROCESSING



Results have shown that there was no consistent improvement of all classifiers due to systematic preprocessing. There was only a slight increase in accuracy of Naïve Bayes with a slightly greater improvement in the F1 score. On the other hand, logistic regression and support vector machine both performed worse than before preprocessing, with logistic regression having its accuracy drop from 58.95% to 56.97% and the F1 score going down from 57.84% to 56.07%. Similar trends were also seen in support vector machine, which had an accuracy fall from 57.83% to 56.71% and F1 score decline from 57.14% to 56.07%.

Such results demonstrate that some of the elements being filtered out during systematic preprocessing carry important information for sentiment classification. Punctuation marks, repetitions of letters, capitalization, hashtags, among other things, represent sentiment expression on social media. Such elements are filtered out through preprocessing, and as a result, some information is lost.

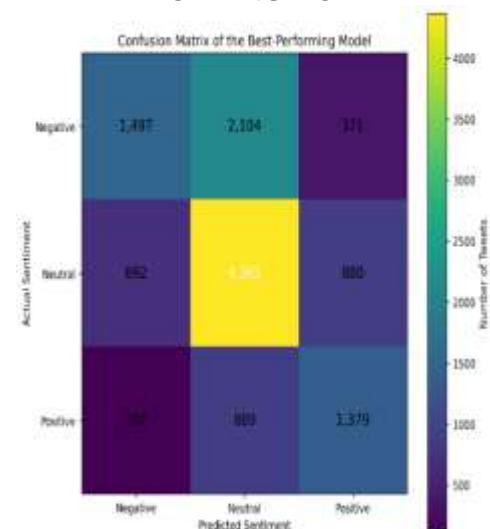
C. Model Comparison

Of all three models, Logistic Regression showed the most effective classification results. Its effectiveness means that a linear model is capable of classifying sentiments in combination with TF-IDF features. Moreover, the effectiveness of this model in classification was especially noticeable for the least preprocessed tweets, which means that it is advantageous to use the minimum amount of preprocessing of social media texts. SVM performed almost equally well as the rest two models and took the second place in both experiments. Its effectiveness was comparable to that of Logistic Regression; however, the accuracy of this model was most affected by preprocessing. Naive Bayes showed the lowest accuracy among the rest two models. On the other hand, its response to preprocessing was not similar to the one of the other two models.

D. Confusion Matrix Analysis

The confusion matrix of the best-performing model, Logistic Regression with minimally processed text, is presented in Table III.

TABLE III: CONFUSION MATRIX OF THE BEST-PERFORMING MODEL



From the confusion matrix, it is seen that the classifier classified 1,497 negative sentiments, 4,365 neutral sentiments, and 1,379 positive sentiments. Neutral sentiments were the class with the highest number of correctly classified tweets. This might be due to the presence of many neutral sentiments in the test dataset. There was a classification problem between negative and neutral sentiments. Of the 3,972 negative sentiments, 2,104 were classified as neutral. In addition, 889 positive sentiments were classified as neutral. This shows that neutral sentiments were the most difficult barrier between negative and positive sentiments.

The reason for the existence of the classification problem among sentiments is the type of language used in social media. It could be the case where one tweet contains different sentiments, indirect meanings, sarcastic sentiments, or context-based sentiments. The representation of the importance of a term does not reveal the context.

E. Discussion of Findings

It has been shown that the effect of text preprocessing on social media sentiment classification is model-dependent. It was expected that cleaning from noise will positively affect all three classification models; however, the obtained results were not as clear-cut as this expectation. Preprocessing was helpful for Naive Bayes (especially regarding the F1-score), while it decreased the accuracy of Logistic Regression and SVM. This result proves that heavy preprocessing is not the right way for social media text classification because features perceived as noisy still contain valuable data related to the user's sentiments.

The research has also revealed that a basic machine learning framework allows performing sentiment classification effectively without using complex deep learning models. The most efficient method of classification was Logistic Regression in combination with TF-IDF feature extraction which required only a moderately complex implementation and produced high accuracy of prediction. It can be concluded that the efficiency of preprocessing depends on the selected classifier. Instead of applying any set of rules for cleaning social media text regardless of the task, preprocessing methods need to be validated experimentally. As a result of the current study, the use of minimally processed text was found to be the most effective strategy for Logistic Regression and SVM, while preprocessing improved the work of Naive Bayes classifier. The key finding of the study is that effective preprocessing does not equal maximal removal of the text elements.

