Twitter Sentiment Classification: A Comparison of Word Embeddings, TF-IDF, and Transformers

Lilian Noé , Esther Barriol , Sophie Urrea

EPFL – CS-433 Machine Learning

***Abstract*—We address the EPFL CS-433 text sentiment classification task, which consists in predicting whether a tweet originally contained a positive or negative smiley, based solely on its textual content. We evaluate multiple text representations and classification models, including averaged GloVe embeddings [1], TF-IDF features, and transformer-based approaches. Classical linear and non-linear classifiers are compared under a consistent validation protocol, and their performance is analyzed on a large-scale Twitter dataset. Among the evaluated models, a DeBERTa [2] transformer achieve the strongest validation performance, outperforming linear models and embedding-based baselines.**

I. Introduction

Sentiment classification of short social media texts poses challenges due to informal language, limited context, and high lexical variability. The EPFL text classification challenge proposes a standard baseline based on word embeddings averaged at the tweet level. In this project, we study the impact of text representation choices (averaged word embeddings, TF-IDF features, and contextualized transformer representations), classifier selection, and hyperparameter tuning. In addition, we explore a transformer-based model to assess the benefits of contextual representations over static embeddings.

II. Dataset and Task

The dataset consists of Twitter messages that originally contained either a positive or negative smiley, which has been removed prior to release. Tweets are provided as whitespace-tokenized sequences of words, and the test set does not include labels. The task is framed as a binary classification problem, where each tweet must be assigned a sentiment label.

The training set contains approximately 2.5 million labeled tweets, evenly split between positive and negative classes. Model selection and hyperparameter tuning are performed using a fixed stratified validation split derived from the training data. The test set consists of unlabeled tweets provided by the challenge and is used only for final blind evaluation through the AIcrowd platform.

III. Method

A. Dataset and Task Formulation

The task consists in predicting whether a tweet originally contained a positive or negative smiley, based solely on its textual content. Tweets are provided as whitespace-tokenized sequences of words, and the problem is formulated as a binary classification task. The test set does not include labels and is used exclusively for final evaluation on the AIcrowd platform.

B. Vocabulary Construction

We build vocabularies from the training corpus using frequency-based filtering to reduce noise and control feature dimensionality. Because TF-IDF and GloVe rely on different preprocessing pipelines, the exact vocabulary construction differs between representations.

a) GloVe vocabulary.: For the embedding-based pipeline, we construct an explicit vocabulary by counting token occurrences over the full training tweets (`train_pos_full.txt` and `train_neg_full.txt`). We retain only tokens appearing at least 5 times in the corpus. This number is chosen to remove the most mispelling possible, and keep only the most relevant words. This threshold removes extremely rare words that are unlikely to yield reliable co-occurrence statistics, while keeping sufficient lexical coverage. The resulting vocabulary is indexed and stored as `vocab.pkl`.

b) TF-IDF vocabulary.: For TF-IDF models, the vocabulary is learned implicitly during vectorizer fitting on the training split. We discard tokens appearing in fewer than 2 documents (`min_df`=2) to not take into account tokens that appear only one time, and also tokens appearing in more than 95% of documents (`max_df`=0.95). This is useful to not take into account tokens that are omnipresent, and thus not interesting to work with. We also cap the vocabulary size to 300k features for computational purposes.

C. Co-occurrence Matrix Construction (GloVe)

To train GloVe embeddings, we build a word co-occurrence matrix from the training tweets. Each tweet is mapped to vocabulary indices, and we record co-occurrences for all pairs of in-vocabulary tokens within the same tweet. The resulting sparse co-occurrence matrix is stored in COO format and duplicate entries are summed to obtain global co-occurrence counts.

D. GloVe Embedding Learning

Given the word co-occurrence matrix constructed from the training corpus, we learn word embeddings using the GloVe framework.

GloVe optimizes a weighted least-squares objective on the logarithm of co-occurrence counts, allowing the model to capture global distributional statistics.

In our implementation, embeddings are trained with a fixed dimensionality of 20 using a learning rate of 10^{-3} for 10 epochs. The weighting function parameters are set to $n_{\max} = 100$ and $\alpha = 3/4$. These values are chosen to ensure stable training while maintaining reasonable computational cost. The resulting embeddings assign a fixed-length dense vector to each word in the vocabulary.

