A Comparative Study of Traditional Machine Learning and Transformer-Based Models

For Twitter Emotion Analysis

Caner Tunç

Software Engineering / Mugla Sıtkı Kocman University

canertunc982@gmail.com

April 15, 2025

Contents

1 Dataset and Preprocessing

1.1 Dataset Structure and Examination

The dataset used in this study consists of texts collected from Twitter that are labeled with emotion tags (sadness, joy, love, anger, fear, surprise). The general structure of the dataset, class distribution, and basic statistics were examined.

1.2 Data Cleaning and Preprocessing

The following preprocessing steps were applied to the dataset:

- Cleaning discrete data from outliers via the IQR method
- Removal of inappropriate content, URLs, mentions, HTML tags, emojis, and special characters
- Converting text to lowercase and removing punctuation marks
- Removal of stopwords
- Application of lemmatization/stemming processes

1.3 Exploratory Data Analysis (EDA)

Exploratory data analysis was performed on the dataset to examine class imbalances. Additionally, word clouds were created to analyze the most frequently used words and text lengths.

2 Feature Extraction and Data Balancing

2.1 Feature Extraction

The following feature extraction methods were used for machine learning models:

- TF-IDF (Term Frequency-Inverse Document Frequency) vectors
- Word2Vec vectors

The datasets created with Word2Vec vectors and TF-IDF were trained separately in different machine learning models. When comparing the results, it was observed that models trained with TF-IDF generally performed better. Therefore, the metric comparison table includes machine learning models trained with TF-IDF.

The main reason TF-IDF gives better results is that it more distinctly represents the content of the text by directly taking into account word frequencies. Especially its emphasis on the importance of words within a document and across the entire dataset allows the model to learn more distinct discriminative features in some classification problems. On the other hand, although Word2Vec represents semantic similarities between words well, the context-dependent nature of the trained vectors and their sensitivity to variations specific to the dataset can limit performance in some cases.

2.2 Data Balancing

Undersampling method was used to address class imbalances in the dataset. With this method, the aim was to improve the learning performance of the models by balancing the number of examples between classes.

3 Modeling and Performance Evaluation

3.1 Traditional Machine Learning Models

The following traditional machine learning models were used in the study:

- Decision Tree
- Logistic Regression
- Random Forest
- K-Nearest Neighbors (KNN)
- Naive Bayes
- XGBoost

Basic parameters and hyperparameter optimization were applied for each model.

3.2 BERT Model

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based language model developed by Google that has produced the most successful results in the NLP field in recent years. In this study, the pre-trained version of BERT was fine-tuned for sentiment analysis.

3.3 Performance Metrics

The performance of the models was evaluated using the following metrics:

- Accuracy
- F1-Score
- Confusion Matrix
- Precision and Recall values

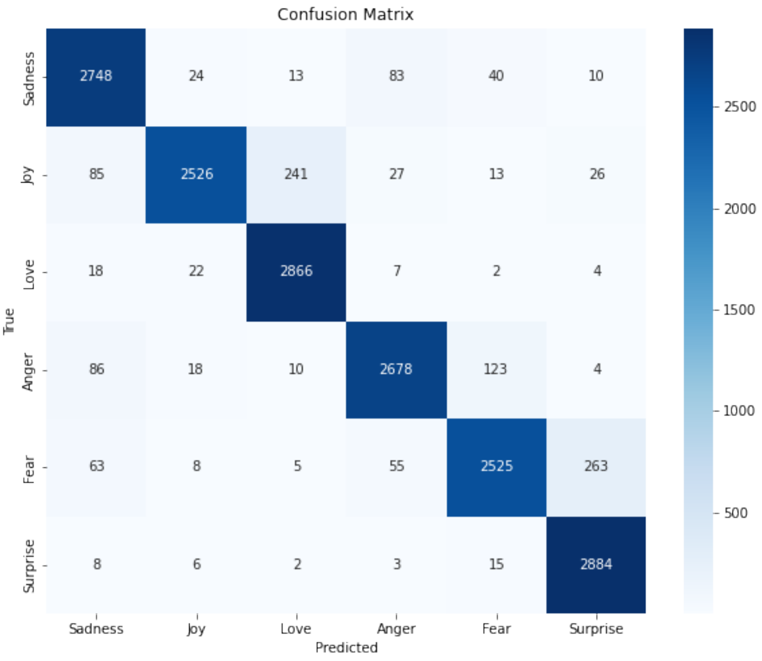
4 Comparison and Interpretation of Results

4.1 Comparison of Model Performances

Table 1: Model Performance Comparison

Model	**Accuracy**	**F1-Score**
Decision Tree	0.87	0.87
Logistic Regression	0.90	0.91
Random Forest	0.91	0.91
KNN	0.40	0.43
Naive Bayes	0.88	0.89
XGBoost	0.89	0.90
BERT	**0.93**	**0.93**

As shown in Table 1, the BERT model achieved higher accuracy and F1-score compared to other models. This result demonstrates BERT's superiority in language understanding and capturing contextual relationships.



