Original Research Article

# Twitter Sentiment Analysis – How Do Models Trained on A Topic-Specific Dataset of Tweets Generalize to A Second, Topic-Diverse Dataset of Tweets?

Milan Stukavec

*Park Lane International School, Uvoz 9, 118 00 Prague, Czech Republic*

## ABSTRACT

Models for sentiment analysis that are trained on one-domain dataset often result in inability to perform well in other domains due to insufficient labeled data and restricted domain knowledge. Understanding how well sentiment classifiers generalize is critical for multi-domain applications where the inability to handle the domain shift may severely impact the stakeholders. The study examines the ability of machine learning models (LinearSVC and Logistic Regression) trained on the "Social Dilemma" movie review dataset to generalize to a new, previously unseen dataset with a non-specific thematic focus. Both models were optimized using hyperparameter tuning and TF-IDF vectorizer, and then tested on the new dataset. The results showed that both models were able to achieve accuracy scores and F1-Scores between 85 and 90% on the first dataset, but when applied to the second dataset, both performance indicators dropped significantly. This was likely due to the shift in topic, vocabulary and context. The study concluded that the degree of generalization ability to an unseen dataset for sentiment analysis depends more on the degree of topic proximity between the training and new datasets than on the optimization and selection of the ML model, highlighting the importance of dataset choice in cross-domain applications.

**Keywords:** Sentiment analysis; Natural Language Processing; Cross-domain generalisation; TF-IDF vectorisation; Logistic Regression; Linear Support Vector Classifier; Twitter dataset

## INTRODUCTION

Sentiment analysis (SA) is the process of determining the emotional tone of a language and the polarity of text. It is widely used in marketing, social media, healthcare, and other fields, where it helps improve customer experience, enhance product development, or track brand reputation. It can be classified into 3 most common types: fine-grained, aspect-based, and emotion detection (1). Sentiment analysis of social media content has gained popularity in recent years given the importance of considering public opinion of the masses for better understanding of consumer feedback (2). It can also serve the education sector, for example, when researchers develop models to enable universities to identify the most discussed topics by students on their social media, including unfavorable comments that serve as suggestions for improvements (3). X, formerly Twitter, is a popular social media platform with 586 million users as of January 2025, measured by ad reach (4). This platform covers a wide range of topics and age groups withing the population; thus, it has been subject to SA

**Corresponding author:** Milan Stukavec, E-mail: stukavecmilan@gmail.com.

**Accepted** December 23, 2025
https://doi.org/10.70251/HYJR2348.413750

soon after its launch in 2006 (4).

Cross-domain sentiment generalization refers to the ability of ML models for sentiment analysis to perform consistently well on unseen data across a variety of domains. However, since most of these models are domain-specific and fail to identify slang's polarity because of a frequent use of slang language and abbreviations in messages on social media, they are not effective when applied on data from new domains (5). As a result, businesses and organizations such as policy analysis, market research, public health monitoring, or misinformation detection that need to address different customer groups and sentiments across multiple domains may face challenges in doing so if their models fail to generalize (5). Since the real-world data are heterogeneous, attempts have been made to use a multi-domain approach, but they faced major challenges with the lack of labeled datasets across multiple domains (6). Ultimately, the decision-making process of businesses and organizations may be severely impacted by the inability to generalize, limiting the understanding of customer opinions across different platforms (6).

Kolchyna et al. investigated lexicon-based methods, machine learning methods (such as SVM and Naive Bayes), and their combination on a dataset of tweets for sentiment analysis (7). It indicated the importance of incorporating emoticons, abbreviations, and social media slang phrases into lexicons for Twitter analysis (7). Similarly, Gupta et al. analyzed the performance of various machine learning models (including SVM, Naive Bayes, and Bayesian Logistic Regression) on a Twitter sentiment analysis with the use of Bag-of-Words (BoW) and Frequency Inverse Document Frequency (TF-IDF) vectorization (8). Their findings showed lower accuracy scores for Naive Bayes (0.66) and SVM (0.85) models compared to Kolchyna et al's which may have been caused by differences in data collection (e.g., via API vs. Kaggle/NLTK datasets) (7, 8). Moreover, Bayesian Logistic Regression had a greater accuracy score (0.74) than Naive Bayes, which may be attributed to differences in dataset characteristics (7, 8). Qi & Shabrina examined the sentiment of tweets related to the COVID-19 situation in England using LinearSVC, Multinomial Naive Bayes, and Random Forest models, each trained with three feature representations: BoW, TF-IDF, and Word2Vec (9). It used GridSearch for hyperparameter tuning, removed hashtags and usernames, in line with my pre-processing of data (9). The results showed that LinearSVC with either BOW or TF-IDF performed best among other conditions (accuracy score of 0.71), while models trained with Word2Vec features showed a significantly lower performance (accuracy ranging from 0.43 to 0.56) (9). The study supports the use of the TF-IDF vectorizer for word encoding (9). Since the dataset was smaler than mine (1000 tweets per sentiment), it can be argued that TF-IDF performs better on smaller datasets compared to Word2Vec (9). A. Prabhat and V. Khullar's study compared Naive Bayes and Logistic Regression for tweet sentiment analysis in terms of accuracy, precision, and throughput (8). The results showed that Logistic Regression had better accuracy value (0.68) than Naive Bayes (0.67) (10).

Text analysis is the process of finding patterns in text and semantic relationships between words, sentences, or sections of texts. It falls under machine learning, as its models are trained to detect and compare patterns in the data provided. However, this poses many challenges, as text cannot be read naturally by computers, and finding semantic relationships between binary representations of text is difficult with small data sets (11). Recent advances in text analysis are mainly due to Large Language Models (LLMs) that differ from classic supervised and unsupervised branches of ML. LLMs use more complex neutral network algorithms with multiple layers, nodes, and hyperparameters; train on much larger datasets; and can be self-learned by automatically labeling data. Supervised models, unlike unsupervised ones, require labeling, the process of assigning a ground-truth value to a given entry in the dataset. During training, the algorithm is fed data and truth values, which are then compared to each other when testing the data (12). This method generally yields higher performance on a dataset. However, it may be prone to overfitting, the idea of the model not performing well on a new set of data.

