Emotion analysis of Reddit's users' comments on migration issues using fine-tuned DistilBERT

Kalina Maria Piskorska

Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics
{piskorska2}@uniba.sk

Abstract. The project explored the possibilities of emotion classification on social media data concerning societal issues. The specific task consisted of fine-tuning the DistilBERT model, with a logistic regression baseline, on a dair-ai/emotion dataset. A logistic regression classifier provided an initial benchmark for emotion prediction, while DistilBERT leveraged the contextual language processing abilities to provide more effective predictions. Overall test F1-score of the logistic regression model was 0.87 and for DistilBERT - 0.92.

After training, both models were used to research the affective states of Reddit users on the Europe subreddit by analyzing their predictions on real-world data scraped from the website (for this project's purpose). The comparison of the results, with some quantitative and qualitative analyses clearly demonstrated DistilBERT's superiority in capturing nuanced emotional expression in user comments. The project allowed me to make first steps and experiment in the field of emotion analysis using natural language processing, providing information and lessons for the future (potential master's thesis topic).

1 Introduction

Social media has become a prominent place for discussion on important societal issues. One of them is the phenomenon of mass migration. Reddit's r/europe subreddit [1] is a good source for exploring user sentiment regarding the issue. The training dataset was the dair-ai/emotion dataset available on Hugging Face [2]. It contains 20 000 examples of English Twitter posts classified into 6 emotion classes - sadness, joy, love, anger, fear and surprise. It was split into training, validation and test subsets in 80/10/10 proportion. The number of texts classified into different groups were not even in the training set which could have affected predictions. The normalized counts (what percentage of the whole set the class constituted) can be seen on Figure 1.

The scraped Reddit dataset consisted of 33 812 comments from 100 different posts on r/europe. The posts were chosen from the subreddit with the query word "immigrants" and sorted by most relevant. To access the posts, PRAW was used - the Python Reddit API Wrapper [3] (the application was also formally registered through Reddit API). This data has been preprocessed but not labelled - as it would require enormous additional labour. Therefore, the results of predicting emotions of the Reddit posts are treated more as a curiosity and training than something the models could be evaluated on.

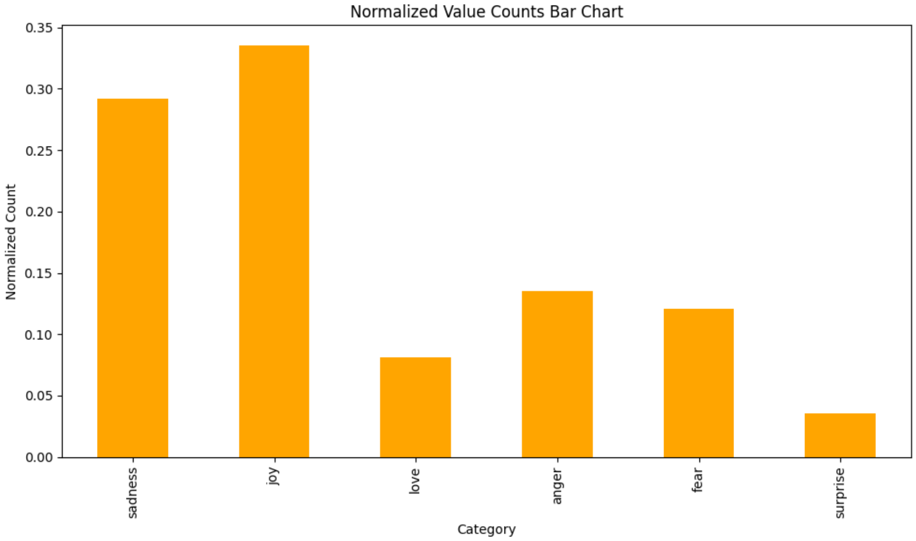

Fig. 1. Percentage of classes in the training dataset

2 Context

Emotion and sentiment analysis have been extensively studied in the field of natural language processing (NLP). Traditional methods, including logistic regression and support vector machines, rely on feature engineering and often use term-frequency inverse document frequency (TF-IDF) or bag-of-words representations. Modern transformer-based models such as BERT and its variants, including DistilBERT [4], offer state-of-the-art results by encoding rich contextual information. Recent studies (e.g., [5,6,7]) demonstrate the superior performance of fine-tuned transformers in emotion classification tasks, making them a natural choice for this project. Furthermore, Reddit has been chosen by many researchers as a source of real-world data for emotion detection using machine learning and deep learning ([8,9]). I could not find a paper specifically targeting attitudes towards immigrants among Reddit users.

