Final project machine learning

Emotion detection using machine learning methods

Louke Boom

1. Introduction

Our lives are becoming increasingly digital. We spend more time online and on social media. AI becomes more prominent in our day to day lives, with people even marrying to their personal chatbots (Heritage, 2025). Social interactions often go hand in hand with emotions. It is already difficult for humans to distinguish emotions from one another. Let alone for a system that does not 'feel.' With these chatbots becoming more integrated into everyday life it is important for them to be able to produce a proper response based on the emotions of their conversational partner. Before said chatbots are even able to produce an answer they must first be able to extract emotions from the input. That is where this research comes into play. For this research, I will talk about two different machine learning methods for emotion detection based on the mteb emotion dataset. One for classical machine learning (logistic regression) and one for deep learning (BERT transformers). First, the current state of the art in the respective research will be treated. Secondly the two different methods will be explained, what preprocessing steps had to be taken, why these methods were chosen and finally a comparison based on the performance of these models will be made.

2. Related research

In recent years there have been multiple instances of research projects that try to do similar things. A previous study on text-based emotion recognition using only classical machine learning algorithms on the ISEAR dataset achieved around 64 % accuracy with Multinomial Naïve Bayes (Ab. Nasir et al., 2020). CARER, similar research achieved around 81 % accuracy, but they did this trough a sophisticated graph-based pattern extraction and engineering (Saravia et al., 2018). Xu et al (2020) achieved a 97.6% accuracy by using CNN with Word2Vec embeddings on binary sentiment classification for microblogs. The research that is explained in this paper is related to the three previous mentioned but is more of a combination of the three. This research has a similar multiclass structure like Ab. Nasir et al., (2020), but the data it uses is more alike to that of Saravia et al., (2018). This research also uses deep learning methods like xu et al., (2020) but uses more modern methods that were not available for the studies at the time of publication.

3. System structure

3.1 Global pipeline structure

The Pipeline of this model is structured as follows. First there is the loading of the data. The data used for this model comes from the mteb/emotion dataset from Hugging Face (2025). Second, we preprocess the data with two parallel preprocessing pipelines: minimal and extensive. This is because the used methods required different forms of preprocessing. Third, we have the feature extraction which for the classical Machine Learning is TF-IDF vectorization with a max of five thousand features and a tokenization for DistilBERT as part of the deep learning method. Forth, there is the model training with balanced class weights for the logistic regression and fine-tuning DistilBERT. Finally, we have evaluation of the models.

3.2 Dataset structure

As mentioned before the data comes from the mteb/emotion set (2025). The data comes from English Twitter messages that consist of six basic emotions: anger, fear, joy, love, sadness, and surprise. These emotions count as the six classes used for this multiclass model. The data contains 19930 entries divided under training, test, and validation samples (with respective sizes of 15956, 1988 and 1986 samples each). This table shows the data distribution per class.

Table 1
Class Distribution Across Dataset Splits

Emotion	Train	Validation	Test	Total
Joy	5,345 (33.5%)	700 (35.2%)	688 (34.6%)	6,733
Sadness	4,663 (29.2%)	550 (27.7%)	579 (29.2%)	5,792
Anger	2,152 (13.5%)	274 (13.8%)	274 (13.8%)	2,700
Fear	1,931 (12.1%)	211 (10.6%)	224 (11.3%)	2,366
Love	1,297 (8.1%)	173 (8.7%)	156 (7.9%)	1,626
Surprise	568 (3.6%)	80 (4.0%)	65 (3.3%)	713
Total	**15,956**	**1,988**	**1,986**	**19,930**

It is important to note that there is imbalance between the classes. Joy and sadness are the two most represented classes while surprise is the least represented. But the distribution of the representation is consistent over train, test, and validation splits. It is also important to note that the samples overall are short. With a range of characters between 11 and 300 and an average length of 97 characters.

3.3 Preprocessing

As mentioned in the pipeline structure there were two different methods of preprocessing namely, minimal, and extensive. For both there were preliminary preprocessing steps taken like lowercasing the text, removing extra whitespaces, URLs, and mentions of usernames as these are not essential for the research that is being performed here.

