Classifying Emotions expressed on Open Social Networks using BERT Transformer Model

Mohona Roy
MSci Data Science
mohona.roy@city.ac.uk

1 Problem statement and Motivation

The intricacies of emotions and their resulting behaviors can lead us through decisions or act as templates for the trajectory of our lives.
There have been several instances of psychological models that attempted to discretize them, the most prominent being Ekman's 6 basic emotions model, categorizing them as anger, disgust, fear, joy, sadness and surprise. With the rise of Open Social Networks (OSN) came microblogging, one that Twitter is best known for. The use of Machine Learning (ML) models to classify emotions have had marketing undertones, such as increasing user engagements. Emotions are too complex to classify. Taking Ekman's model, ambiguities in certain pairs of emotions like surprise with fear, and disgust with anger are present. Past solutions polarize emotions using binary classification methods with lexicon-based approaches. Following OpenAI's ChatGPT, the integration of AI has become prominent in our lives, hence, it is necessary to humanize emotions as they are. This study aims to establish a multi-class supervised model that can distinguish texts into 6 different emotions with high accuracy based on sentence context and semantics. We shall proceed with the following research question:

How does the combination of different feature extraction algorithms, paired with varying models performs on multi-classification problems with subjective ambiguity?

2 Research hypothesis

Our research hypothesis states that a transformer-based model can prove to be the best solution to our research question. First introduced by Devlin et al., 2019, uncased BERT models will not differentiate between capitalization of letters. Casual tones present in tweets provide language variability, which the uncased version of the BERT model is suitable for. BERT is a pre-trained model on a large corpus that can be fine-tuned to suit a smaller dataset. BERT is a bi-directional model; it can have a more concrete grasp on contexts by considering the entire sentence.

3 Related work and background

A similar problem was addressed by Rao et al., 2019, with multi-class sentiment analysis on news articles to understand social emotions. The paper uses ATM, automating applications of ML. Their experiment utilizes a lexical analysis model, with headlines tokenized with corresponding articles. The model's accuracy was 74.78%.

Yassine et al., 2010 referred to the task as "emotion mining" and utilized K-Means clustering to discretize different "friendship strengths" on Facebook data, resulting in 3 centroids. Emojis and acronyms were replaced with descriptions. This process is inline with object-orientation. An SVM was trained resulting in an accuracy of 87%.

Data was obtained in Wikarsa & Thahir, 2015 by using KDD, a data mining technique. A Naïve Bayes classifier was trained resulting in an 83% accuracy.
In Bouazizi & Ohtsuki, 2019, the authors experiment with unigram features by training binary classifiers in increasing numbers (up to the number of classes), indicating a "One vs All" approach. Their accuracy was 60.2%.

A more recent study (Sahu & Shah, 2020) also made the use of unigram modelling but with an SVM, with an accuracy of 85.20%.

A comprehensive approach was observed in Chawla & Mehrotra, 2018 with an ensemble of four classifiers: SVM, Logistic Regression, Naïve Bayes and SGDunder. The model best performed

with a voting-based algorithm with an F-score of 89% when with TF-IDF.

In Mossad et al., 2023, a BERT model was implemented on Arabic tweets using binary classification. Their accuracy was 78.93%, higher than LSTM.

BERT was used in Mahimaidoss & Sathianesan, 2023. Pre-processing included POS tagging to address slangs in tweets; accuracy was 91.58%.

A pre-trained DistilBERT used in Singh & Kumar, 2023 achieved 93.5% accuracy, compared to their baseline Naïve Bayes (78%), and SVM (88%), concluding that traditional supervised ML approaches performed higher than lexicon-based seen so far.

Lastly, in Sharma et al., 2021, an ANN was built using an embedding layer (pretrained GloVe 50D) with an LSTM and a GRU. To address the vanishing gradient problem, resulting in the LSTM performing with a 74% accuracy and GRU with 68%.