3 Description and justification of methods

Data Collection The dair-ai/emotion dataset was chosen as the primary dataset for fine-tuning the DistilBERT model. This dataset, easily accessible through

Hugging Face, is well-organized and derived from social media platforms, making it particularly relevant for this emotion classification task. Its popularity among researchers and the availability of many examples ensure robust model training and evaluation. For testing on real-world data, comments were scraped from the Europe subreddit on Reddit. The Reddit API, accessed via PRAW library, facilitated the collection process. Reddit provides a diverse set of data, capturing a wide range of topics and emotional expressions due to the free and often anonymous nature of user interactions. The search for specific topics is easy with query words. Initially, Twitter (now X) was considered as a data source; however, changes to their API restricted free data scraping, making Reddit a more viable alternative.

The data scraped from Reddit had to be preprocessed first to achieve a clean dataset to run the models on. To do that, all characters were checked and removed if they were not letters (special characters) and lowercased. In addition, to make the texts resemble training data more, longer comments were cut to 512 characters. Some of the scraped comments were also deleted if the content was "[deleted]/[removed]/[censored]". Later on, some of the examples ended up being empty rows - so they also had to be removed before using them as a dataset for the models. A limitation that has not been overcome in this project (but should be in the future attempts) was the fact that even though most comments are in English, some comments might have been in other languages and therefore, the models trained on English sets could not classify them correctly.

Baseline Model A logistic regression model was trained on TF-IDF features derived from the dair-ai/emotion dataset. Logistic regression was selected as it offers a simple, interpretable baseline to establish the lower bound for performance. Its reliance on TF-IDF ensures focus on term importance, though it lacks the capability to capture context beyond individual tokens. The class Logistic Regression from the scikit-learn library by default uses the One-vs-Rest method that allows for multi-class prediction that is suitable for this task.

Fine-tuned model A DistilBERT model was fine-tuned on the dair-ai/emotion dataset using a classification head with a softmax layer. DistilBERT was chosen for its efficiency and ability to encode contextual word representations, which are critical for understanding nuanced emotional expressions.

Key Hyperparameters for Training

`output_dir='./logs/run1/'`

- Using Weights & Biases (a cloud-based platform) for logging metrics in real-time and monitoring storage, checkpoints, etc. It worked together with the Hugging Face's transformers library.

`per_device_train_batch_size=32`

– A larger batch size balances computational efficiency with stability of the training. It's a common choice for NLP tasks.

`per_device_eval_batch_size=128`

– Speeds up evaluation.

`gradient_accumulation_steps=2`

– Helps improve generalization by accumulating gradients before updating weights. It's better for memory constraints.

`learning_rate=2e-5`

– A low learning rate works well for BERT, allowing for fine-tuning without degrading the pre-trained weights too much.

`num_train_epochs=3`

– Not too many runs through the dataset to avoid overfitting.

`seed=42 and data_seed=42`

– Ensures reproducibility.

`dataloader_num_workers=2`

– Helps speed up data loading.

4 Brief description of technical issues

Limited computational power Using Google Colab provides free access to GPUs, however, with more demanding tasks and a beginner's trial-and-error approach - the GPU restrictions can be challenging. There are restricted runtime durations and session timeouts that slowed down some of the steps of the project and complicated experimentation.

Scraping data As mentioned before, initially the data on "immigrants" was supposed to be scraped from Twitter (X). However, after starting the project, I found out that scraping posts from Twitter is not as accessible anymore (the API is not free for bigger samples). It meant that I had to shift my strategy from using snscrape to another library. I have settled on using Reddit data that may have been a better source when it comes to the content in the end (as subreddits allow for targeting a very specific topic and users). Nonetheless, when the training data is from Twitter and the scraped data for predictions is from Reddit, there might be relevant differences in the platforms that affect the results. In addition, the short nature of Twitter posts meant that some of the longer Reddit posts had to be cut off to match the type of training examples.