The extensive preprocessing is used for the logistic regression methods. It does not differ that much from the minimal, but this cleans most special characters and repetition

of characters. This is because classical ML with TF-IDF performs better with cleaner and more standardized text where the feature extraction focuses on word frequencies instead of contextual nuances. The minimal was used for the BERT Transformer, as they are usually trained on more natural text. The use of punctuation and special structure are important for emotional indications and are therefore essential for the training of said transformer.

3.4 Method selection and structural analysis

For the classical machine learning part of this research logistic regression was chosen. The main reason for this decision is because it is an interpretable linear model that is fast to train and manages features like TF-IDF vectors efficiently. For this type of data, it also required quite little fine-tuning and preprocessing of the data. As the results later will show it also outperforms more classical methods like Naïve Bayes or decision Trees. This method also works quite well because the features in the text are linearly separable. TF-IDF was used because it captures both term frequency and inverse document frequency. This makes sure that words that are common are interpreted as les important that more unique words. The features were limited to 5000 to keep the balance between expressiveness an efficiency. This also prevents overfitting on rare features. The features were also configured to consist of unigrams and bigrams. I also implemented L2 regularization and let it have a max of 1000 iterations.

DistilBERT was chosen because it is about 40% smaller and 60% faster than BERT-Base but it keeps most of its performance (95%). This also ensures that the training time of the model is much quicker. The amount of data that is used for this research is also not that large for which DistilBERT is more than enough and larger models might overfit this small amount of data. Transformers are more beneficial as deep learning methods because they negate a few cons of other methods like RNN's. These are much slower and could have issues with distant relations between words and may have vanishing gradients. DistilBERT was configured to have a 2e-5 learning rate with a batch size of 16train, and 32 eval which ran for 3 epochs.

As evaluation metrics for both methods, we will look at accuracy, F1-score (macro and weighted), and F1-scores per class.

4. Discussion of results

The following two tables show us the results of the overall performance of the models and the performance per class in the appendix are the confusion matrices for both models that were used: When we look at these results both models do quite well in

Table 2
Overall Performance Comparison on Test Set

Model	Accuracy	F1-Macro	F1-Weighted	Improvement
Logistic Regression	87.8%	84.2%	88.0%	Baseline
DistilBERT	**92.9%**	**88.8%**	**93.0%**	**+5.1%**

comparison to the similar studies that were aforementioned. BERT does outperform Logistic Regression on all metrics (+ 4,6 – 5,0%). When we look at the class improvement, we also see there an improvement on all classes. Important to note here is that

Table 3
Per-Class Performance Comparison on Test Set

Emotion	Logistic Regression			DistilBERT			Δ F1
	Prec.	Rec.	F1	Prec.	Rec.	F1	
Sadness	0.94	0.88	0.91	0.96	0.97	**0.96**	+0.05
Joy	0.93	0.88	0.90	0.95	0.95	**0.95**	+0.05
Love	0.70	0.91	0.79	0.83	0.88	**0.85**	+0.06
Anger	0.85	0.89	0.87	0.94	0.93	**0.93**	+0.06
Fear	0.90	0.83	0.87	0.93	0.84	**0.88**	+0.01
Surprise	0.60	0.86	0.71	0.68	0.82	**0.74**	+0.03
Macro Avg.	0.82	0.88	0.84	0.88	0.90	**0.89**	+0.05

both models do have difficulties with Surprise. This is mainly because Surprise is much less represented in the data as table 2 shows. For the majority classes (Joy & Sadness) do both models perform very well. Logistic regression does sometimes confuse Love for Joy, but BERT also does this better. This could suggest that BERT has a better semantic understanding than LR.

During training BERT also had a good validation performance. It had a max validation loss of 0.143 and this loss decreased with each epoch. For both models are there also no signs of overfitting which is desirable as well. The confusion matrices in the figures section are the best indication of where the most mistakes were made. While we do see that BERT does have significantly less confusions, it did make similar mistakes to LR. These being that Love if frequently confused with Joy and there also is some confusion between Anger and Sadness. A reason for this occurring could be semantic similarities between these emotions. And of course, Surprise is the class that experienced the most errors which is probably due to the lack of data for this class.

However, both methods have shown that they performed much better than the studies that were explored in the beginning. With similar studies only achieving results of 81% in comparison to our 92%. (Xu et al., used only a binary model and not a multi-class one so it is not a fair comparison with our research, but it shows the possibilities of deep learning). This indicates that the technology that is used to perform these types of tasks has evolved rapidly over the past 5 years and could potentially evolve even more in the future.