Classification is one of the two main types of supervised learning, along with regression. It is applied to problems with defined labels and categorical variables of an output. It predicts group membership for data instances (13). The output variable (y) can take on only a discrete predefined list, and the target variable is one of the possible categories or classes. Many classification models work on the principle of regularization, which balances model complexity against classification error to prevent overfitting.

Support Vector Machine (SVM) is a type of classification ML model whose objective is to find an optimal line, hyperplane, that provides the maximum separation between classes. The closest data points of each class that define the margin are called the support vectors. Although SVMs are known for

binary classification, they can also handle multi-class classification by using One-to-One (OvO) or One-vs-All (OvA) approaches (14). OvA creates a binary SVM model for each possible pair of classes and is best used if the number of classes is small, making it a suitable approach for my experiment—positive vs. neutral, positive vs. negative, and negative vs. neutral. LinearSVC is a linear SVM classifier optimized for speed and high-dimensional data (15). It uses the L2-regularized hinge loss (Eq. 1) that combines a hinge loss term and an L2 regularization term (16). It penalizes both margin violations and large model weights, balancing classification accuracy with a large margin.

$$l_h(\boldsymbol{w};(f,l)) = \max\left(0,\ 1 + \max_{l' \neq l} \sum_{f \in \mathcal{F}} w_{(f,l')} - \sum_{f \in \mathcal{F}} w_{(f,l)}\right) \quad \text{(Eq. 1)}$$

Logistic regression (LR) is the second most widely used algorithm for classification (17). Its output variable always lies between 0 and 1, where the threshold value sets the decision boundary. The number of classes determines the function used to calculate the likelihood of a prediction. Binary classification uses the sigmoid function. However, with multi-class classification, the sigmoid function is replaced by the softmax function (Eq. 2), which takes a vector of all the scores and outputs probabilities for each of the classes that together always add up to 1. The behavior of these functions explains how LR makes predictions. A strong negative input score leads to a value close to zero, thus to a negative class with high confidence. Conversely, strong positive input pushes the function toward a positive class with a high probability. Logistic regression uses the log-loss function (Eq. 3), which evaluates the predicted probabilities against the true class labels (y) (18). To maximize the log-likelihood, meaning to achieve regularization, the log-loss function must be minimized. When the true value is 1 and the model predicts the opposite (a value close to 0), then the log-loss function approaches infinity, and vice versa, thus penalizing the most abnormal predictions (18).

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{(Eq. 2)}$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)} + \left(1 - y^{(i)}\right)\log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right] \quad \text{(Eq. 3)}$$

This paper presents Classic NLP-based SA that uses ML models with feature extraction to determine the sentiment (19) and to compare the performance of Logistic Regression and Support Vector Machine (LinearSVC) for sentiment analysis on two datasets of tweets. Specifically, the study evaluates which model generalizes better to a second dataset of tweets on different topics after being trained and optimized on Dataset One of Social Dilemma movie reviews.

This study aims to provide empirical evidence on the effects of hyperparameter tuning of sentiment classifiers by analyzing whether the challenges associated with domain-specific training dataset configuration can be mitigated through model optimization. This contribution is intended to inform stakeholders by clarifying whether a dataset selection is the most critical factor for model performance or whether alternative optimization strategies are a valid approach to increase model performance.

## METHODS AND MATERIALS

### Dataset

**Table 1.** Dataset identifiers, official dataset names, and source links used in this study.

| Alias | Official Name | Link |
|---|---|---|
| Dataset One | The Social Dilemma Tweets - Text Classification | kaggle.com/datasets/kaushiksuresh147/the-social-dilemma-tweets |
| Dataset Two | Twitter Tweets Sentiment Dataset | kaggle.com/datasets/yasserh/twitter-tweets-sentiment-dataset |

Dataset One contains 20,068 records (Appendix I). It consists of tweets labelled with the token "#TheSocialDilemma." Each entry (row) has 14 columns (12 string columns, 1 boolean column, and 1 integer column), and the only unique sentiment values are "Positive", "Neutral", and "Negative". For the purpose of this study, only the "text" (x value) and "Sentiment" (y value) columns are needed. The dataset was extracted using TwitterAPI after September 9th, 2020, when the movie was published. It is limited only to the hashtags; therefore, it can contain tweets regardless of geographic origin. Table 2 shows a sample of 3 selected rows that are most relevant for SA.

Dataset Two, used to evaluate both models, contains tweets with no particular focus on a certain topic. Rather, it is a selection of 27.5 thousand tweets labelled according to their sentiment. The dataset consists of 4 columns, and a sample selection is displayed in Table 3. Several tweets in the dataset end with trailing ellipses

**Table 2.** Representative sample of three tweets from Dataset One ("The Social Dilemma" dataset), showing tweet text, associated hashtags, and sentiment labels.

| Text | Hashtags | Sentiment |
|---|---|---|
| The problem of me being on my phone most the time while trying to watch #TheSocialDilemma 🤦‍♀️ | TheSocialDilemma | Positive |
| Recommended some watching #TheSocialDilemma on Netflix. #takemoretimeout #losethephone | TheSocialDilemma, takemoretimeout, losethephone | Neutral |
| After watching #TheSocialDilemma whenever I go on IG what pops up definitely makes mad sense smfh | TheSocialDilemma | Negative |

**Table 3.** Representative sample of tweets from Dataset Two, illustrating the dataset's topic diversity and sentiment labels.

| Textid | Text | Selected_Text | Sentiment |
|---|---|---|---|
| 6e0c6d75b1 | 2am feedings for the baby are fun when he is all smiles and coos | fun | positive |
| cb774db0d1 | I`d have responded, if I were going | I`d have responded, if I were going | neutral |
| 549e992a42 | Sooo SAD I will miss you here in San Diego!!! | Sooo SAD | negative |

("…"), which may indicate unfinished text that did not fit into the dataset in its entirety.

Both datasets are imbalanced in terms of their sentiments. In Dataset One, Positive class accounts for 47% of total tweets, Neutral is 35%, and Negative equals 18%. In Dataset Two, Positive is 32%, Neutral 40%, and Negative 28%.