4 Accomplishments

1. Remove all stop-words - Completed
2. Tokenisation- Completed
3. Perform Exploratory Data to comprehend data - Completed
4. Build Rule-based classifier by obtaining the top 5 key words for each class - Completed
5. Train a Random Classifier, Majority Classifier, Naïve Bayes classifier and SVC by obtaining vectors using CountVectorizer. - Completed
6. Train a Random Classifier, Majority Classifier, Naïve Bayes classifier and SVC by obtaining vectors using TF-IDF vectorizer – Completed
7. Experiment with FastText vectorization – Not completed due to resources requirements
8. Perform an in-depth analysis of performance metrics on the test data for each – Completed
9. Obtain and adjust BERT - Completed
10. Tokenize dataset with BERT tokenizer - Completed
11. Train tuned BERT model– Completed
12. Conduct performance evaluation with validation set – Completed
13. Perform an in-depth analysis on test data – Completed

5 Approach and Methodology

1. As discussed, emotions are extremely ambiguous and can be easily misclassified (Sharma et al., 2021). Words have varying meanings based on how they are spelt. 'and' and 'aaaaand' have different senses albeit the same meaning. To address this, we will not perform any stemming for tone preservation.
2. The high computational complexity is a potential challenge. BERT is a bidirectional model that traverses each document in both ways. To reduce computational complexity, the sentences were all padded to 140 characters. Twitter had this limit at the time the data was collected.
3. While the non-linearity of a BERT model is expected to perform better than the baseline models, BERT is generally very difficult to interpret. This will be addressed by testing the model's performance on a validation set while it is running, so that we can evaluate the model at each epoch.
4. The first baseline model used is a Rule-based classifier. This would allow us to better understand if the presence of certain keywords act as an orientation towards the ground truth. A Random classifier is used to be compared to with the majority class classifier to understand how much the unbalanced labels could have an effect. A Multinomial Naïve Bayes classifier and an SVM classifier were used to assess the linearity between the classes.
5. The libraries imported are datasets, matplotlib, sklearn, collections, wordcloud, nltk, torch and transformers.

6 Dataset

The dataset used is a Twitter emotions dataset obtained from https://huggingface.co/datasets/dair-ai/emotion, accessed via "load_dataset("emotion")". The version used in

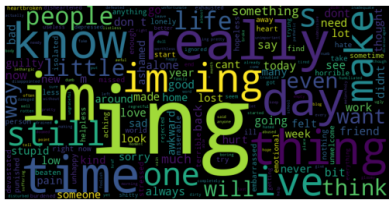
this study is normalised allowing an efficient use of computational resources. Coming in 16000 texts in the training set and 2000 in the validation and testing sets, the labels are:

- 0: sadness
- 1: joy
- 3: anger
- 4: fear
- 5: surprise

Some examples of text include:

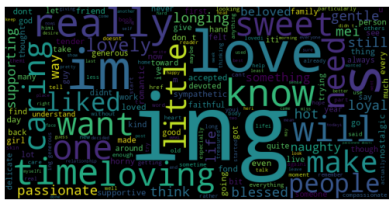
- "i feel pathetic most of the time" (0)
- "i talk to dogs as i feel they cannot understand words but they can read emotions and know how to be supportive I decided I should go home" (2)

Ambiguities between different emotions make this dataset challenging to work with. However, the dataset is suitable for this task due to labels being close to Ekman's model. The data first appeared in Saravia et al., 2018.