Characteristics of data Reddit users exhibit particular behaviors, there is a specific culture on that platform that can complicate the models' ability to correctly predict the users' emotions. One of the characteristics is the frequent use of sarcasm. Sarcasm is hard for models to recognise and I have noticed instances where DistilBERT's predictions are the exact opposite of the user's intent because of that. The fact that the discussions concerned political topics also influences the tone and the way users express emotions which is not covered by the more basic dair-ai/emotion training set.

5 Experimental evaluation

For the first stage of the project, Logistic Regression and DistilBERT were trained and tested on the dair-ai/emotion dataset. Below in tables 1-3, one can see the results of the training (for DistilBERT) and tests (both models). As expected, DistilBERT outperformed the baseline model, achieving testing accuracy of 92.1%, while Logistic Regression had the testing accuracy score of 87%. The difference in results can be explained by DistilBERT's ability to take into account contextual information and therefore the more subtle differences in emotional expressions of the users.

Table 1. Test results of Logistic Regression on dair-ai/emotion set

Class	Precision	Recall	F1-score
sadness	0.90	0.93	0.92
joy	0.84	0.96	0.89
love	0.82	0.59	0.69
anger	0.90	0.82	0.86
fear	0.89	0.79	0.84
surprise	0.85	0.52	0.64
accuracy	0.87		

Table 2. Training scores of DistilBERT on dair-ai/emotion set

Epoch	Training Loss	Validation Loss	F1-score
1	—	0.283	0.918
2	0.9946	0.187	0.933
3	0.9946	0.174	0.930

In the graph below (Figure 2), one can also notice that there were clear differences in the correct predictions of the fine-tuned DistilBERT model per class. The class "surprise" was particularly troublesome and had the lowest percentage of correct predictions in testing stage. That might be due to the differences in

Table 3. Test results of DistilBERT on dair-ai/emotion set

Class	Precision	Recall	F1-score
sadness	0.97	0.96	0.96
joy	0.94	0.94	0.94
love	0.79	0.82	0.80
anger	0.92	0.93	0.93
fear	0.87	0.92	0.89
surprise	0.79	0.62	0.69
accuracy	0.9210		

the number of examples per class in the training set, as displayed in Figure 1 - the class "surprise" had the least number of representatives.

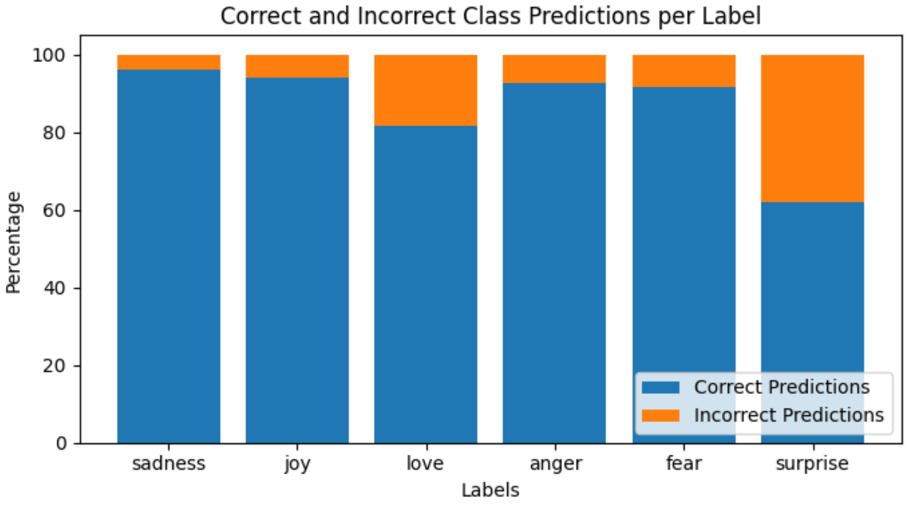

Fig. 2. Correct and incorrect prediction by DistilBERT per class

Finally, the models were run on the comments from Reddit to predict their emotional class. The dataset was not labeled, therefore, the performance could not be evaluated by comparing the predictions to true labels. Nevertheless, the agreement between the models' predictions was investigated. The predictions of the two models were the same in 47.383% of the cases. That shows quite large dissimilarities in classification. In Figure 3, the labels assigned by the models can be compared in terms of frequency. Clearly, the Logistic Regression model predicted the "joy" label for most of the comments, while fine-tuned DistilBERT had better prediction distribution among the classes, even though "joy" was still the most frequently predicted class.