Figure 1 shows a flow chart that represents anchor steps in the experiment. The goal is to find the extent of the Optimized Model's ability to generalize to Dataset Two. It applies to both machine learning algorithms used in this experiment, and Table 5 further breaks down these steps into six conditions (see *Results & Discussion*).

## Pre-Processing of Data

Dataset Two was left unprocessed since its role in this investigation is to provide an evaluative tool for the models to test if they can generalize to raw, unseen data. This design increases the ecological validity of this experiment, as it mimics a real-world environment.

In contrast, Dataset One was preprocessed and cleaned prior to training because its purpose was to serve as the learning base for the classification models. Standard text normalization steps, such as token removal and text simplification in SA significantly improve the performance of ML classifiers (20). Therefore, a dimensionality reduction in Dataset One to keep only "text" and "Sentiment" columns was performed. Therefore, all "tags" (words starting with "@") and

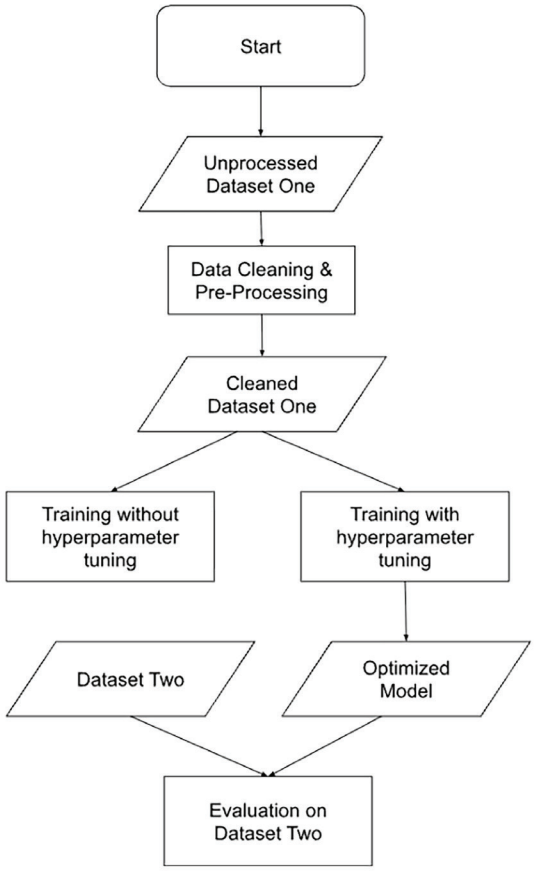


**Figure 1.** Overview of the experimental workflow, illustrating data preprocessing, model training, hyperparameter tuning, and evaluation on a topic-diverse dataset.

hashtag names (words starting with '#') were removed from Dataset One. Since Dataset One contains only three unique values for Sentiment, a choice was made to map these values to 1 for "Positive", 0 for "Neutral", and -1 for "Negative" without using any vectorization or encoding technique. By default, the dataset was split in an 80/20 ratio for training and testing.

Vectorization approaches use two types of vectorizations encoding: one-hot encoding and BoW encoding. One of the main limitations of the BOW technique is that it does not capture the similarity between different words with the same meaning (21). On the other hand, a bag of n-grams splits a sequence of words into n tokens (n consecutive words) that allow it to capture more complex relationships and word order. For the purpose of SA and given the maximum allowed character count (l40) of a tweet, the primary objective is to identify words (low n-grams of words) that most constitute the sentiment. Therefore, the TF-IDF word encoding vectorizer was used, as it quantifies the importance of a word relative to other words in the text corpus. This approach is used in both Kolchyna et al. and Qi & Shabrina studies to account for conventional phrases in text that may carry necessary meaning to the sentiment.

**Hyperparameter Tuning**

Before testing, it is important to find a suitable set of hyperparameters that maximize the performance of a model. This technique is called hyperparameter tuning and involves algorithms that test different combinations of hyperparameters and then compare the individual attempts to identify the best result. GridSearch is a well-known method that tries all combinations of the provided hyperparameters to find the best conditions for the estimator. It is usually followed with cross-validation, with k-fold validation being very popular, which divides the data into k folds with 1 test set and k-1 training sets, iterates over each combination, and averages the results as a cross-validation model. In Scikit-learn, both LinearSVC and Logistic Regression models have predefined default parameters that can be adjusted depending on the expected result, data file type, and model type.

Table 4 shows a final set of hyperparameters tested via GridSearchCV and the most optimal values. When using GridSearch with 5-fold stratified cross validation, the TF-IDF vectorizer must be placed inside a Pipeline, a Scikit-learn utility which ensures that each fold is vectorized independently in order to prevent data leakage.

A fixed random_state with the random seed value of "42" was used to ensure deterministic reproducibility of the experiments, which led to identical outcomes yielded by repeated runs. In addition, stratification was used consistently at all stages. However, performance variability across different random seeds was not assessed.

The C parameter controls regularization strength by penalizing misclassified points, where C and regularization strength have a mutually dependent inverse relationship. Class_weight is used to automatically adjust weights of classes to reflect their proportional frequency because of the imbalanced datasets. The final set of C values was determined by plotting a validation curve (see *Results & Discussion*). Solver specifies the optimization algorithm, and since this is a multiclass classification, Scikit-learn recommends using either 'lbfgs' or 'saga'.

For the TF-IDF vectorizer, the hyperparameter

**Table 4.** Hyperparameters tested using GridSearchCV for LinearSVC, Logistic Regression, and the TF-IDF vectorizer, along with the selected optimal values.

| Model | Hyperparameter | Values Tested in GridSearchCV | Optimal Value |
|---|---|---|---|
| LinearSVC | C | [1.5, 1.6, 1.7] | 1.6 |
| | class_weight | "balanced" | balanced |
| | random_state | Same fixed integer used in experiments | 42 |
| Logistic Regression | C | [15, 20, 25] | 20 |
| | class_weight | "balanced" | balanced |
| | random_state | Same fixed integer used in experiments | 42 |
| | solver | lbfgs, saga | lbfgs |
| TF-IDF Vectorizer | max_df | [0.25, 0.5, 0.75] | 0.25 |

max_df is tuned to determine what percentage of terms is kept after removing frequently recurring terms. Out of the total tested values, 0.25 provided the best results, suggesting that the sentiment of a tweet is likely determined with less frequent and unique words that convey the meaning. After the tuning, the TF-IDF matrix has 15,051 tweets and 20,586 unique words (each tweet is represented as a vector of 20,586 numbers). Because tweets contain very few characters, each tweet includes only a small fraction of the 20,586 possible words, so most values in its vector are '0'.

