Transformer based ensemble for emotion detection

Aditya Kane* and **Shantanu Patankar***
Pune Institute of Computer Technology, Pune
{adityakane1, shantanupatankar2001}@gmail.com

Sahil Khose and **Neeraja Kirtane**
Manipal Institute of Technology, Manipal
{sahilkhose18, kirtane.neeraja}@gmail.com

Abstract

Detecting emotions in languages is important to accomplish a complete interaction between humans and machines. This paper describes our contribution to the WASSA 2022 shared task which handles this crucial task of emotion detection. We have to identify the following emotions: sadness, surprise, neutral, anger, fear, disgust, joy based on a given essay text. We are using an ensemble of ELECTRA and BERT models to tackle this problem achieving an F1 score of 62.76%. Our codebase [1] and our WandB project[2] is publicly available.

1 Introduction

Even after engineering a 175B parameter language model like GPT-3 (Brown et al., 2020) we are far from artificial general intelligence. Emotion is a concept that is challenging to describe. However, as human beings, we understand the emotional effect that situations could have on other people. It is interesting to see how we can infuse this knowledge into machines. This work explores whether it is possible for machines to map emotions to situations consciously. Emotion in text has been studied for a while and has given interesting insights. The dataset that we are using is an extended version of the (Ekman, 1992) dataset. Our team, MPA_ED, participated in the WASSA 2022 Shared Task on Empathy Detection and Emotion Classification, Track 2: Emotion Classification (EMO), which consists of predicting the emotion at the essay level. This paper has the following contributions:

We propose three new datasets generated using various sampling techniques which overcome the class imbalance. We present our ensemble based solution consisting of multiple ELECTRA and BERT (Devlin et al., 2018) models to solve the emotion classification task. We provide a detailed analysis of the performance of the cluster of models and reflect on the shortcomings of the models as well as the dataset generated that affected the performance.

2 Related Work

Emotion detection and sentiment analysis has been an extensive research topic since the inception of natural language processing. It has been studied in great detail by faculties of both computer science and neurobiology (Okon-Singer et al., 2015). Murthy and Kumar (2021) presents an extensive review of the modern emotion classification techniques. The work by Alhuzali and Ananiadou (2021) remains the current state-of-the-art on emotion classification on the renowned SemEval dataset (Mohammad et al., 2018). BERT remains the best performer on the GoEmotions dataset (Demszky et al., 2020)

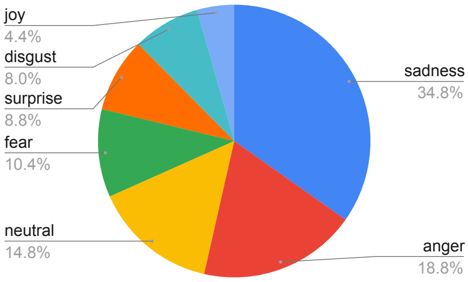


Figure 1: Class distribution in emotions

3 Data

The dataset consists of 1860 data points. Each data point has an essay and its emotion. The emotions are classified into seven types: anger, disgust, fear, joy, neutral, sadness, and surprise. The validation and test split has 270 and 525 data points respectively. The classes for the training data expresses high imbalance, as shown in Fig 1 . Here

*First authors

[1] https://bit.ly/WASSA_shared_task

[2] https://wandb.ai/acl_wassa_pictxmanipal/acl_wassa

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 250 - 254
May 26, 2022 ©2022 Association for Computational Linguistics

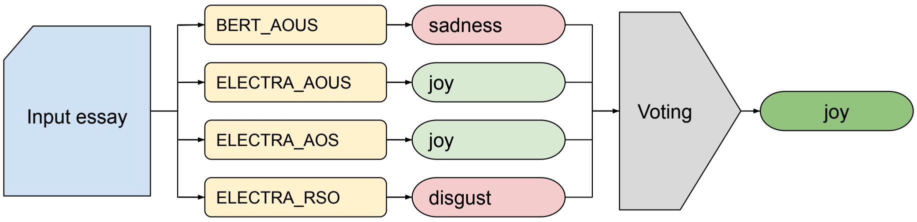


Figure 2: Ensemble pipeline

we see that the emotion "sadness" has the maximum number of data points, whereas "joy" has the least number of data points. The distribution is highly skewed and hence data augmentation is required to mitigate that. We performed basic preprocessing like removing punctuation, numbers, multiple spaces, and single line characters.