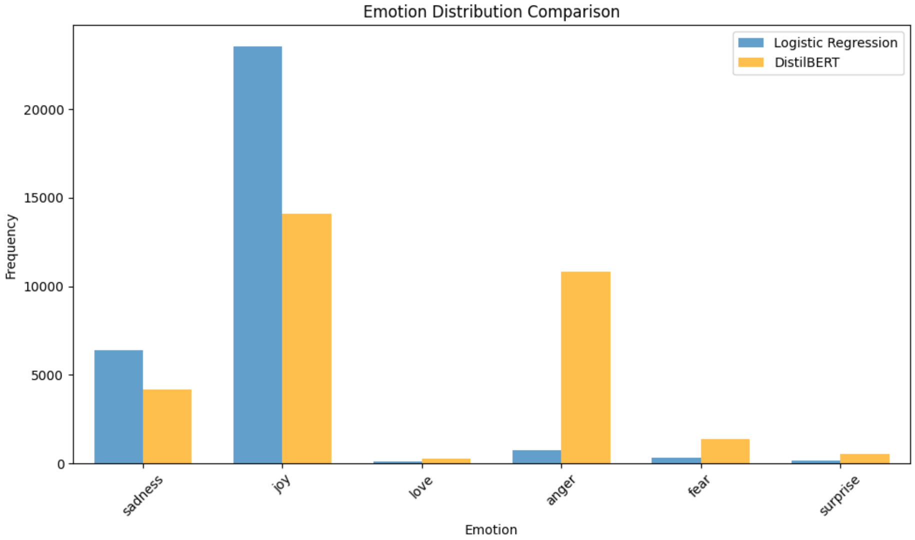

Fig. 3. Comparison of number of models' predictions per class

For another method of comparison between the models, a confusion matrix (Figure 4) was created to visualise the agreement on specific instances. In a case of high agreement, the diagonal line of the table would be highlighted in darker blue - however, in this task the agreement was quite low. Since most of the predictions of the Logistic Regression model were "joy" and "sadness", the first two columns representing these classes for LR show more saturation. There was simply no other option since LR classified small numbers of the examples as the other classes.

The models had the most similar predictions and were in agreement for the "joy" class the most. At the same time, a lot of the instances that DistilBERT classified as anger, Logistic Regression classified as joy. It shows that the BERT architecture picks up on more nuanced emotions.

When checking the predictions file and analysing the comments - one can see that sarcasm was particularly difficult for the models to detect e.g., "sure this will not cause future issues" was classified as joy by both models. Still, even in those more complicated situations, BERT performed better - e.g., for a comment "the repeated stupidity of political leaders across western europe continues to astound me", it classified it as "anger", while LR "joy". This might partly show why there is such a disagreement between models for "joy" versus "anger" classes. Sometimes the Internet/Reddit-specific slang confused the models - such as "spain is cooked" or "wtf" - where both models classified it as "joy".