Regarding ethics, this investigation uses publicly available datasets that contain user-generated text that was originally posted on the social platform X. However, there were no attempts to identify users, nor to share the analysis with third parties. The role of ethics can be a limitation of sentiment analysis if data is manipulated without its author's permission.

## RESULTS AND DISCUSSION

There are several metrics and evaluation tools to evaluate the ML models. This study uses accuracy (Eq, 4), precision (Eq. 5), recall (Eq. 6), confusion matrix, and the F1-Score (Eq. 7) to assess the models. They all depend on the idea of correctness, where each prediction falls into one of four possible categories: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Based on these categories, the evaluation metrics can be calculated. Accuracy is simply a fraction of correct predictions, and it is an easy way to measure performance. Precision measures how accurate the positive predictions are. Recall, on the other hand, measures the fraction of positives found in a particular set.

The F1-Score is a harmonic mean of the latter two metrics, incorporating them together. However, there is a trade-off between precision and recall: favoring precision may harm recall, as the pursuit to make positive predictions to avoid false positives may make the model miss TPs, and vice versa.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad \text{(Eq. 4)}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{(Eq. 5)}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{(Eq. 6)}$$

$$\text{F1} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad \text{(Eq. 7)}$$

Table 5 summarizes the six conditions and their names for better analysis of the results. Each model was first trained with no set parameters to serve as a control group. There were two experimental groups: one with hyperparameter tuning for Dataset One and one for Dataset Two that has been trained on the models after hyperparameter tuning.

To identify the optimal C parameter for both models, validation curve was plotted, and the range of values was iteratively narrowed to approximate the vertex. Next, three most plausible values were tested via GridSearch to find the final parameter. This approach was chosen because of the limited computational capacity of my laptop, which did not allow for testing tens of possible values.

The final validation curve for Condition Four can be seen in Figure 2 with the narrowed-down domain ($10 \leq C \leq 25$) and vertex around C=18, supporting the optimal value of C of 20. However, this value implies high complexity of the model that excessively tried to fit the training data of Dataset One. This could cause overfitting and a poor performance on Dataset Two (Condition Six) as the model was unable to generalize well to the new dataset, leading to an accuracy score of 0.54 and average F1-Score of 0.53 (Table 6). Moreover, it could result in the bias-variance trade-off, a situation when the model learns noise, which prevents it from finding patterns and relationships in data outside the dataset.

Table 6 shows the overall results for all six conditions, measured by the four evaluation metrics described above.

**Table 5.** Definition of the six experimental conditions used to evaluate model performance, including baseline training, hyperparameter tuning, and cross-dataset evaluation.

| | **No hyperparameter tuning (Dataset One)** | **Hyperparameter tuning (Grid Search with 5 CV on Dataset One)** | **Evaluation on Dataset Two** |
|---|---|---|---|
| LinearSVC | Condition One | Condition Two | Condition Five |
| Logistic Regression | Condition Three | Condition Four | Condition Six |

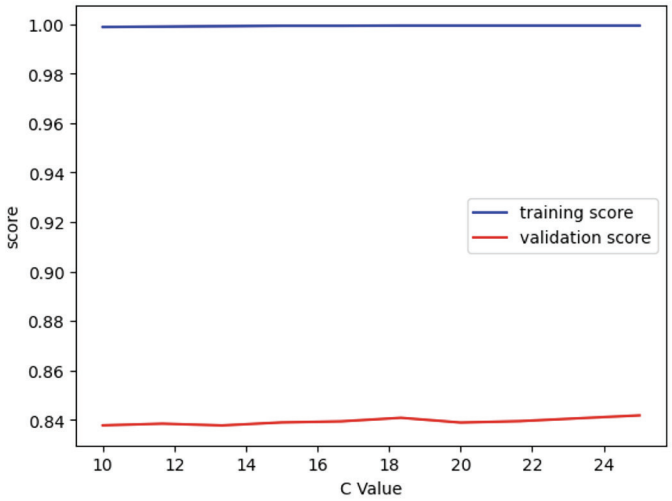


**Figure 2.** Validation curve for Logistic Regression (Condition Four), showing model performance across different values of the regularization parameter C and supporting the selected optimal value.

All numbers are rounded to two decimal places; the accuracy score and F1-Score are considered the primary performance indicators, as the latter one combines both precision and recall. All results are reported as point estimates from single single, deterministic runs using a fixed random seed; performance variability and confidence intervals were not assessed.

Although lower than LinearSVC, Logistic Regression's performance outperformed ML models in a couple of referential studies (see *Introduction & Literature Review*). To decide between Logistic Regression and Naive Bayes as a second model to LinearSVC, a dry-run test of Naive Bayes model trained on Dataset One without hyperparameter tuning was conducted that revealed a lower accuracy score (0.84) and F1-Score (0.82) compared to both Condition One and Condition Three, making Logistic Regression a preferred choice over a more common Naive Bayes model for sentiment analysis.