To overcome the class imbalance, GoEmotions dataset is used, which is a similar dataset with 27 emotions. We suggest three data augmentation techniques using the dataset described as follows:

- **Augmented Over-UnderSampling (AOUS)**: If X denotes the number of data points per class, in this method, if the data points in a particular class are greater than X, we undersample the data by randomly removing the essays. Otherwise, the data is oversampled by simply adding Reddit comments with maximum lengths from GoEmotions dataset (sorted by lengths) (Fig 3). As the average length of comments in GoEmotions dataset is 12 and average length of essays in WASSA dataset is 84, the comments with maximum length are chosen for oversampling. We take X as 400 in our experiments.
- **Random synthetic oversampling (RSO)**: We observe a significant difference in the average comment length of GoEmotions dataset and the average essay length in the WASSA dataset. To avoid disturbing the length distribution of the WASSA dataset after oversampling, we create synthetic essays by concatenating multiple random comments with same emotion (Fig 4). We match the distribution of lengths of the synthetically generated essays from GoEmotions dataset with the distribution of the original dataset using “Systematic Sampling.” We eliminate the deficit in each class by adding synthetically generated essays.
- **Augmented Oversampling (AOS)**: X denotes the highest number of data points per class. If the number of data points is less than X, the data is oversampled by adding comments from GoEmotions dataset with the highest lengths. (Fig 3)

Model	**Dataset**	**macro F1**
BERT$_{base}$	AOUS	59.19%
ELECTRA$_{base}$	AOS	58.94%
ELECTRA$_{base}$	RSO	59.06%
ELECTRA$_{base}$	AOUS	59.67%
Ensemble	val	62.76%
	test	53.41%

Table 1: Validation metrics

The data distribution post augmentation is balanced with number of samples in AOS, RSO and AOUS datasets equal to 4528, 4828 and 2800 respectively.

4 System Description

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is a transformer-based (Vaswani et al., 2017) language model developed by Google.

ELECTRA (Clark et al., 2020) is a variation of BERT, having a different pre-training approach. It requires less compute time compared to BERT.

We performed ablations with many of the present well-known language models — ALBERT (Lan et al., 2019), XLNET (Yang et al., 2019), RoBERTa (Liu et al., 2019) and found BERT and ELECTRA to perform the best.

5 Ensemble Methods

We conducted extensive experimentation and observed some models to perform substantially better than others. We shortlisted the models based on the validation F1-score. We decided to ensemble these models for better performance. We shortlisted four models and used majority voting as our ensemble method: BERT with AOUS, ELECTRA with AOS, ELECTRA with RSO, ELECTRA with AOUS.

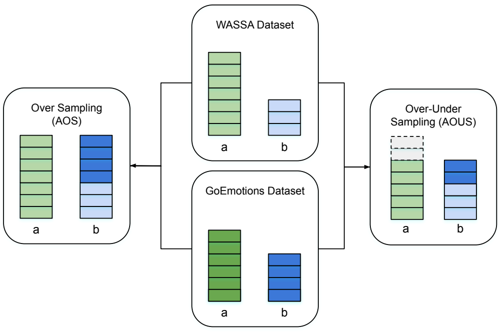


Figure 3: AOS and AOUS

We used the ensemble of the models in Table 1. The confusion matrices of each of these models are shown in Fig 5. The confusion matrix of the resultant ensemble is shown in 6. Note that all confusion matrices are normalized by the number of true samples in each class of the evaluation dataset. We deduce the following observations:

1. When the true label is "disgust," all models confuse the emotions "anger" and "disgust". All models have below average performance on "anger" and "disgust".
2. Models trained on AOUS dataset (c, d in Fig 5) are less prone to confusion in multiple close classes like "disgust", "fear" and "sadness" .
3. The emotions "anger" and "disgust" do not benefit from the ensemble, whereas "fear" suffers a bit. However we observe, the emotions "neutral", "sadness" and "surprise" experience significant gains from this process.