E. Tweet Representation

We consider two complementary tweet-level text representations: averaged word embeddings and TF-IDF features.

a) Averaged Word Embeddings.: Each tweet is represented by averaging the embeddings of all vocabulary words it contains. This Average Word Embedding (AWE) representation provides a simple and computationally efficient way to map variable-length text sequences to fixed-dimensional feature vectors. While this approach discards word order and syntactic structure, it serves as a strong and widely used baseline for text classification tasks and aligns with the reference solution proposed for the challenge.

b) TF-IDF Representation.: In addition to dense embeddings, we use Term Frequency–Inverse Document Frequency (TF-IDF) features to construct sparse lexical representations of tweets. TF-IDF assigns higher weights to words that are frequent within a tweet but rare across the corpus, thereby emphasizing sentiment-bearing terms while down-weighting common grammatical words. This representation is particularly well suited for sentiment analysis of short texts, where lexical cues such as negations and expressive phrases play a central role. Tweets are tokenized using whitespace tokenization without additional normalization. Words appearing fewer than two times in the corpus and extremely frequent words are discarded to reduce noise. TF-IDF features are extracted using the `TfidfVectorizer` implementation from Scikit-learn, following standard practices [3], [4], [5].

F. Classification Models

We train several supervised classifiers on the tweet-level feature representations described above. The following models are considered:

- **Logistic Regression**, which provides a linear decision boundary and serves as a strong baseline for both sparse and dense high-dimensional representations.
- **Support Vector Machines (SVM)**, which aim to maximize the margin between sentiment classes and are particularly effective in high-dimensional feature spaces.
- **Random Forests**, which introduce non-linearity through ensemble-based decision trees and allow us to assess the benefit of non-linear decision boundaries.

These classifiers are evaluated with both TF-IDF features and averaged word embedding representations in order to analyze how different inductive biases interact with sparse lexical features and dense embedding-based representations.

G. Validation Strategy and Model Selection

Model selection and hyperparameter tuning are performed using a fixed stratified train/validation split with a validation ratio of 20%. Given the large size of the dataset (approximately 2.5 million samples), this strategy provides statistically stable performance estimates while avoiding the computational cost of k-fold cross-validation. All models are evaluated on the same split to ensure fair and consistent comparison. Accuracy is used as the primary criterion for model selection, while the F1-score is reported as a complementary metric to account for class balance. The best-performing hyperparameter configurations selected on the validation split for each model are summarized in Table II.

H. Hyperparameter Optimization

Hyperparameters are selected using grid search on the validation set. For Logistic Regression and linear Support Vector Machines, we tune the regularization parameter C, which controls the strength of ℓ_2 regularization. For Random Forests, we optimize the number of trees, the maximum tree depth, the maximum number of features considered at each split, and the minimum number of samples per leaf. For the GloVe embedding-based pipeline, hyperparameter tuning is limited to the embedding dimensionality and the number of training epochs, while the remaining parameters are kept fixed to their default values.

I. Transformer-Based Extension

In addition to static word embeddings, we investigated transformer-based architectures built on BERT, a framework specifically designed for natural language processing. BERT-style models are particularly well suited to the Twitter domain, as they capture contextual semantics and long-range dependencies, which are essential for interpreting short, nuanced texts and accurately classifying their sentiment.

The transformer models trained in this project were DeBERTa and DistilBERT, which respectively correspond to an enhanced and a distilled variant of BERT. Unlike traditional embeddings such as GloVe, these models generate contextualized word representations, where the embedding of a token depends on its surrounding context. Each model relies on its own tokenizer and is trained using a validation set to monitor performance and retain the checkpoint achieving the highest validation score.

Transformer training involves a large number of hyperparameters that are internally optimized during fine-tuning. While this results in significantly better performance than linear models, it also comes with a substantially higher computational cost. As a consequence, extensive experimentation was not feasible. In particular, the DeBERTa-v3-large model could only be trained three times and was therefore primarily used to compare learning rates. In contrast, the lighter DistilBERT model allowed additional experiments on the learning rate, number of epochs, and validation set size.

Given the character constraints of tweets, the maximum token length was set to 128, which is sufficient to cover

the vast majority of samples. This value represents a good trade-off between predictive performance and tokenization efficiency.

As the performances of transformers tightly depend on the amount of data they can train on, we used a validation split of only 2% to keep the most available training data. We considered overfitting as improbable, as the validation still contained 50'000 tweets.

The exact models used were DistilBERT-base-uncased and microsoft/deberta-v3-large. The best overall performance was obtained with DeBERTa-v3-large, trained on Google Colab with a learning rate of 1e-5. Due to limited computational resources, the number of training epochs was restricted to one. To maximize efficiency, the training process was adapted to fully leverage an NVIDIA A100 GPU available in the Colab environment.

IV. Results

Model	Validation Accuracy	F1-score
GloVe + Logistic Regression	0.586	0.672
GloVe + Linear SVM	0.587	0.662
GloVe + Random Forest	0.642	0.669
TF-IDF + Logistic Regression	0.817	0.776
TF-IDF + Linear SVM	0.869	0.872
TF-IDF + Random Forest	0.742	0.786
DistilBERT	0.895	0.896
DeBERTa	**0.913**	**0.914**

TABLE I
VALIDATION PERFORMANCE (ACCURACY AND F1-SCORE) OF ALL EVALUATED MODELS.