For LinearSVC, the hyperparameter tuning resulted

**Table 6.** Performance metrics for all six experimental conditions, including accuracy, average F1-score, and class-specific precision, recall, and F1-scores.

| Condition | Accuracy | Average F1-Score | Class | Precision | Recall | F1-Score |
|---|---|---|---|---|---|---|
| Condition One | 0.90 | 0.88 | −1 | 0.89 | 0.74 | 0.81 |
| | | | 0 | 0.89 | 0.94 | 0.91 |
| | | | 1 | 0.9 | 0.92 | 0.91 |
| Condition Two | 0.87 | 0.85 | −1 | 0.86 | 0.7 | 0.77 |
| | | | 0 | 0.87 | 0.89 | 0.88 |
| | | | 1 | 0.87 | 0.91 | 0.89 |
| Condition Three | 0.86 | 0.83 | −1 | 0.91 | 0.6 | 0.72 |
| | | | 0 | 0.84 | 0.91 | 0.87 |
| | | | 1 | 0.86 | 0.92 | 0.89 |
| Condition Four | 0.86 | 0.84 | −1 | 0.83 | 0.71 | 0.77 |
| | | | 0 | 0.85 | 0.9 | 0.87 |
| | | | 1 | 0.89 | 0.9 | 0.89 |
| Condition Five | 0.53 | 0.52 | -1 | 0.58 | 0.36 | 0.44 |
| | | | 0 | 0.56 | 0.49 | 0.52 |
| | | | 1 | 0.50 | 0.75 | 0.60 |
| Condition Six | 0.54 | 0.53 | -1 | 0.59 | 0.37 | 0.45 |
| | | | 0 | 0.55 | 0.51 | 0.53 |
| | | | 1 | 0.52 | 0.74 | 0.61 |

in worsened accuracy and average F1-Score compared to Condition One. Although there was only a slight decrease between accuracy and F1-Score (-3.33% and -3.41%, respectively), the hyperparameter tuning through GridSearch did not improve the model. Hyperparameter tuning had no effect on accuracy but added one basis point to the average F1-Score for Logistic Regression, although it came entirely from an improvement in the negative class (+6.94%).

The precision/recall trade-off curves for Condition Two and Condition Four are shown in Figure 3 and Figure 4, respectively. Considering both curves and the results in Table 6, this trade-off is supported, as maximizing recall leads to poor precision and vice versa. With a few exceptions, both models prioritized precision over recall (precision > recall). This means that both models attempted to minimize the number of false positives (as few incorrectly classified sentiments as possible), which could have led to some false negatives being omitted.

In fact, this claim is supported by the confusion matrices for all conditions. Figure 5 shows a confusion matrix for Condition Two, where the diagonal line with cells of most color intensity depicts true positiveness for a particular class, while the other two cells in a row are false negatives for that class. The FNs for each class (264 for -1, 181 for 0, and 247 for 1) illustrate the precision–recall trade-off and the model's stricter classification boundary. The negative class's FNs in terms of their percentage representation are higher than in other classes, resulting in a significantly lower recall score (0.7) for this class compared to the other two (0.89 and 0.91). This phenomenon can also be seen in other conditions in Table 5.

However, Condition Five and Condition Six show a substantial decline in all metrics with a larger spread between class scores. This suggests that Dataset Two features dissimilar characteristics to Dataset One as both

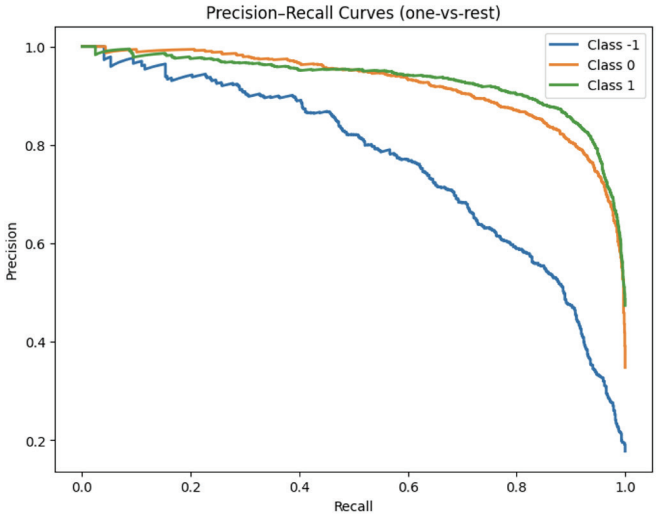


**Figure 4.** Precision–recall trade-off curve for Logistic Regression under Condition Four, highlighting class-specific differences in prediction behavior.

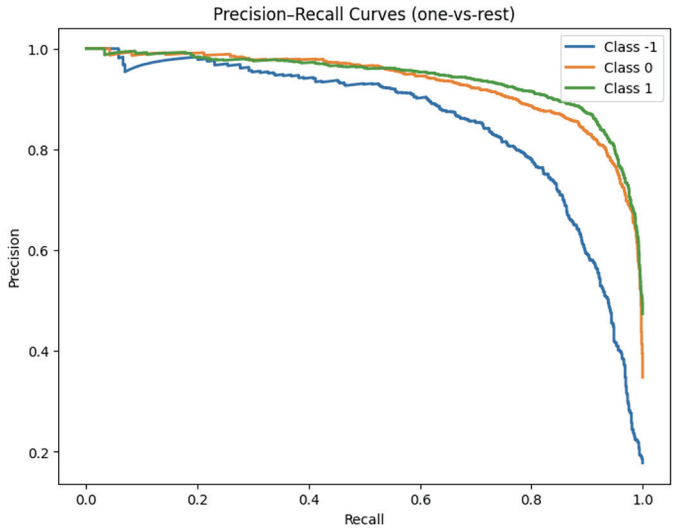


**Figure 3.** Precision–recall trade-off curve for LinearSVC under Condition Two, illustrating the balance between precision and recall across sentiment classes.

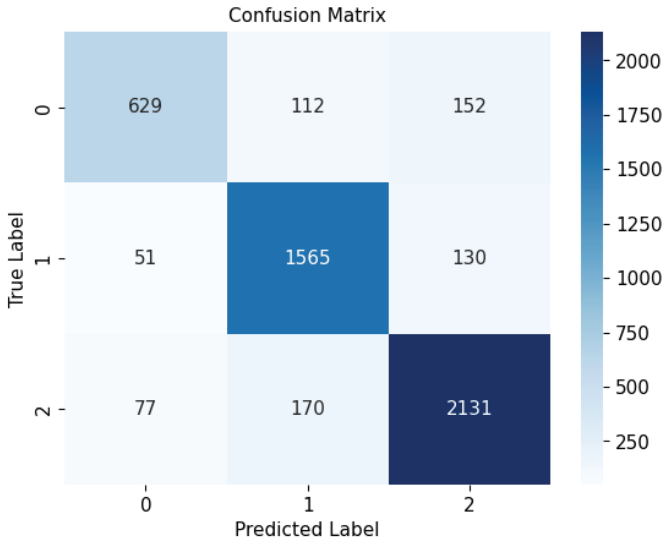


**Figure 5.** Confusion matrix for LinearSVC under Condition Two, visualizing true and false predictions across sentiment classes for Dataset One.

