Developing a Text Classifier for Emotion Detection in Social Media

Introduction to the Domain-Specific Area

Emotion detection in text is a growing field within Natural Language Processing (NLP), focused on identifying the underlying emotional tone of written language. As digital communication dominates online platforms, recognising emotions in short texts such as tweets or chat messages has practical relevance in domains like customer service, mental health support, social media monitoring, and digital well-being. These systems help automate emotional understanding, enabling applications like empathetic chatbots, content moderation tools, and mental health tracking systems.

Recent literature has explored both classical and deep learning approaches for emotion classification. For example, Colnerič and Demšar (2018) compared traditional machine learning with BERT-based models, showing that although deep learning offers improved accuracy, it requires greater computational resources. For resource-constrained applications, classical models such as logistic regression remain appealing due to their interpretability and efficiency. This project evaluates such classical approaches in detecting emotion in real-world, short-form social texts.

Description of the Selected Dataset

This project uses the dair-ai/emotion dataset, accessed via Hugging Face Datasets. It consists of approximately 20,000 English-language tweets labelled with one of six basic emotions: sadness, joy, love, anger, fear, or surprise. The dataset is split into 16,000 samples for training, 2,000 for validation, and 2,000 for testing. Each record includes a text field and a corresponding integer label (0–5).

The dataset is pre-cleaned and tokenised, requiring minimal additional processing. Its structure reflects real-world social media communication, where messages are often informal and brief. However, limitations exist, notably class imbalance—emotions like joy and sadness are more prevalent than surprise or love. This imbalance can skew model performance and calls for careful evaluation metrics. Furthermore, ethical considerations around public tweet usage (e.g., consent, representation) should be acknowledged, especially for emotionally sensitive content.

Table 1 shows the mapping between numeric labels and their corresponding emotion categories.

Label	Emotion
0	Sadness
1	Joy
2	Love
3	Anger
4	Fear

Label	Emotion
5	Surprise

Objectives of the Project

The primary goal of this project is to develop and evaluate a machine learning-based text classifier for emotion detection in short-form, informal texts. The system aims to accurately assign one of six emotion labels to each input tweet using interpretable, efficient modelling techniques.

Specific objectives include:

- Applying minimal pre-processing appropriate for informal text
- Converting textual input into numerical features using TF-IDF
- Implementing logistic regression as a baseline model
- Comparing results against a simpler Naïve Bayes model
- Analysing class-level performance and model limitations

This project's broader contribution is a reproducible, low-resource pipeline suitable for integration into real-world emotion-aware applications. Its focus on transparency and accessibility makes it a useful baseline for future research and deployment in low-compute environments.

Evaluation Methodology

Evaluation focuses on four standard metrics: accuracy, precision, recall, and F1-score. These are calculated per class to ensure detailed insight into each emotion category. Additionally, macro and weighted averages are used:

- **Macro-average** treats all classes equally, making it ideal for assessing fairness in imbalanced datasets.
- **Weighted-average** adjusts scores based on class frequency, offering a more realistic overall performance estimate.

Accuracy measures the proportion of correctly predicted labels. Precision evaluates the correctness of positive predictions, while recall measures how well the model captures actual positives. F1-score balances the two, offering a harmonic mean that accounts for both false positives and false negatives.

Metrics are computed on the held-out test set of 2,000 samples. Confusion matrices are also used to visualise classification patterns and misclassifications. This methodology ensures that the evaluation captures both overall success and weaknesses specific to underrepresented emotion categories.

The confusion matrix in Figure 1 is used to visualise prediction distributions per emotion class.