Model	Best hyperparameters
GloVe + Logistic Regression	$C = 10^{-5}$
GloVe + Linear SVM	hinge, $C = 3 \times 10^{-4}$
GloVe + Random Forest	$n = 200$, depth=40, feat=$\sqrt{}$
TF-IDF + Logistic Regression	$C = 0.01$
TF-IDF + Linear SVM	$C = 0.1$, ngram$(1, 2)$, vocab=300k
TF-IDF + Random Forest	$n = 200$, depth=None
DistilBERT	$learning\_rate = 2e - 5$
	$num\_train\_epochs = 2$
DeBERTa	$learning\_rate = 1e - 5$

TABLE II
BEST-PERFORMING HYPERPARAMETER CONFIGURATIONS SELECTED ON THE VALIDATION SPLIT.

Table I reports the validation accuracy and F1-score of all evaluated models.

Models based on averaged GloVe embeddings exhibit poor performance across all classifiers, with validation accuracies close to 0.6. These results indicate that, in this setting, averaged static embeddings fail to capture sufficient discriminative information for sentiment classification.

In contrast, TF-IDF representations lead to a substantial performance improvement for all classifiers. In particular, the TF-IDF + LinearSVM model achieves strong validation performance, outperforming all GloVe-based approaches by a large margin. This highlights the importance of sparse lexical features and explicit word frequency information for sentiment analysis of short social media texts.

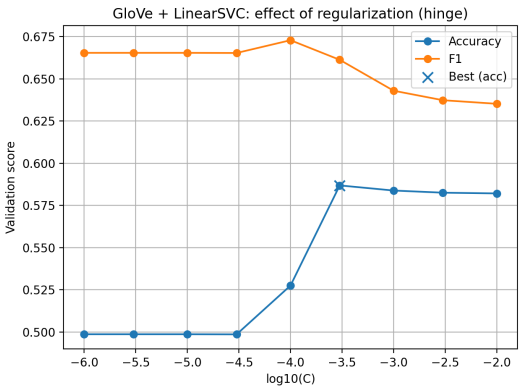

Fig. 1. Validation accuracy and F1-score of LinearSVM trained on averaged GloVe embeddings as a function of the regularization parameter C (hinge loss). Performance quickly saturates for small values of C and slightly degrades as regularization is relaxed, indicating limited linear separability in the averaged embedding space.

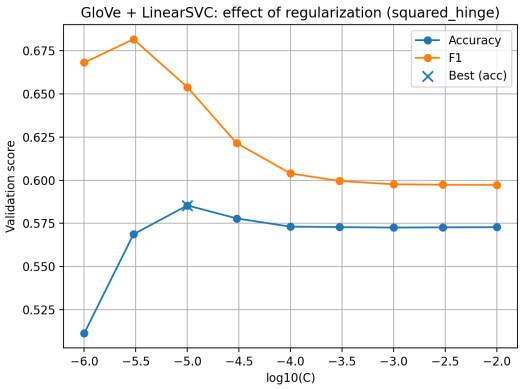

Fig. 2. Validation accuracy and F1-score of LinearSVM trained on TF-IDF features as a function of the regularization parameter C. Performance steadily improves as regularization is relaxed, highlighting the strong discriminative power of sparse lexical features for linear models on short texts.

Figures 1 and 2 illustrate the effect of the regularization parameter C for LinearSVM. While performance with GloVe embeddings saturates rapidly, TF-IDF features benefit from relaxed regularization, confirming their higher linear separability.

Transformer-based models achieve the highest validation scores. Both DistilBERT and DeBERTa significantly outperform classical approaches, with DeBERTa reaching the best overall performance. Nevertheless, the performance gap between TF-IDF + LinearSVM and transformer models remains moderate, considering the substantially higher computational cost of fine-tuning large transformer architectures.

V. Ablation Study

To better understand the contribution of individual design choices, we perform an ablation study on the **TF-IDF +**

LinearSVM model, which achieves the best validation performance among all non-transformer approaches.

The goal of this analysis is to isolate the effect of key representation parameters while keeping all other components fixed. In particular, we focus on the influence of lexical feature construction, which plays a central role in sparse text representations.

A. Ablation settings

We analyze the following factors:

- **N-gram range**: We vary the maximum length of word n-grams used in the TF-IDF representation, comparing unigrams $(1, 1)$, unigrams + bigrams $(1, 2)$, and unigrams + bigrams + trigrams $(1, 3)$.
- **Vocabulary size**: We vary the maximum number of TF-IDF features retained, using vocabulary sizes of 50k, 100k, and 300k terms.