models were unable to categorize sentiments of tweets that concerned different topics than movies, lacking the essential vocabulary. The reduced model robustness led to almost identical values of the performance indicators under both conditions, leading to a conclusion that good performance on Dataset One does not necessarily guarantee a similar outcome for different conditions, regardless of the type of model. In fact, Logistic Regression performed slightly better than LinearSVC despite the opposite trend in all previous conditions. Figure 6 shows a confusion matrix for Condition Five, which contrasts significantly with the confusion matrix representing the same model's (LinearSVC) performance on Dataset One. The FNs for each class (5015 for -1, 5677 for 0, and 2152 for 1) reveal the limitations of single-domain training dataset. Since more negative tweets are misclassified than correctly classified and the model tends to label neutral tweets as positive more often than as negative, it can be concluded that the classifier has a positive bias. This can be attributed to slang words and abbreviations used in online messages, mild positive language in neutrally and negatively labeled tweets, and lack of essential vocabulary to capture diverse range of topics in Dataset Two. Compared to Figure 8, the performance on Dataset Two is uneven across sentiment classes, limiting its effectiveness in multi-domain applications.

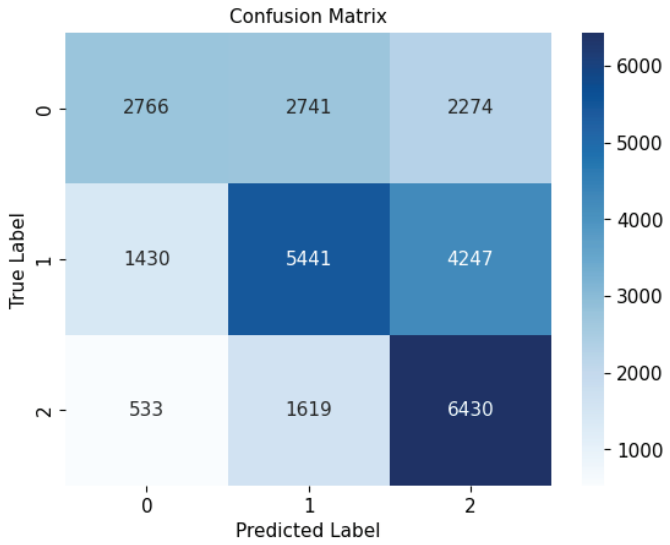


**Figure 6.** Confusion matrix for LinearSVC under Condition Five, demonstrating reduced classification performance and class imbalance effects when evaluated on Dataset Two.

To further analyze the performance of the models in Condition Two and Condition Four, false positive (FP) results were examined to identify possible reasons for misclassification. FPs occur when the predicted value differs from the actual value (e.g., the model incorrectly predicts 1, and the actual value is either -1 or 0). For example, for positive sentiment (class 1), there were 318 FPs in Condition Two and 283 FPs in Condition Four. Table 7 shows the percentage of similarity between FPs for all three classes in both conditions. Given that the similarity score is high and both models had difficulty with the same tweets, conclusions can be drawn about the ML technique and the data. Both linear classifiers tended to misclassify the same "borderline" cases, probably due to noisy or inconsistent labels, class imbalance, ambiguous sentiment, very short or long texts, and lack of vocabulary. A noisy dataset may include sarcastic tweets, tweets containing negations, and slang words.

**Table 7.** Percentage overlap of false positive predictions between LinearSVC and Logistic Regression across sentiment classes, indicating shared misclassification patterns.

| Class | Percentage of Common False Positives |
|---|---|
| 1 | 80.13% |
| 0 | 94.96% |
| -1 | 88.46% |

Table 8 shows a sample of common FNs for positive sentiment (out of a total of 255) with their actual labels. Both LinearSVC and Logistic Regression assigned them a value of 1. This random selection of tweets reveals the limitations of Dataset One and both models. First, the trailing ellipses in the tweets may have introduced ambiguity during training, causing the models to assign these tweets to the positive class because the real sentiment was not part of the dataset. This is further supported by the actual labels, which can be questioned, especially for the negative class. Another explanation is the mislabeling of the dataset. For example, the last tweet in Table 8 does not necessarily show signs of negative sentiment. Since it appears to be in full length, the quality of Dataset One is debatable.

## CONCLUSION

Given that the models behaved similarly on both datasets, i.e., the performance metrics were similar and

**Table 8.** Sample of tweets misclassified by both LinearSVC and Logistic Regression, along with their true sentiment labels, illustrating sources of ambiguity and labeling challenges.

| Tweet | True Label |
|---|---|
| I'm watching and I might go off the grid for some time. I knew Facebook was messed up but not THIS MUCH | 0 |
| Good morning  I watched and it's worth a watch on  it's soooo devastating but necessary. | 0 |
| It's genius level to manipulate people by showing them documentary about how they are being manipulated!… https://t.co/RUltqhW9xG | 0 |
| How does it CONTROL OVER us?!<br>Would that be turned to something that could truly help unifying the world, in a good… https://t.co/xTSxOS19au | 0 |
| Everyone should watch like NOW. Seriously. Then maybe consider deleting some of your social media… https://t.co/71lO6fv5FX | -1 |
| Social media has no moral compass. No true understanding of good and bad. | -1 |
| What killed >200K Americans from COVID19? SOCIAL MEDIA PLATFORMS! You killed these people … https://t.co/aU8XpkIE6y | -1 |
| I watched and it made me realize I'm not popular enough for a social media addiction | -1 |

Note: Emojis were present in the original tweets but are omitted here for formatting compatibility.

both models failed to generalize to Dataset Two, the results support the main conclusion of this experiment: the degree of generalization ability to an unseen dataset for sentiment analysis depends more on the degree of topic proximity between the training and new datasets than on the optimization and selection of the ML model.