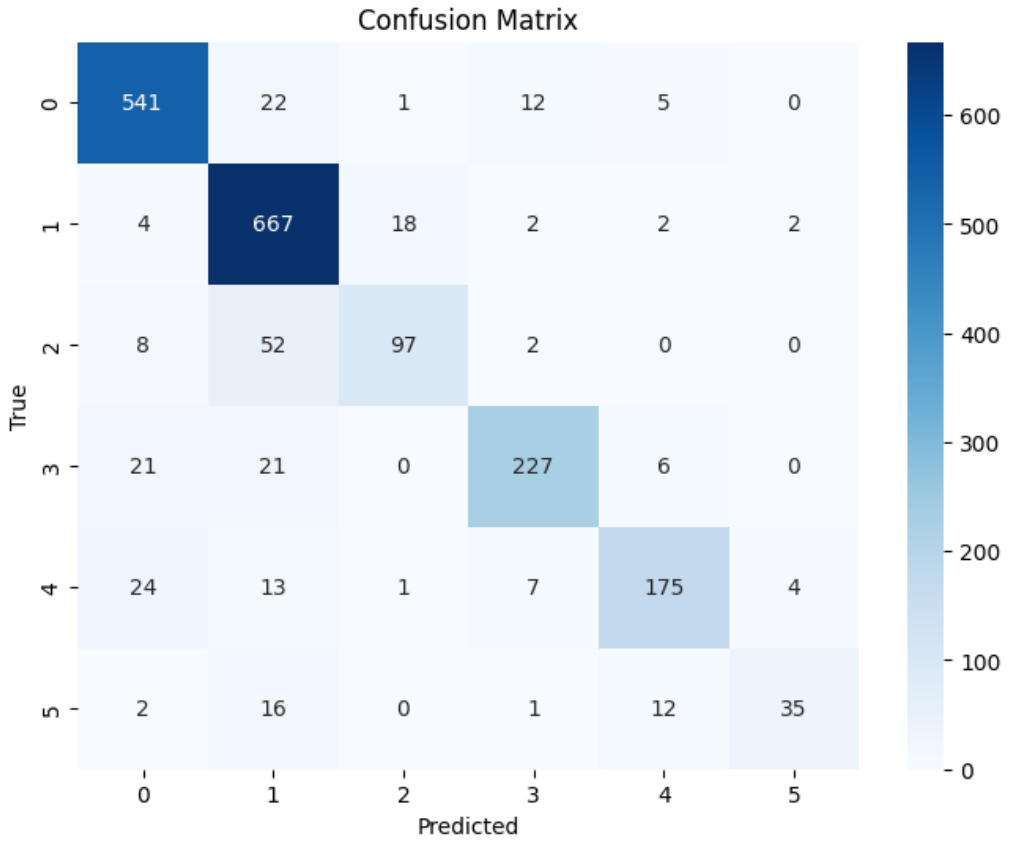

Figure 1: Confusion matrix for Logistic Regression classifier. Most classes are accurately predicted, especially joy and sadness, while love and surprise show more misclassifications.

Evaluation (of Model Performance)

The logistic regression model outperformed Naïve Bayes across all metrics. Logistic Regression achieved 87% accuracy and a macro-average F1-score of 0.81. In contrast, Naïve Bayes achieved 73% accuracy and a macro F1-score of 0.50.

Naïve Bayes performed adequately on majority classes like sadness and joy but performed poorly on minority emotions such as surprise and love. This is largely due to its feature independence assumption, which fails to model subtle linguistic dependencies. Logistic Regression, by contrast, handled sparse TF-IDF features more effectively and showed greater flexibility in distinguishing between nuanced expressions of emotion.

The confusion matrix for Naïve Bayes showed overprediction of dominant classes like joy, while Logistic Regression displayed a more balanced and distributed error pattern. These findings confirm logistic regression’s superior generalisation and suitability for emotion classification on imbalanced, short-form text data.

As shown in Figure 1, the logistic regression classifier demonstrates balanced performance across most emotions. Figure 2 illustrates that Naïve Bayes, by contrast, misclassifies minority emotions such as 'surprise' more frequently.

Metric	Logistic Regression	Naïve Bayes
Accuracy	0.87	0.73
Macro F1-Score	0.81	0.50
Weighted F1-Score	0.87	0.68

Table 1. Comparison of evaluation metrics for both classifiers.