For all experiments, the LinearSVM regularization parameter C is fixed to its optimal value selected on the validation split. All other hyperparameters are kept constant to ensure that each ablation isolates the effect of a single factor.

B. Results

Table III reports the validation accuracy and F1-score obtained for each ablation setting.

Ablation factor	Setting	Validation Accuracy / F1
N-gram range	(1,1)	0.8400 / 0.8440
N-gram range	(1,2)	0.8685 / 0.8715
N-gram range	(1,3)	0.8706 / 0.8735
Vocabulary size	50k	0.8613 / 0.8647
Vocabulary size	100k	0.8654 / 0.8685
Vocabulary size	300k	0.8685 / 0.8715

TABLE III
ABLATION STUDY ON TF-IDF + LINEARSVM. ONE FACTOR IS VARIED AT A TIME WHILE KEEPING ALL OTHER HYPERPARAMETERS FIXED.

C. Discussion

The results highlight the importance of short-range lexical patterns for sentiment classification. Moving from unigrams to bigrams yields a substantial performance gain, confirming that expressions such as negations and short phrases carry crucial sentiment information. Extending the representation to trigrams provides only marginal improvements, suggesting diminishing returns for higher-order n-grams.

Similarly, increasing the vocabulary size consistently improves performance, but with decreasing marginal benefit beyond 100k features. This indicates that most discriminative information is captured by a moderately large vocabulary, while larger feature spaces mainly increase computational cost without significant performance gains.

Overall, this ablation study demonstrates that the strong performance of TF-IDF + LinearSVM primarily stems from effective lexical feature engineering rather than increased model complexity.

VI. DISCUSSION

Our experiments show that TF-IDF representations are particularly effective for sentiment classification on short Twitter texts. Across all classifiers, TF-IDF features consistently outperform averaged GloVe embeddings, with the strongest results obtained using a LinearSVM.

Averaged GloVe embeddings discard word order and frequency information, which are crucial for sentiment detection. As a result, performance saturates quickly and shows limited separability in the averaged embedding space, even when increasing model capacity or relaxing regularization.

In contrast, TF-IDF explicitly captures term importance and short lexical patterns. The clear improvement from unigrams to bigrams confirms the importance of local expressions such as negations and common sentiment phrases, while higher-order n-grams provide only marginal gains.

Transformer-based models achieve the highest performance by leveraging contextualized representations. However, due to their high computational cost, we were only able to run a limited number of experiments. The obtained results are consistent with the literature, where transformer models are known to outperform classical approaches on sentiment analysis tasks. Despite this, TF-IDF combined with LinearSVM remains a strong and competitive baseline, offering an excellent trade-off between performance and computational efficiency.

VII. CONCLUSION

We compared averaged word embeddings, TF-IDF features, and transformer-based models for Twitter sentiment classification. Our results show that TF-IDF representations combined with a linear SVM significantly outperform averaged GloVe embeddings, highlighting the importance of sparse lexical features for short texts. Transformer models achieve the best performance by leveraging contextual representations, but at a much higher computational cost.

VIII. ETHICAL RISKS

This project involves automatic sentiment classification of public Twitter data. We identify **misclassification and downstream misuse** as the main ethical risk associated with this task.

Risk description and stakeholders. Sentiment classifiers may incorrectly label tweets, especially in the presence of sarcasm, ambiguous language, or informal expressions. Directly impacted stakeholders include Twitter users whose content could be misinterpreted, while indirect stakeholders include organizations or institutions relying on automated sentiment analysis for decision-making (e.g. content moderation, marketing analysis, or political monitoring). Incorrect predictions could lead to unfair moderation decisions, misleading analyses, or incorrect inferences about public opinion.

Severity and likelihood. The severity of this risk is moderate, as sentiment labels alone are not inherently harmful, but can become problematic if used in high-stakes contexts. The likelihood of misclassification is non-negligible, given the noisy and informal nature of Twitter language.

Risk evaluation and mitigation. We evaluated this risk by inspecting validation errors and observing systematic failure cases, such as very short tweets or tweets containing sarcasm. Due to the academic scope of the project and the benchmark nature of the dataset, we did not deploy the model in real-world settings. We explicitly limit our claims to performance on the provided dataset and avoid presenting the system as suitable for decision-critical applications.

Additional ethical considerations include dataset bias, as Twitter users are not representative of the general population, and privacy concerns. However, the dataset consists of anonymized, pre-processed tweets released for educational purposes, and no personal identification or sensitive attribute inference is performed.

Overall, while ethical risks are limited in this academic context, careful consideration is required when transferring similar models to real-world applications.

REFERENCES

[1] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[2] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations (ICLR)*, 2021.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of EMNLP*, 2002.

[5] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," in *CS224N Project Report*, 2009.