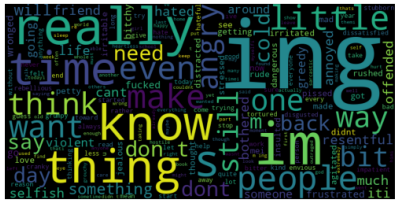


Most common words for "sadness" class

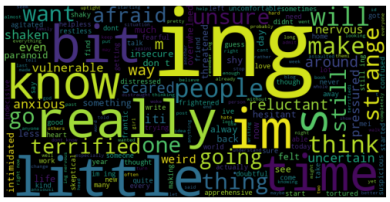
Most common words for "joy" class



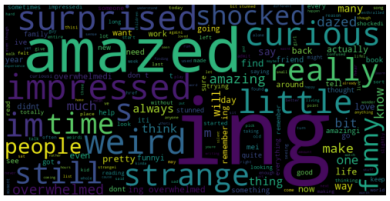
Most common words for "love" class



Most common words for "anger" class



Most common words for "fear" class



Most common class for "surprise" class

6.1 Dataset preprocessing

Tokenization was done on white spaces. When the word cloud was visualized, the presence of certain words was unique for certain labels. Hence individual words were taken as tokens. Stop words were removed from the vocab. The text had been normalized in advance.

7 Baselines

The baselines used were paired with vectors obtained with a TF-IDF vectorizer and CountVectorizer. First chosen to be rule-based, random, then linear. The rule-based baseline was used to analyze how "stereotypical" they were. The Random and Majority models were used to inform how imbalanced labels would influence. The linear models were used to understand the overall relationship between the words and the labels.

8 Results, error analysis

Due to its performance, the SVM classifier with CountVectorizer is chosen for comparison with the primary uncased BERT model, scoring 93% across all metrics.

Weighted:	Precision	Recall	F1-score	Accuracy
Rule-based	26%	33%	21%	33%
CV + Random	25%	24%	24%	24%
CV + Majority	12%	35%	18%	35%
CV + Naïve Bayes	77%	77%	73%	77%
CV + SVM	89%	88%	88%	88%
TF-IDF + Random	25%	25%	25%	25%

TF-IDF + Majority	12%	35%	18%	35%
TF-IDF Naïve Bayes	73%	65%	56%	65%
TF-IDF SVM	1%	11%	2%	11%

Error Analysis

Rule-based classifier

- "i didn't feel humiliated" misclassified as 1 but true label is 0
- "i can go from feeling so hopeless to so damned hopeful just from being around someone who cares and is awake" misclassified as 4 but true label 0 (due to mention of word "hopeless")

Random classifier with CountVectorizer

- "i am feeling grouchy" misclassified as 4 but true label is 0
- "i am ever feeling nostalgic about the fireplace i will know that it is still on the property" misclassified as 4 but true label is 1

Majority-based classifier with CountVectorizer

- "i didn't feel humiliated" misclassified as 1 but true label 0
- "i can go from feeling so hopeless to so damned hopeful just from being around someone who cares and is awake" misclassified as 1 but true label is 0

Naïve Bayes with CountVectorizer:

- "i cant walk into a shop anywhere where i do not feel comfortable" misclassifies as 1 but true label is 4
- "i felt anger when at the end of a telephone call" misclassified as 0 but true label 3

SVC with CountVectorizer

- "i don t feel particularly agitated" misclassified as 3 but true label is 4
- "im not sure the feeling of loss will ever go away but it may dull to a sweet feeling of nostalgia at what i shared in this life with my dad and the luck i had to have a dad for years: misclassified as 1 but true label 0

Random classifier with TF-IDF

- "im feeling rather rotten so im not very ambitious right now" misclassified as 3 but true label 0
- "i left with my bouquet of red and yellow tulips under my arm feeling more optimistic than when i arrived" misclassified as 3 but true label 1

Majority-based classifier with TF-IDF

- "im feeling rather rotten so im not very ambitious right now" misclassified as 1 but true label is 0
- "i never make her separate from me because i don t ever want her to feel like i m ashamed with her" misclassified as 1 but true label is 0

Naïve Bayes with TF-IDF

- "i cant walk into a shop anywhere where i do not feel uncomfortable" misclassified as 1 but true label is 4
- "i felt anger when at the end of a telephone call" misclassified as 1 but true label 3

SVM with TF-IDF

- "im feeling rather rotten so im not very ambitious right now" misclassified as 4 but true label is 0
- "i never make her separate from me because i don t ever want her to feel like i m ashamed with her" misclassified as 4 but true label is 0

BERT

- "i explain why i clung to a relationship with a boy who was in many ways immature and uncommited despite the excitement i should have been feeling for getting accepted into the masters program at the university of virginia" misclassified as 2 but true label is 1
- "i feel if i completely hated things i d exercise my democratic right speak my mind in what ever ways possible and try to enact a change" misclassified as 0 but true label is 3.