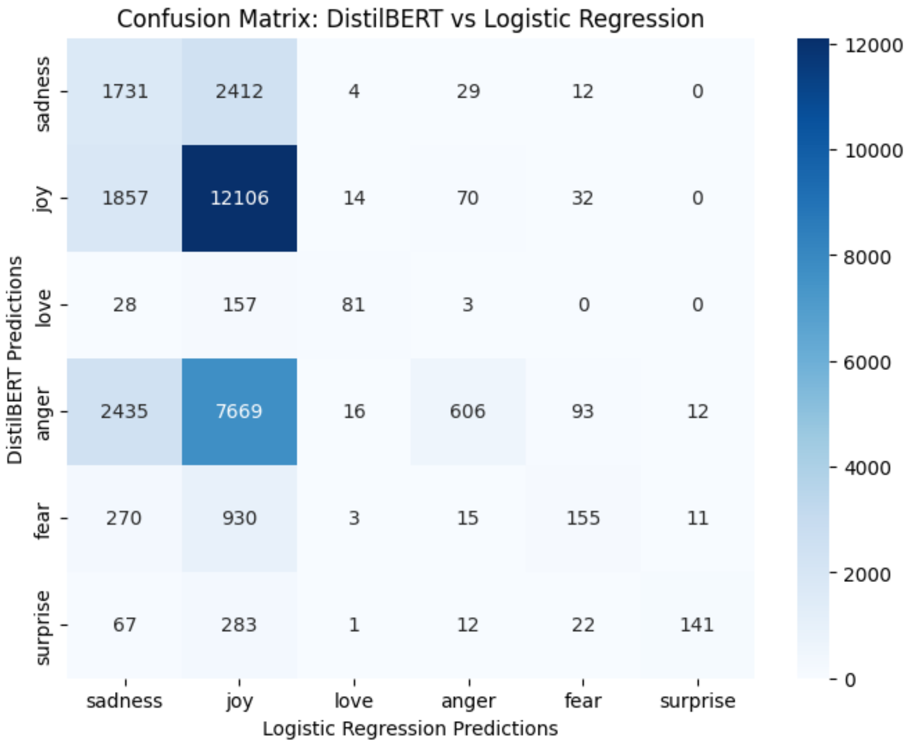

Fig. 4. Confusion matrix

6 Conclusion

This project was successful as it achieved its primary goal of providing an opportunity for hands-on practice and exploration of emotion analysis using NLP techniques. It offered me significant learning opportunities, particularly in handling real-world challenges such as data preprocessing, fine-tuning transformer models, and evaluating performance on scraped, unlabeled social media data. The use of DistilBERT demonstrated the power of transformer-based models for capturing nuanced emotional expressions compared to traditional methods like logistic regression.

Looking at the predictions, the models, even fine-tuned DistilBERT, still made some obvious mistakes. In hindsight, several enhancements could be made to improve the overall quality and robustness of the project.

- Use of larger transformer models: Exploring models such as RoBERTa or BERT-large could further enhance performance, due to their greater capacity for contextual understanding.

- Addressing class imbalance: Implementing strategies or finding datasets with better representation of different classes could help mitigate the impact of class imbalances observed in this project.
- Obtaining/creating a labeled Reddit dataset: Collecting and annotating a dataset from Reddit comments would allow for more precise evaluation and model tuning specific to the domain. Furthermore, Reddit comments often relate to the previous comments - it is a thread and just analysing one comment can not be enough context, that should be taken into account.

By incorporating these improvements, future iterations of the project could achieve even greater accuracy and generalizability, making the model more effective for analyzing real-world social media discussions and ideally, provide real value for research in the field of cognitive science.

7 Appendix

Files used for the project can be found here:
https://github.com/kalina-piskorska/ml-uniba-project-2025.git

References

1. Europe subreddit, `https://www.reddit.com/r/europe/`, last accessed 2025/01/05
2. Saravia, E., Liu, H. C. T., Huang, Y.-H., Wu, J., Chen, Y.-S.: CARER: Contextualized Affect Representations for Emotion Recognition. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October-November 2018, 3687–3697. Association for Computational Linguistics (2018)
3. PRAW Documentation, `https://praw.readthedocs.io/en/stable/`, last accessed 2025/01/05
4. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv **abs/1910.01108** (2019)
5. Acheampong, F. A., Nunoo-Mensah, H., Chen, W.: Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review **54**(8), 5789–5829 (2021)
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
7. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification?. In: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings **18**, 194–206. Springer International Publishing (2019)
8. Ren, L., Lin, H., Xu, B., Zhang, S., Yang, L., Sun, S.: Depression detection on reddit with an emotion-based attention network: algorithm development and validation. JMIR Medical Informatics **9**(7), e28754 (2021)
9. Turcan, E., McKeown, K.: Dreaddit: A reddit dataset for stress analysis in social media. arXiv preprint arXiv:1911.00133 (2019)