	Sadness	Joy	Love	Anger	Fear	Surprise
Sadness	2748	24	13	83	40	10
Joy	85	2526	241	27	13	26
Love	18	22	2866	7	2	4
Anger	86	18	10	2678	123	4
Fear	63	8	5	55	2525	263
Surprise	8	6	2	3	15	2884

Figure 1: BERT Model Confusion Matrix

Diagonal values in the confusion matrix demonstrate that our model can detect the six basic emotions (sadness, joy, love, anger, fear, and surprise) with high accuracy. The success rate is particularly high in the sadness and surprise categories. However, the confusion between joy and love emotions indicates an area we could focus on in future studies to improve the distinction between these two emotions.

4.2 Strengths and Weaknesses of Models

4.2.1 Analysis of Machine Learning Models

Among the traditional machine learning models (Decision Tree, Logistic Regression, Random Forest, KNN, Naive Bayes, XGBoost), XGBoost showed the best overall performance. Notably, Random Forest and Logistic Regression achieved the highest F1-score and accuracy, both reaching 0.91. These models generally stand out with their faster training times and lower computational resource requirements. However, in models trained with Word2Vec embeddings, overfitting was observed.

Advantages:

- Faster training times
- Lower computational resource requirements
- Easy interpretability (especially Decision Tree and Logistic Regression)
- Reasonable performance even on small datasets

Disadvantages:

- Difficulty in capturing complex linguistic relationships
- Inability to understand contextual meanings of words
- High dependence on feature engineering
- Lower overall performance

4.2.2 Analysis of the BERT Model

The BERT model demonstrated superior performance compared to other models due to its strong structure in language understanding and capturing contextual relationships.

Advantages:

- Ability to understand contextual meanings of words
- Creation of richer representations through bidirectional learning
- Higher accuracy and F1-score
- Less dependence on feature engineering

Disadvantages:

- Longer training times
- High computational resource requirements (need for GPU)
- More difficult to interpret due to complex model structure
- Risk of overfitting on small datasets

4.3 Why is BERT More Successful?

The main reasons why BERT is more successful compared to other models are:

1. **Contextual Word Representations**: BERT can understand different meanings of words according to context. For example, it can distinguish between different meanings of the word "bank" in "river bank" and "financial bank" expressions.

2. **Bidirectional Learning**: BERT performs bidirectional learning by considering both the words to the left and right of a word. This allows it to create richer and more comprehensive representations.

3. **Transfer Learning**: BERT is a model that has been pre-trained on millions of texts. Thanks to this pre-training, effective results can be obtained even on small datasets.

4. **Subword Tokenization**: BERT can analyze words it has never seen before by dividing words into subword units.

4.4 Areas for Improvement

The areas for improvement in our study are:

1. **Increasing the Number of Epochs**: The model was trained with 2 epochs, but looking at the metrics and learning curve, it was concluded that this number of epochs was not sufficient for the model to fully converge. To improve the model's performance, the number of epochs should be increased (e.g., to 4 or 5 epochs).

2. **Hyperparameter Optimization**: Optimizing the hyperparameters used in the BERT model (e.g., learning rate, batch size, sequence length) with grid search or random search methods could improve performance. Although some machine learning models were trained with random search, if there had been no time constraint, all models could have been trained with grid search or random search methods in a wider range of hyperparameters to obtain better results.

3. **Comparison of Different Transformer-Based Models**: Models could be trained with different Transformer architectures such as RoBERTa, DistilBERT, XLNet other than BERT, and their performances could be compared. With these comparisons, the most suitable Transformer model could be determined.

4. **Hybrid Data Balancing Approach**: Only the undersampling method was used in this study. However, a hybrid system could be developed using oversampling methods such as SMOTE together, and the performances of these methods could be compared.

5. **Spell Checking**: Correcting spelling errors in the dataset with spell checking methods could give healthier results especially in word-based vectorization methods such as TF-IDF. However, due to time constraints, this process could not be performed.

6. **Exploring and Optimizing Word Embeddings**: Different word embedding methods could be experimented with and compared, and the Word2Vec model could be optimized more thoroughly to achieve better results.

5 Conclusion

In this study, various machine learning models and the BERT transformer model were comparatively examined for sentiment analysis on Twitter texts. The results revealed that the BERT model showed superior performance compared to other models.

The success of the BERT model stems from its bidirectional learning capability, contextual understanding capacity, and transfer learning advantages. However, BERT has disadvantages such as requiring higher computational resources and longer training times.

In future studies, to further improve the performance of the BERT model, the number of epochs can be increased, training can be done with different transformer-based models for comparison, better hyperparameter optimization for machine learning algorithms, different data balancing strategies, spell checking method can be used.