Lab 3

AML

Elia Faure-Rolland
Tarek Saade
Ana Martinez
Jonas Thalmeier

04/06/2025

EURECOM
Sophia Antipolis

1 Introduction

This report presents a comparative study of two approaches for classifying tweet sentiment into positive, neutral, and negative categories: a traditional TF-IDF + Logistic Regression baseline and a modern BERT-based fine-tuned model. The task is based on a dataset from Figure Eight's Data for Everyone platform, which contains labeled tweets reflecting user sentiment. Additionally, we explore an optional bonus task aimed at identifying the most sentiment relevant words within tweets, a challenging problem due to the subjective nature of language and annotation inconsistencies.

2 Methodology

2.1 Data Preprocessing

The dataset used in this study consists of tweets labeled with sentiment categories, positive, neutral, and negative, originally sourced from Figure Eight's crowdsourced platform. The dataset was partitioned into training (70%), validation (15%), and test (15%) subsets to ensure rigorous evaluation. The preprocessing pipeline involved systematic cleaning of raw tweet data. We removed URLs, user mentions, and non-alphanumeric characters while intentionally preserving emoticons such as ":)" and ":-(" due to their strong correlation with sentiment expression.

2.2 Model Architectures

We implemented and compared two fundamentally different approaches to sentiment classification: a traditional TF-IDF with logistic regression baseline and a modern BERT-based deep learning model. These approaches represent distinct paradigms in natural language processing with significant differences in their underlying mechanisms and capabilities.

The **Term Frequency-Inverse Document Frequency (TF-IDF)** approach is a well established method in text processing that converts raw text into numerical features while preserving important linguistic patterns. The TF component measures how frequently a term appears in a document, while the IDF component downweights terms that appear frequently across all documents. This combined metric helps identify terms that are particularly unique and informative for classification while reducing the impact of common but uninformative words.

In our implementation, we used a TF-IDF vectorizer with a logistic regression classifier using unigrams and bigrams. The model processes text by extracting both individual words (unigrams) and consecutive word pairs (bigrams), and creates a numerical representation where terms that are frequent in specific tweets but uncommon overall receive higher weights, making them more influential in the classification process. The weighted features are then fed to a logistic regression classifier that learns to associate these textual patterns with sentiment categories.

In contrast to the traditional approach, we employed **Bidirectional Encoder Representations from Transformers (BERT)**, a state-of-the-art transformer-based language model that has revolutionized NLP through its deep contextual representations. Unlike traditional methods that process text sequentially, BERT uses a multi-layer bidirectional transformer architecture that can simultaneously consider all words in a sentence and their complex interrelationships.

Our second approach leverages transfer learning through fine-tuning of the `bert-base-uncased` model over three training epochs. Fine-tuning BERT on the specific task allows the model to adapt its general language understanding to the nuances of the dataset at hand. For our sentiment analysis task, we added a classification layer on top of BERT's [CLS] token representation. During fine-tuning, we trained for three epochs with a learning rate of 2e-5 and batch size of 64.

This comparative design allows examination of both traditional and state-of-the-art NLP techniques.

3 Results

3.1 Sentiment Classification Performance

Classification efficacy was evaluated using macro F1 scores, with results summarized in Table 1. The BERT model demonstrated superior performance, achieving a 0.78 F1 score on both validation and test sets. This represents a significant 10 percentage point improvement over the TF-IDF baseline, which scored 0.69 and 0.68 respectively. However, this performance gain comes with substantial computational tradeoffs - BERT requires nearly 85 times longer for inference than the lightweight TF-IDF approach.

Table 1: Classification Performance (Macro F1)

Model	Validation F1	Test F1	Inference Time
TF-IDF	0.69	0.68	0.14s
BERT	0.78	0.78	11.87s

3.2 Bonus Task: Relevant Word Identification

For sentiment-relevant word identification, we evaluated two distinct approaches: attention-based token selection and Integrated Gradients attribution.

The first method involves selecting the most significant tokens based on their attention values. This method reveals which words the model prioritizes in its internal computations, though these highlighted terms may not always directly contribute to the final sentiment classification.

The second method, on the other hand, performs token selection based on the type of task being addressed by examining how each token influences the model's loss function through gradient analysis.