6 Experiments and Results

Our training setup was fairly straightforward. Language model backbone followed by fully connected layer and Softmax is used. CrossEntropy loss was used. We employed the Adam optimizer with $1e{-}5$ learning rate and batch size of 8. We fixed the seed for `numpy` and `torch` to 3407.

Some of the observations made during our extensive experimentation is as follows:

1. **Batch size 8 outperforms larger batch sizes**: We observed improvements across all models and datasets using a batch size of 8 over 32 or 64. We speculate this is because smaller batch size helps in generalization as the stochasticity of individual batches increase.
2. **ELECTRA fine-tuned on the AOUS dataset outperforms other models**: ELECTRA performs better than BERT for all our augmented datasets. We believe models finetuned on AOUS dataset perform better because AOUS dataset has 400 labels per class, making the dataset balanced while limiting the adulteration induced by the GoEmotions dataset.

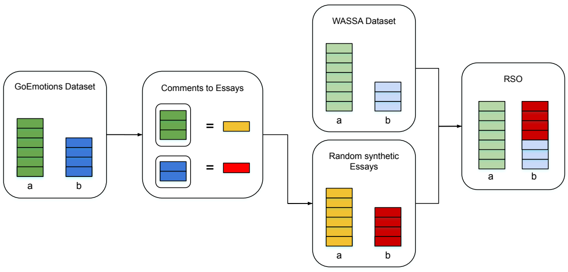
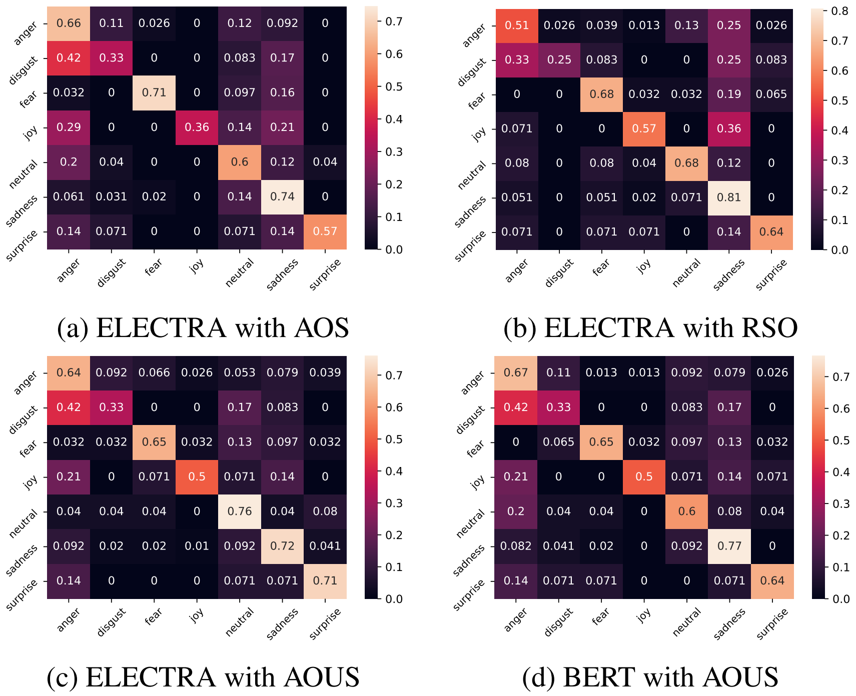