Overall, the same sentences were misclassified by the baseline models due to the ambiguities mentioned. BERT can capture these effectively, making wrong predictions only where the context dramatically shifts.

9 Lessons learned and conclusions

In this study, we have pursued a range of 9 baselines in decreasing order of randomness to establish a benchmark for our primary transformer's model. The corpus has been understood to have orientations towards certain emotions based on the way specific words are used. We have conducted an error analysis for each

baseline model and concluded that ambiguities between the classification of certain emotions were far too complex to be captured by traditional ML algorithms. However, an uncased BERT transformer model can capture the casual tones in texts, although drastic shifts in context play a weakness in this model. Due to the primary model outperforming the baseline models across several metrics, we conclude that our hypothesis regarding the suitability of an uncased BERT for multi-class sentiment detection is proven correct. Among the baseline models, the SVM classifier with TF-IDF vectorizer is proven to be a close second.

Due to time and computational resource constraints, experimentations with FastText embeddings were not completed, the rationale behind which is its pr-etrained attributes and its ability to stem unknown words down to its roots. This experimentation is recommended for future work. Additionally, experimentations with POS tagging and varying n-grams are also encouraged to gain insights on different performances.

References

Sharma, T., Diwakar, M., Singh, P., Lamba, S., Kumar, P., & Joshi, K. (2021). *Emotion Analysis for predicting the emotion labels using Machine Learning approaches*. In 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) (pp. 1-6). Dehradun, India: IEEE.

Singh, A., & Kumar, S. (2023). *A Comparison of Machine Learning Algorithms and Transformer-based Methods for Multiclass Sentiment Analysis on Twitter*. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-9). Delhi, India: IEEE.

Mahimaidoss, N. K., & Sathianesan, G. W. (2024). *Emotion Identification in Twitter Using Deep Learning Based Methodology*. Journal of Electrical Engineering & Technology, 19(6), 1891–1908.

Bouazizi, M., & Ohtsuki, T. (2019). *Multi-class sentiment analysis on twitter: Classification performance and challenges*. Big Data Mining and Analytics, 2(3), 181-194.

Rao, Y., Li, Q., Wenyin, L., Wu, Q., & Quan, X. (2014). *Affective topic model for social emotion detection*. Neural Networks, 58, 29-37.

Yassine, M., & Hajj, H. (2010). *A Framework for Emotion Mining from Text in Online Social Networks*. In 2010 IEEE International Conference on Data Mining Workshops (pp. 1136-1142). Sydney, NSW, Australia: IEEE.

Chawla, S., & Mehrotra, M. (2018). *An Ensemble-Classifier Based Approach for Multiclass Emotion Classification of Short Text*.

Wikarsa, L., & Thahir, S. N. (2015*). A text mining application of emotion classifications of Twitter's users using Naïve Bayes method*. In 2015 1st International Conference on Wireless and Telematics (ICWT) (pp. 1-6). Manado, Indonesia: IEEE.

Sahu, L., & Shah, B. (2022). *An Emotion based Sentiment Analysis on Twitter Dataset.* In 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET) (pp. 1-4). Bhopal, India: IEEE.

Mossad, N., Mohamed, Y., Fares, A., & Zaky, A. B. (2023). *Arabic text sentiment analysis and emotion classification using transformers*. In 2023 11th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC) (pp. 131-137). Alexandria, Egypt: IEEE.

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., & Chen, Y.-S. (2018). CARER: *Contextualized Affect Representations for Emotion Recognition*. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3687-3697). Brussels, Belgium: Association for Computational Linguistics.