5. Conclusion

This research has shown that modern methods become more capable of classifying emotions in comparison to technologies that were available only five years ago. It shows what possibilities we could have towards the future with classifying emotions which is something many humans have difficulties with.

While this experiment went rather well it is important to note that the data that was used consists of twitter messages which often are quite emotionally loaded and therefore could mean that this improved the working of the models. For future research it could be relevant to try and uses this model on more ambiguous and longer texts in comparison to the test data and see how these methods perform then. I do think that the results would be less spectacular than they were during this experiment.

6. important note to research

Claude.ai was used to help me with multiple smaller tasks throughout this research. It was used to help me correct code when it was not working, give me an explanation of which methods of preprocessing were required for the data in order to have the two models work properly, and I used it to create the tables into the LaTeX format that was used to write this paper. Finally, it was used to make the GitHub Readme.

7. Figures

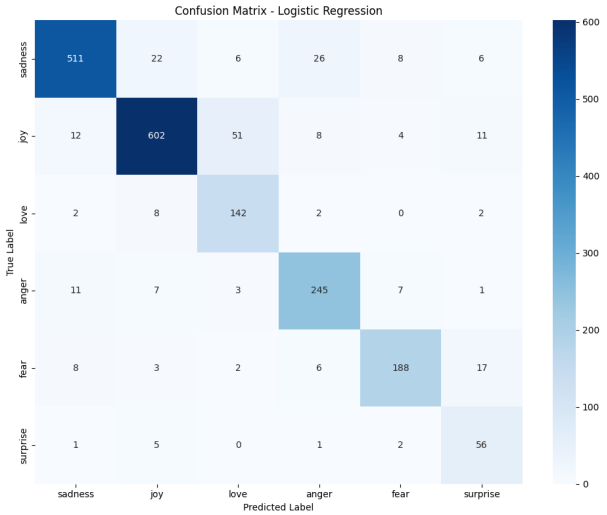

Figure 1
Confusion matrix LR

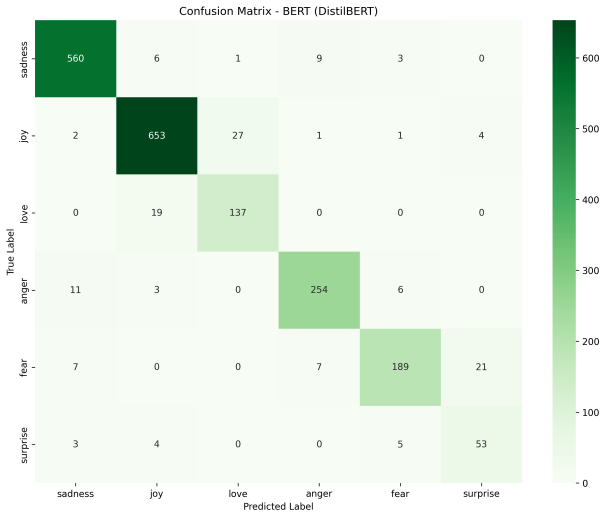

Figure 2
Confusion matrix BERT

References

[1] Ahmad Fakhri Ab. Nasir, Eng Seok Nee, Chun Sern Choong, Ahmad Shahrizan Abdul Ghani, Anwar P. P. Abdul Majeed, Asrul Adam, and Mhd. Furqan. Text-based emotion prediction system using machine learning approach. *IOP Conference Series: Materials Science and Engineering*, 769(1):012022, 2020.

[2] Anthropic. Claude.

[3] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, et al. MMTEB: Massive Multilingual Text Embedding Benchmark, 2025.

[4] Stuart Heritage. 'I felt pure, unconditional love': the people who marry their AI chatbots, 2025.

[5] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive Text Embedding Benchmark, 2022.

[6] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, 2018. Association for Computational Linguistics.

[7] Dongliang Xu, Zhihong Tian, Rufeng Lai, Xiangtao Kong, Zhiyuan Tan, and Wei Shi. Deep learning based emotion analysis of microblog texts. *Information Fusion*, 64:1–11, 2020.

[8] mteb/emotion● Datasets at Hugging Face, 2025.