V. CONCLUSION

This work focused on the impact of text preprocessing on the sentiment classification of social media posts using machine learning approaches. For the experimental part of this research, the subset containing the sentiment data was extracted from the TweetEval database. The research compared three popular classifiers, namely, Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine. The extraction of features from the text was performed with the help of the TF-IDF algorithm. The evaluation of the three models was carried out on both minimally preprocessed and systematically preprocessed text samples. Experimental results have demonstrated that text preprocessing was beneficial for improving the performance of all three classifiers to the same degree. Only Multinomial Naive Bayes classifier showed a certain increase in performance due to the application of systematic text preprocessing, especially in terms of F1-score. On the contrary, the use of minimally preprocessed text led to better outcomes for Logistic Regression and SVM classifiers. Logistic Regression yielded the highest results with the minimally preprocessed text, showing an accuracy of 58.95% and an F1-score of 57.84%. Such results indicate that preprocessing of text is not always beneficial in terms of improved sentiment classification of social media data. There is some information in such texts that can be useful for classification, namely, information contained in punctuation, hashtags, repeating of the characters and other informal elements of the text. Thus, the selection of preprocessing should be made according to the specificities of the dataset and classification algorithm, but not as a mandatory step of text cleaning. The research has proved that traditional ML

algorithms offer a rather simple solution to the problem of social media sentiment analysis. The combination of TF-IDF and Logistic Regression offered the best results in the absence of deep learning architecture. However, the difficulties in distinguishing between negative/positive and neutral tweets demonstrate the limitation of the traditional features representation. The future works may include the research of the effect of selective preprocessing on the sentiment classification of social media data.

REFERENCES

- [1] Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.
- [2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1–2, pp. 1–135, 2008.
- [3] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, Eds., *A Practical Guide to Sentiment Analysis*. Cham, Switzerland: Springer, 2017.
- [4] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014.
- [5] A. Giachanou and F. Crestani, "Like it or not: A survey of Twitter sentiment analysis methods," *ACM Computing Surveys*, vol. 49, no. 2, Art. no. 28, Jun. 2016.
- [6] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decision Analytics Journal*, vol. 3, Art. no. 100073, Jun. 2022.
- [7] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 Task 4: Sentiment analysis in Twitter," in *Proc. 11th Int. Workshop Semantic Evaluation (SemEval-2017)*, Vancouver, BC, Canada, 2017, pp. 502–518.
- [8] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 Task 4: Sentiment analysis in Twitter," in *Proc. 10th Int. Workshop Semantic Evaluation (SemEval-2016)*, San Diego, CA, USA, 2016, pp. 1–18.
- [9] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1644–1650.
- [10] Cardiff NLP, "TweetEval: Repository for the TweetEval benchmark," GitHub. [Online]. Available: <https://github.com/cardiffnlp/tweeteval>. Accessed: Jul. 10, 2026.
- [11] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.
- [12] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98*. Berlin, Germany: Springer, 1998, pp. 137–142.
- [13] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in *AAAI-98 Workshop on Learning for Text Categorization*, Madison, WI, USA, 1998, pp. 41–48.
- [14] C. C. Aggarwal and C. Zhai, Eds., *Mining Text Data*. New York, NY, USA: Springer, 2012.
- [15] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good, the bad and the OMG!," in *Proc. 5th Int. AAAI Conf. Weblogs and Social Media (ICWSM)*, Barcelona, Spain, 2011, pp. 538–541.
- [16] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford University, Stanford, CA, USA, CS224N Project Report, 2009.
- [17] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011.

- [18] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for Twitter sentiment analysis," *Expert Systems with Applications*, vol. 110, pp. 298–310, Nov. 2018.
- [19] D. Effrosynidis, S. Symeonidis, and A. Arampatzis, "A comparison of pre-processing techniques for Twitter sentiment analysis," in *Research and Advanced Technology for Digital Libraries*. Cham, Switzerland: Springer, 2017, pp. 394–406.
- [20] [F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [21] "Treatment episode data set: discharges (TEDS-D): concatenated, 2006 to 2009." U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, August, 2013, DOI:10.3886/ICPSR30122.v2