Figure 4: RSO

3. **Multi-task learning has poor performance**: We experimented with multi-task learning where empathy and distress tasks (Track 1) and emotion classification task (Track 2) were trained together with a shared backbone. We observed that the training was erratic, and the training loss did not converge.
4. **Models are sensitive to data imbalance**: When trained on the original dataset with class imbalance, the model is biased towards predicting classes with more training samples. We used data augmentation techniques mentioned in Section 3 to tackle this issue. After handling the class imbalance with data augmentation, the macro F1 score of the BERT model increased from 32.19% to 59.19%.
5. **Emotion "joy" vs "surprise"**: These are the only two positive emotions in the dataset. We expected all of the models to confuse these emotions as they are semantically similar. However, to our "surprise", we observed the models performed spectacularly on these two emotions. We think this is because "surprise" and "joy" have distinct appearances in the corpus. "surprise" examples have some sort of exclamation or a questioning tone in them. This leaves us with "joy", which happens to be the only positive emotion along with "surprise" in the corpus.
6. **Randomly created synthetic essays provide little understanding**: We observed the model trained on RSO augmented data often predicts



(a) ELECTRA with AOS

(b) ELECTRA with RSO

(c) ELECTRA with AOUS

(d) BERT with AOUS

Figure 5: Confusion matrices of our models.

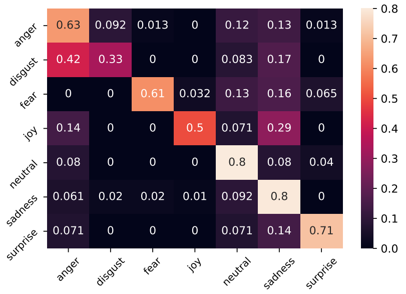


Figure 6: Confusion matrix of our final ensemble.

other emotions as "sadness" (Fig 5 (b)). We speculate this is because there was no addition of synthetically generated data for the "sadness" class as it is the largest class. We further hypothesize the synthetic data in RSO, being randomly concatenated, disrupts the context of the entire essay as a whole. However, we still use the model in our final ensemble since it performed well amongst the population. We think this occurs due to multiple factors being simultaneously at play. Further investigation is a promising future direction.

The validation confusion matrix of all the four models are displayed in Fig 5 and their results in Table 1. We present the following statistics. (True Positive (TP), standard deviation (σ), mean (μ))

1. **The highest TP** μ is for **"sadness"** and **"fear"** emotion with 76 and 67.25 values respectively. Interestingly both of these emotions also have the least TP σ with 3.92 and 2.87 values respectively.
2. **The least TP** μ is for **"disgust"** and **"joy"** emotion with 31 and 48.5 values respectively. "joy" also accounting for the highest TP σ with 8.81 value which infers that all the models are agreeing on different datapoints to classify as "joy". Whereas "disgust" has one of the least TP σ with 4.0 just following "fear" and "sadness", this suggests that all the models are able to agree on a very small sample space of the class data to be classified as "disgust".

7 Conclusion

In this work, we have explored an application of BERT and ELECTRA as a means to the task of emotion classification. Various data sampling techniques were used to overcome the large imbalance in data. In the end the best metrics were achieved by using majority voting of the 4 best models as an ensemble. We foresee multiple future directions, including multi-task learning of multiple tasks with a shared backbone, pretraining on the entire GoEmotions dataset, as well as studying and rectifying spurious behaviour of "anger" and "disgust" labels.

References

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Ashritha R Murthy and K M Anil Kumar. 2021. A review of different approaches for detecting emotion from text. *IOP Conference Series: Materials Science and Engineering*, 1110(1):012009.

Hadas Okon-Singer, Talma Hendler, Luiz Pessoa, and Alexander J. Shackman. 2015. The neurobiology of emotion–cognition interactions: fundamental questions and strategies for future research. *Frontiers in Human Neuroscience*, 9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.