In Amolik et al., the researchers claimed that they could improve accuracy by increasing the size of the dataset. These findings could be applied to my study as well, whose moderate dataset size is both a limitation and a subject for further improvement (22). Qi & Shabrina support my findings, as their ML model's performance was significantly lower on a dataset of tweets with various topics, showing that a model trained on a niche dataset can lead to overfitting (7). Interestingly, a higher model's complexity in Condition Four compared to Condition Two (measured by the degree of the C parameter) had a negligible effect on the overall model's performance on the unseen data (7). On the other hand, in Kolchyn et al., the F-Score significantly increased after feature engineering and hyperparameter tuning (5). All these findings underscore the significance of the choice of a training dataset.

Although the topic closeness between the datasets plays a crucial role, Table 8 reveals a deeper causative truth about the importance of high-quality training data. As such, the incomplete tweets and wrong labelling in Dataset One could worsen the effects of hyperparameter tuning and lead to a rarely high value of the C parameter for Logistic Regression, as it could try to fit the flawed training data perfectly. This finding confirms the importance of data sets not only in terms of their semantic and lexical diversity but also in terms of data quality.

Due to the limited computing capacity of the equipment used in this experiment, one of the weaknesses of this study is the small number of hyperparameters tested using GridSearch and the long execution time, which leads to fewer attempts at model optimization. Although this could improve model performance, the expected impact is uncertain.

To improve the reliability and validity of this experiment, the lexicon-based approach could be used as a control group for a better analysis of the results, according to the referential studies. From the results, it is evident that machine learning for sentiment analysis has applicability in NLP for most industries but requires a conscious and deliberate understanding of the purpose and content of topics that are to be before the training. If tailored ML models for most types of sentiment analysis were to exist, the identification of an emotional tone of text would likely experience an improvement and deployment for various business purposes, such as review management and understanding customer feedback.

## FUNDING SOURCES

The author did not receive any funding for the conduct of the research and preparation of the article.

## CONFLICT OF INTEREST

The author declares no conflicts of interest related to this work.

## REFERENCES


1. Saju B, Jose S, Antony A. Comprehensive study on sentiment analysis: Types, approaches, recent applications, tools and APIs. In: 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA). IEEE; 2020; p. 186–93. doi:10.1109/ACCTHPA49271.2020.9213209.
2. Poecze F, Ebster C, Strauss C. Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts. *Procedia Comput Sci*. 2018; 130: 660–6. doi:10.1016/j.procs.2018.04.117.
3. Mehmood S, Ahmad I, Khan M, Khan F, Whangbo T. Sentiment Analysis in Social Media for Competitive Environment Using Content Analysis. *Comput Mater Continua*. 2022; 71: 5603–18. doi:10.32604/cmc.2022.023785.
4. DataReportal. Essential X stats. Available from: https://datareportal.com/essential-x-stats (accessed on 2025-10-23).
5. Pervaiz K, Azam M, Nasim F, Noor S, Ayub K. Cross-domain sentiment analysis: a multi-task learning approach with shared representations. *J Comput Biomed Inform*. 2024; 7 (2). doi:10.56979/702/2024.
6. Abdullah NA, Feizollah A, Sulaiman A, Anuar NB. Challenges and recommended solutions in multi-source and multi-domain sentiment analysis. *IEEE Access*. 2019; 7: 144957–144971. doi:10.1109/ACCESS.2019.2944420.
7. Kolchyna O, Souza TTP, Treleaven P, Aste T. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. arXiv. 2015. Available from: https://arxiv.org/abs/1507.00955.
8. Gupta V, Joshi N, Katre N. Study of Twitter sentiment analysis using machine learning algorithms on Python. *Int J Comput Appl*. 2017; 165 (9): 14–7. doi:10.5120/ijca2017914022.
9. Qi J, Shabrina N. Sentiment analysis using Twitter data: A comparative application of lexicon- and machine-learning-based approach. *Soc Netw Anal Min*. 2023; 13 (1): 30. doi:10.1007/s13278-023-01030-x.
10. Prabhat A, Khullar V. Sentiment classification on big data using Naïve Bayes and Logistic Regression. In: Proceedings of the 2017 International Conference on Computer, Communication and Computational Intelligence (ICCCI); 2017; p. 1–5. doi:10.1109/ICCCI.2017.8117734.
11. Ittoo A, Nguyen LM, van den Bosch A. Text analytics in industry: Challenges, desiderata and trends. *Comput Ind*. 2016; 78: 96–107. doi:10.1016/j.compind.2015.12.001.
12. Cunningham P, Cord M, Delany SJ. Supervised Learning. In: Cord M, Cunningham P, editors. Machine Learning Techniques for Multimedia. Berlin: Springer; 2008. doi:10.1007/978-3-540-75171-7_2.
13. Soofi AA, Awan A. Classification Techniques in Machine Learning: Applications and Issues. *J Basic Appl Sci*. 2017; 13: 459–65. https://doi.org/10.6000/1927-5129.2017.13.76
14. Suthaharan S. Support Vector Machine. In: Machine Learning Models and Algorithms for Big Data Classification. Vol. 36. Boston: Springer; 2016. doi:10.1007/978-1-4899-7641-3_9.
15. sklearn.svm.LinearSVC (Class documentation). Available from: https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html (accessed 2025-09-15).
16. Moore R, DeNero J. L1 and L2 regularization for multiclass hinge loss models. In: Proceedings of the Machine Learning in Speech and Language Processing (MLSLP 2011); 2011; p. 1–5.
17. 7 machine learning algorithms to know: A beginner's guide. Available from: https://www.coursera.org/gb/articles/machine-learning-algorithms (accessed 2025-10-21).
18. sklearn.linear_model.LogisticRegression (Class documentation). Available from: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed 2025-09-15).
19. Kanakaraj M, Guddeti RMR. NLP based sentiment analysis on Twitter data using ensemble classifiers. In: 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN). IEEE; 2015; p. 1–5. doi:10.1109/ICSCN.2015.7219856.
20. Palomino MA, Aider F. Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis. *Appl Sci*. 2022; 12 (17): 8765. doi:10.3390/app12178765.
21. Sineglazov V, Savenko I. Comparative Analysis of Text Vectorization Methods. *Electronics Control Syst*. 2023; 2 (76): 21–7. https://doi.org/10.18372/1990-5548.76.17663
22. Amolik A, Jivane N, Bhandari M, Venkatesan M. Twitter sentiment analysis of movie reviews using machine learning techniques. *Int J Eng Technol*. 2016; 7: 2038–44.

