Sentiment Analysis Research Project

CS 583

Submitted by Sanjna Chippalaturthi and Asritha Pidikiti

Introduction

The present work deals with the sentiment analysis of tweets concerning former U.S. President Barack Obama and politician Mitt Romney. It is categorizing the tweets into positive, negative, and neutral sentiments, which is expected to enable an understanding of public attitudes regarding their campaign times. Sentiment analysis is useful since it employs Natural Language Processing (NLP) techniques to mine subject information from the text data to attune opinions easily.

The present project is a application of machine learning models that include the setting up of the logistic regression, naive bayes, and the support vector machines (SVM) to preprocess the raw datasets collected from Twitter, prepare numerical features from the dataset, and finally train a model for accurate sentiment classification.

This document provides the different methods and experiments undertaken, as well as results obtained from the mentioned sentiment analysis project, where the comparative experimentations of the models and their effectiveness with respect to sentiment nuances are reported.

Preprocessing

Dataset included the tweets concerning Obama and Romney, marked for sentiment classification. Each tweet is assigned a sentiment label using negative, neutral, positive, and mixed as -1, 0, 1, and 2 respectively. The preprocessing for machine learning began by cleaning the text of the tweets and got rid of any special characters, mentions, hashtags, and hyperlinks such as noise, then just their meaningful content remained. Converts whole text into lower case for uniformity purposes.

Next is tokenization that breaks apart a tweet into individual words. Next is removal of stopwords such as "and," "the," and "is," those words that do not contribute towards a sentiment analysis. Stemming and lemmatization were then done in order to incrust words in their root forms like "running" was reduced to "run" thereby limiting redundancy. All this ensured that the text data was simplified and standardized for more manageable analysis by models.

Finally, preprocessed text was numericalized using TF-IDF vectorization. Words turned into numbers by this method kept the weight of each word in the paper as compared to their occurrence in the entire set of tweets. The ready structured dataset of numerical features on sentiment labels was set to train-test split to make it balanced and fair for model evaluation. This thorough preprocessing prepares the data for sentiment classification based on machine learning techniques.

Model Choice and Justification:

Logistic Regression has been preferred over Naive Bayes for sentiment analysis due to the high performance achieved regarding such techniques working with feature-rich numerical representations. It is most effective in either binary or multiclass classification problems and works rather well when the observations include informative features drawn from text data and metadata such as tweet length, number of hashtags, and mentions. Also, since it permits probabilistic predictions, it is most convenient for situations that require confidence scores or for the next downstream decision processes.

Although Naive Bayes provides high computational efficiency and simplicity, the major disadvantage of Naive Bayes is the assumption of feature independence, which is typically not the case in text data. Some n-grams or metadata features may tend to have correlation; hence, the model produces portions of this relationship making Logistic Regression a preferred model for this dataset. This logic coupled with applying class imbalanced learning improves performance for Logistic Regression over Naive Bayes, especially for SMOTE, which generates synthetic samples from underrepresented classes. The method of over-sampling enhanced class separation while allowing the model to learn more balanced decision boundaries, thus giving Logistic Regression an edge against Naive Bayes.

The preprocessed text data was fed to TF-IDF Text Vectorization with unigrams and bigrams to capture context and behavior patterns in the tweeting. Metadata features such as tweet length, number of hashtags, and mentions formed a feature set together with the TF-IDF matrix. The entire combination of features was used to train the Logistic Regression model with 'balanced' class_weight to help negate the class imbalances in case of using the model to fit. The result from all preprocessing and feature engineering steps made Logistic Regression able to harness text and metadata to render a strong classifier that could make both balanced and accurate predictions of sentiment.

Experimental Results:

Obama - Table:

Model	Accuracy	Precision	Recall	F1
Logistic regression	0.85	0.87	0.86	0.85
SVM	0.78	0.73	0.74	0.72
Naive Bayes	0.62	0.63	0.62	0.62

Romney - Table:

Model	Accuracy	Precision	Recall	F1
Logistic regression	0.71	0.73	0.70	0.72
SVM	0.57	0.56	0.57	0.55
Naive Bayes	0.58	0.58	0.58	0.52

Conclusion

The outcome of the sentiment classification experiment explains the variation with which different models of machine learning exist in efficiency and efficacy to carry out the project of analyzing tweets with respect to Barack Obama and Mitt Romney.

The analysis pointed out that, for the Obama dataset, the best model is Logistic Regression, which achieved an analysis accuracy of 85%, precision weight of 87%, and an F1-score of 85%. Therefore, it proves to have a great capability in predicting and capturing the sentiment patterns within this dataset. SVM performed reasonably well in the regard of accuracy at 78% and F1 score of 72% while Naïve Bayes brought up the rear in the performance with an accuracy of 62% and an F1 score of 62%.

The Romney dataset performed relatively well overall for all models. Again, Logistic Regression performed the best: 71% accuracy and 72% F1 score. Naive Bayes and SVM achieved similar results of 58% and 57% accuracy, respectively. However, Naive Bayes performed slightly better in precision and recall compared to SVM.

In conclusion, Logistic Regression seems to be the strongest and the most reliable method in this study on sentiment analysis because it captures fairly both precision and recall on both datasets, showing that it generalizes well on the complexities of the text. The performance of SVM and Naive Bayes is, however, weaker for the Romney dataset, revealing that these two models may not capture granularity in this specific kind of data.

Future works will examine deeper models such as deep learning architectures e.g., BERT, and other features like engineering techniques to improve the performance of these models, particularly on datasets which have a stronger sentiment distribution.

References

[1] Aliman, G., et al. "Sentiment analysis using logistic regression." *Journal of Computational Innovations and Engineering Applications* 7.1 (2022): 35-40.

[2] Prabhat, Anjuman, and Vikas Khullar. "Sentiment classification on big data using Naïve Bayes and logistic regression." *2017 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 2017.

[3] Goel, Ankur, Jyoti Gautam, and Sitesh Kumar. "Real time sentiment analysis of tweets using Naive Bayes." *2016 2nd international conference on next generation computing technologies (NGCT)*. IEEE, 2016.

[4] Aftab, Shabib, Ali Umer, and Shah Khusro. "Sentiment Analysis of Tweets Using SVM." 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, 2017.