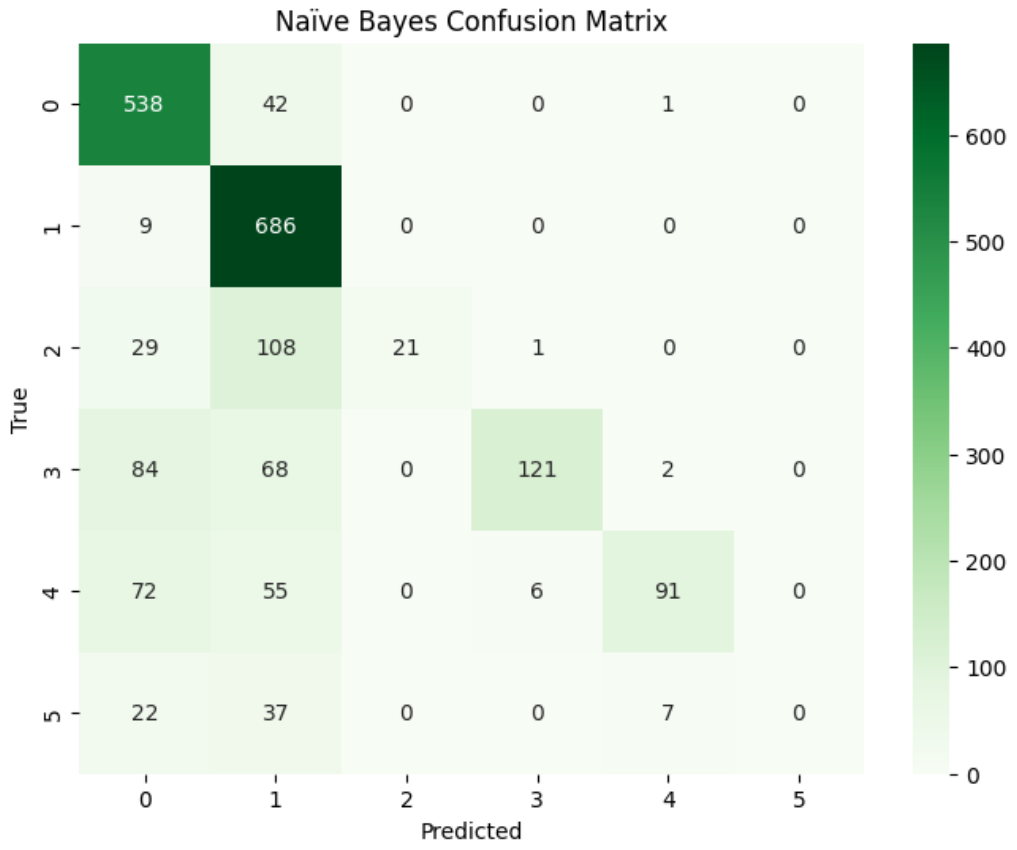

Figure 2: Confusion Matrix – Naïve Bayes

Evaluation of the Project and Its Results

This project successfully developed and evaluated a reproducible, interpretable emotion detection pipeline using classical machine learning. Its primary contribution lies in validating the effectiveness of low-compute models—especially logistic regression—for accurately classifying emotions from brief, informal social media text.

The logistic regression model demonstrated robustness to class imbalance and strong performance without reliance on GPUs or extensive hyperparameter tuning. In contrast, Naïve Bayes, while

computationally efficient, showed structural limitations in capturing nuanced or minority emotions. The model's shortcomings reinforced the value of more flexible classifiers even in lightweight settings.

The project design supports reusability and extensibility. The codebase is clean, modular, and designed for reproducibility. Future directions may include: integrating transformer-based architectures like BERT, experimenting with data augmentation or synthetic oversampling to mitigate imbalance, and exploring multilingual datasets for broader applicability. Despite its simplicity, this project offers a solid, transparent starting point for emotion-aware applications in academia and industry.

Reference

Colnerič, N., & Demšar, J. (2018). Emotion recognition on Twitter: Comparative study and training a unified model. In 2018 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 1-8). IEEE. https://ieeexplore.ieee.org/document/8295234

Hugging Face. (2024). dair-ai/emotion dataset. Hugging Face Datasets. https://huggingface.co/datasets/dair-ai/emotion

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.