**Appendix I:** Sample from Dataset One

| user_name | user_location | user_description | user_created | user_ followers | user_ friends | user_ favourites | user_ verified |
|---|---|---|---|---|---|---|---|
| Stephanie | Canada | | 2017-02-15 14:48:31 | 61 | 251 | 1823 | FALSE |
| JBernard | Portland, OR | | 2011-09-06 23:31:59 | 92 | 1198 | 8070 | FALSE |
| Stephanie | Canada | | 2017-02-15 14:48:31 | 61 | 251 | 1823 | FALSE |
| ☆ | Rubyy / dhoni | Reality sucks . | 2020-06-08 9:00:07 | 262 | 280 | 342 | FALSE |
| A RICH | Dublin, Ireland | Techno Dj. | 2010-10-26 14:39:08 | 2481 | 429 | 8619 | FALSE |
| Barbara Chesire-Chabbaga | Nairobi, Kenya | BeeComing | 2010-12-04 20:16:44 | 858 | 616 | 2290 | FALSE |
| Ari | Texas, USA | Mechanical Engineer • Bolivian | 2009-08-04 0:09:20 | 518 | 459 | 13302 | FALSE |
| Ali Hayati | Kuwait | Peanut Butter addict. قفـانك enthusiast. I make sounds. instagram: ali_hayati | 2011-05-05 9:58:36 | 1017 | 358 | 1328 | FALSE |
| good~tamez rishtedar 🤦 | | baati ko uska Diya mil gaya @ maindiyatubaati | 2019-05-28 11:57:47 | 1214 | 1227 | 8058 | FALSE |

**Continued Appendix I:** Sample from Dataset One

| date | text | hashtags | source | is_ retweet | Sentiment |
|---|---|---|---|---|---|
| 2020-09-16 20:34:40 | I can't say I'm surprised and some of these things I am conscious of but I'm so turned off #TheSocialDilemma | TheSocialDilemma | Twitter for iPhone | FALSE | Positive |
| 2020-09-16 20:33:12 | @daithaigilbert @slpng_giants We must press our elected officials to write legislation that employs reasonable and… https://t.co/BsXmdsZcA7 | | Twitter for iPhone | FALSE | Positive |
| 2020-09-16 20:32:27 | Just watched #TheSocialDilemma documentary Definitely think everyone needs to watch this and the irony of me tweeting about it... | TheSocialDilemma | Twitter for iPhone | FALSE | Neutral |
| 2020-09-16 20:27:40 | Watched #TheSocialDilemma | TheSocialDilemma | Twitter Web App | FALSE | Neutral |
| 2020-09-16 20:26:55 | Ok I'll give it a go. Watching now #TheSocialDilemma | TheSocialDilemma | Twitter for iPhone | FALSE | Positive |
| 2020-09-16 20:21:53 | You will see this tweet if the Twitter AI thinks it's profitable for you to see it. #TheSocialDilemma | TheSocialDilemma | Twitter for iPhone | FALSE | Neutral |
| 2020-09-16 20:09:18 | Me 15 minutes into #TheSocialDilemma on Netflix https://t.co/BcTOyD4Ads | TheSocialDilemma | Twitter for iPhone | FALSE | Neutral |
| 2020-09-16 20:07:21 | Well, damn. #TheSocialDilemma https://t.co/ rmxCndS6Z7 | TheSocialDilemma | Twitter for iPhone | FALSE | Neutral |
| 2020-09-16 19:51:51 | Felt same after watching the #TheSocialDilemma https://t.co/fsODAWXdcK | TheSocialDilemma | Twitter for Android | FALSE | Neutral |

Note: Emojis were present in the original tweets but are omitted here for formatting compatibility.

**Appendix II.** Sample from Dataset Two

| textID | text | selected_text | sentiment |
|---|---|---|---|
| cb774db0d1 | I`d have responded, if I were going | I`d have responded, if I were going | neutral |
| 549e992a42 | Sooo SAD I will miss you here in San Diego!!! | Sooo SAD | negative |
| 088c60f138 | my boss is bullying me... | bullying me | negative |
| 9642c003ef | what interview! leave me alone | leave me alone | negative |
| 358bd9e861 | Sons of ****, why couldn`t they put them on the releases we already bought | Sons of ****, | negative |
| 28b57f3990 | http://www.dotheboucny.com/smf - some shameless plugging for the best Rangers forum on earth | http://www.dotheboucny.com/smf - some shameless plugging for the best Rangers forum on earth | neutral |
| 6e0c6d75b1 | 2am feedings for the baby are fun when he is all smiles and coos | fun | positive |
| 50e14c0bb8 | Soooo high | Soooo high | neutral |
| e050245fbd | Both of you | Both of you | neutral |
| fc2cbefa9d | Journey!? Wow... u just became cooler. hehe... (is that possible!?) | Wow... u just became cooler. | positive |
| 2339a9b08b | as much as i love to be hopeful, i reckon the chances are minimal =P i`m never gonna get my cake and stuff | as much as i love to be hopeful, i reckon the chances are minimal =P i`m never gonna get my cake and stuff | neutral |
| 16fab9f95b | I really really like the song Love Story by Taylor Swift | like | positive |
| 74a76f6e0a | My Sharpie is running DANGERously low on ink | DANGERously | negative |
| 04dd1d2e34 | i want to go to music tonight but i lost my voice. | lost | negative |
| bbe3cbf620 | test test from the LG enV2 | test test from the LG enV2 | neutral |
| 8a939bfb59 | Uh oh, I am sunburned | Uh oh, I am sunburned | negative |
| 3440297f8b | S`ok, trying to plot alternatives as we speak *sigh* | *sigh* | negative |
| 919fa93391 | i`ve been sick for the past few days and thus, my hair looks wierd. if i didnt have a hat on it would look... http://tinyurl.com/mnf4kw | sick | negative |
| af3fed7fc3 | is back home now gonna miss every one | onna | negative |
| 40e7becabf | Hes just not that into you | Hes just not that into you | neutral |

Note: Emojis were present in the original tweets but are omitted here for formatting compatibility.