While attention-based selection identifies tokens deemed important by the model, it does not necessarily consider their impact on the classification outcome. As a result, it may highlight words that are not particularly relevant to the sentiment classification decision. In contrast, selecting tokens based on gradient values can provide more targeted and task-relevant insights. Since the gradient is directly affected by the classification loss, this method emphasizes tokens that have a stronger impact on the final prediction.

To account for the variable number of sentiment-relevant words in different tweets, we adopt a threshold-based evaluation approach rather than using a fixed word count across all examples. For each method, multiple threshold levels are compared to determine how many of the top-ranked words (based on either attention weights or gradient values) should be considered as significant for the sentiment classification. This flexible approach recognizes that some tweets may contain only one or two strongly indicative words, while others might require several terms to fully capture their sentiment.

To measure performance, we compute Jaccard similarity scores between the words selected by each method (at various threshold levels) and the human-annotated ground truth. This allow us to assess how well each technique aligned with human judgment across different selection strictness levels, rather than assuming all tweets should have the same number of relevant words identified.

Figure 1 illustrates the relationship between threshold values and Jaccard similarity scores for both methods.

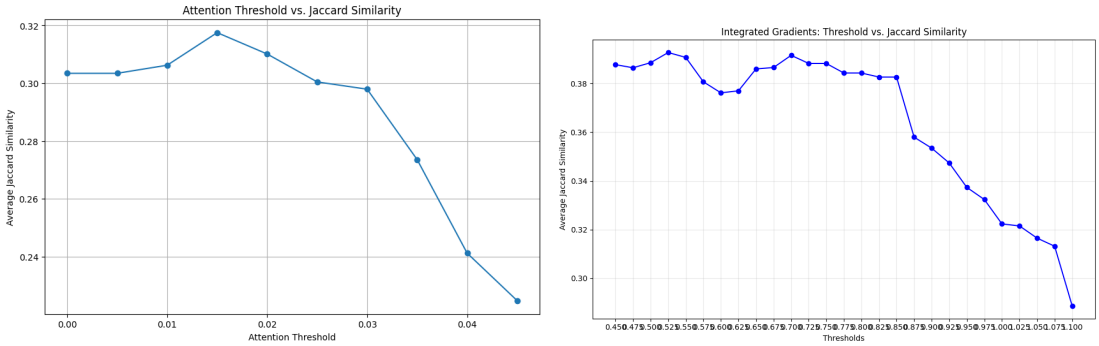

Figure 1: Jaccard scores vs. threshold for attention (left) and gradient methods (right)

Several critical observations emerged from this analysis. First, neutral tweets were intentionally excluded from evaluation since the entire tweet is considered relevant in such cases. Both methods achieved relatively low maximum Jaccard scores of 0.163, indicating limited alignment with human annotations. We noted substantial inconsistencies in the ground truth annotations themselves, which complicates the evaluation. Furthermore, neither method demonstrated a clear optimal threshold for balancing precision and recall in relevant word identification.

4 Discussion

The results confirm BERT's superiority for sentiment classification despite its computational demands. The relevant word identification task proved considerably more challenging than initial classification. Our analysis suggests annotation inconsistencies in the ground truth data may fundamentally limit evaluation reliability. The Jaccard similarity metric itself may be suboptimal for this task, as it doesn't account for semantic relevance or partial overlaps.

Future work should consider several promising directions. Twitter-specific language models like BERTweet could better capture platform-specific linguistic patterns. Alternative evaluation metrics such as F1-overlap or ROUGE might better assess partial matches. Most importantly, human evaluation of selected text spans would provide more reliable assessment than automated comparison with potentially inconsistent annotations. Sequence labeling approaches framing the problem as token classification rather than attribution analysis might yield better results.

5 Conclusion

This study confirms the superior performance of BERT over traditional TF-IDF approaches for sentiment classification on Twitter data. BERT achieved significantly higher macro F1 scores in both the validation and the test sets. However, this improvement comes with a substantial increase in computational cost.

The secondary task of identifying sentiment-relevant words within tweets proved to be considerably more challenging. Both attention-based and gradient-based attribution methods struggled to align with human annotations, reflecting the subjective nature of the task and inconsistencies in the labeled data. Additionally, the use of Jaccard similarity as the evaluation metric may not fully capture the semantic relevance of selected words.

Future work should address these limitations by developing more consistent and robust evaluation

strategies. On the modeling side, sequence labeling architectures and span prediction models offer a promising alternative to attribution-based methods.