Classifying Sentiments in Tweets on the Sentiment140 Dataset

Davi Reis e Sarah Pimenta

I. Introduction

We analyze the Sentiment140 dataset to assess the performance of a Logistic Regression classifier for sentiment analysis. Various aspects are explored, such as pre-processing, model evaluation, dataset size impact, and topic-based analysis, to identify improvement strategies.

II. Dataset

The Sentiment140 dataset contains 1.6 million tweets labeled as positive or negative, collected from Twitter. Each tweet includes fields such as sentiment label, tweet ID, and content, though the query field is not relevant for this study [1].

III. Classification pipeline

The classification pipeline involved several pre-processing steps. First, all text was converted to lowercase for consistency. Then, we applied regular expressions to remove URLs, mentions, and non-alphabetic characters, focusing on meaningful textual content. Finally, lemmatization using WordNetLemmatizer was performed to reduce words to their base forms, standardizing variations.

IV. Evaluation

The Logistic Regression classifier was evaluated using accuracy, as the dataset is balanced. After testing with multiple random states (42, 52, and 62), the model achieved an average accuracy score of 0.7646. The model tends to predict positive sentiments more frequently, leading to a higher number of false positives (43,599) compared to false negatives (31,739).

Prominent words identified by the model include "*proud*" and "*smile*" for positive sentiments and "*sad*" and "*disappointing*" for negative sentiments. These words make sense in the context of sentiment classification, as they are strongly associated with emotional expressions. However, reliance on individual words rather than the full context can lead to misclassifications, particularly in cases of sarcasm. For instance, sarcastic tweets using positive words like "*great*" to convey negative sentiments can cause the model to incorrectly classify them as positive, contributing to the elevated number of false positives.

While the model captures sentiment indicators effectively, it struggles with context dependency, especially in cases like sarcasm. Words like "*missin*" could be used both positively ("Missin' the good times") or negatively ("Missin' my flight"). The classifier, which relies on the Bag-of-Words approach, assigns weights based on word frequency without considering the surrounding context, limiting its ability to understand nuances like sarcasm or ambiguity.

V. Dataset size

Our assessment of dataset size reveals that increasing the training set from 10% to 90% improves test accuracy from 0.755 to 0.767. However, as seen in Figure 1, the accuracy gains become marginal beyond 70%, indicating limited room for further improvement through additional data. The training accuracy also decreases from 0.78 to 0.77, showing reduced overfitting as the dataset size grows. Given the diminishing returns, increasing the dataset size further is unlikely to significantly boost accuracy. From a business perspective, expanding the dataset may not be a feasible solution, as the costs associated with obtaining more data would likely outweigh the small accuracy improvements. Alternative methods, such as model tuning or feature engineering, may offer more substantial benefits.

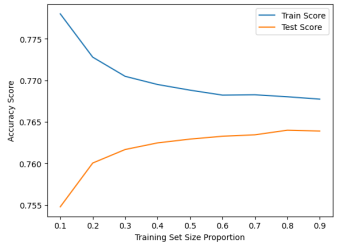

Fig. 1. Improvement of test accuracy with increased dataset size. The accuracy gains become marginal beyond 70%.

VI. Topic analysis

We used Latent Dirichlet Allocation (LDA) to identify key topics [5] and assess classifier performance. Topics 3 and 1 had the highest accuracy (0.783 and 0.770), showing better pattern recognition. Topic 4 had the lowest accuracy (0.748), indicating more difficulty in classification.

In addition to Logistic Regression, we implemented a two-layer classifier. First, we applied LDA to group the documents by topic. Next, a topic-specific classifier was applied to each group. This approach improved the overall accuracy from 0.7646 to 0.7711, demonstrating that topic-specific models can better capture nuances in each category.

References

[1] A. Kazanov, "Sentiment140 Dataset," Kaggle, 2015. [Online]. Available: https://www.kaggle.com/datasets/kazanova/sentiment140. [Accessed: 28-Sep-2024].

[2] L. Hu, J. Cao, and Z. Zhao, “A survey of federated learning for edge computing: Research problems and solutions,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 223-239, Jan. 2023, doi: 10.1109/JSAC.2022.3218501.
[3] W. Chen, H. Liu, L. Liu, and X. Zeng, “Metaverse: Security and privacy concerns,” *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3603-3610, March 2022, doi: 10.1109/JIOT.2021.3067656.
[4] GitHub Repository Link: https://github.com/sarahp31/ai_nlp.git
[5] Blei, David Ng, Andrew Jordan, Michael. (2001). Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3. 601-608.