Team: Ana Bernal, Leah Tesfa, Andrés Mondragón **Project Mentor TA:** Jiahao Huang

Embeddings Evaluation for Emotion Classification in Social Media Text

1) Abstract

In our project, we tackle the challenge of emotion classification in textual data sourced from platforms like social media. The motivation behind our project stems from the growing importance of understanding human emotions in online interactions. By automating the process of emotion classification, our system can assist in various applications, including sentiment analysis, social media monitoring, and even the development of empathetic AI systems capable of understanding and responding to emotions effectively. We primarily work with a Kaggle dataset, *Emotions* [1], which provides us with a collection of English Tweets and six emotion labels: sadness (0), joy (1), love (2), anger (3), fear (4), and surprise (5). While this was a sufficient baseline, we dedicated our first contribution to supplementing this dataset by gathering more data that reflects feelings of loneliness (6), jealousy (7), and awkwardness (8). This allowed us to account for more nuanced emotions. As our second contribution, we evaluated the effectiveness of various vectorization and embedding techniques when used with classifiers trained on this dataset in the task of emotion classification. We calculated evaluation metrics for two traditional vectorization models and four Transformer models and evaluated their performance when used with two classification models. Perhaps most interestingly, we find that the most commonly used Transformer model, BERT, did not outperform other models. This is likely due to the high volume of data that its architecture requires. For this reason, we find that DistilBERT, a distilled version of BERT, performed better in our classification task using both SVM and Logistic Regression.

2) Introduction

This project focuses on the task of using supervised Machine Learning techniques to classify emotions in textual data. Emotions play a pivotal role in communication, and understanding them in online interactions can provide valuable insights into user sentiments, behaviors, and societal trends. Our system takes as input a dataset with text data that expresses diverse emotions from platforms like Twitter and Reddit, and aims to accurately classify the underlying emotions or feelings conveyed in the text. The output of our system is a classification label indicating the predominant emotion or feeling expressed in the text.

Initially using a Kaggle dataset, *Emotions [1]*, that provides us with a collection of English Tweets demonstrating six emotions (anger, fear, joy, love, sadness, and surprise), we increased our dataset to encompass three more feelings. To achieve this, we scraped text posts made on subreddits for emotions of loneliness, jealousy and awkwardness. After preprocessing to ensure our two datasets are compatible, we use classification models to accurately predict emotions/feelings associated with the text. Our labeled dataset containing these text samples along with corresponding emotion or feeling labels is used for the training and testing. With this dataset we train our classification models to learn the relationships between input text features and the associated emotion labels. To evaluate the performance of our proposed system, we employ metrics such as accuracy and F1 score.

3) Background

One prior work we consider relevant for this project is “*Finding Good Representations of Emotions for Text Classification*” *[2]*, by Ji Ho Park. In this paper, the author proposed improved representations of text, in both word and sentence levels, called emotional word vectors (EVEC). The representations are learned from a convolutional neural network model with an emotion-labeled corpus. The data used comes from Twitter and uses hashtags in tweets to automatically annotate emotions, a methodology known as distant supervision. The emotion labels used came from a highly cited psychology paper, “*Emotion knowledge: Further exploration of a prototype approach*” *[3]* by Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O’Connor.

Ultimately, we found a dataset that utilizes very similar data to the one in the paper. *Emotions [1]* is a dataset on Kaggle that provides us with a collection of English Tweets annotated with six fundamental emotions: anger, fear, joy, love, sadness, and surprise. Given the results from other papers mentioned in the paper by Park, such as *[4]*, the distant supervision method of using hashtags has already been proven to provide reasonably relevant emotion labels. Yet, while the six emotions identified in the dataset used in the Park paper and the dataset from Kaggle can provide a fundamental understanding of human emotions, we believe that there could be additional emotions to account for when it comes to social media data.

For this reason, we decided to expand the dataset to account for new emotions and also experiment with different vectorization techniques to evaluate their performance when being used as inputs for classification models.

4) Summary of Our Contributions

In our goal to advance emotion classification within textual data sourced from social media platforms, we are making two significant contributions. Firstly, we recognize the inherent complexity and nuance of human emotions, acknowledging that the conventional six emotion labels may not fully capture the breadth of emotional experiences expressed online. To address this limitation, we are expanding the emotion labels by incorporating additional feelings, thus enriching the emotional spectrum considered in our classification task. Additionally, we are evaluating and identifying the reasons for the effectiveness of various vectorization and embedding techniques when used as input for classifiers to enhance the accuracy and robustness of emotion classification models. Through these contributions, we aim to improve the applicability and performance of emotion classification in real-world scenarios.

5) Detailed Description of Contributions

1. Contribution 1:

Part of our project required us to gather more data since we believed that the current six emotion labels sadness (0), joy (1), love (2), anger (3), fear (4), and surprise (5), were insufficient to account for other more nuanced emotions. Before anything, we should emphasize the difference between emotion and feeling: “Emotions originate as sensations in the body. Feelings are influenced by our emotions but are generated from our mental thoughts.” *[5]* Given that we believe that accounting for feelings in addition to emotions would improve the applicability of our analysis in the context of human-machine communications, we identified other feelings that are significantly more prevalent in our current social media landscape. Keeping in mind that the original six emotions were identified in a psychological study from 1987 *[3]* and our chosen prior work essentially only considers these six emotions, our contribution was to expand the labels by adding three more feelings: loneliness (6), jealousy (7), and awkwardness (8). To do this, we leveraged the Reddit API to scrape text posts and comments made on the subreddits “r/loneliness”, “r/retroactivejealousy”, and “r/socialskills”. Due to the changes in the Twitter API, which makes it hard to access Twitter data, we naturally gravitated towards Reddit, since many users also express their emotions through the posts and comments they make and the platform allows for a larger character limit so people are more likely to express their thoughts more. Similarly to how distant supervision was used in the Twitter data where the proxy for knowing whether a tweet talked about a particular emotion was the use of a hashtag, our proxy for the feeling was the subreddit name. Thus, for loneliness (6) we used data from “r/loneliness”, for jealousy (7) we used “r/retroactivejealousy”, and for awkwardness (8) we used “r/socialskills”. The subreddit descriptions gave us confidence that posts made within them would talk about the associated feelings we identified. While the larger character limit for Reddit posts would be beneficial for other NLP tasks where computational resources are less constrained, in our case this would create inconsistency in the data. The Twitter data we sampled and used for our models has a 328 character limit for tweets, yet, “in most cases, the text content of a Tweet can contain up to 280” characters. [6] Thus, in order to make both the process of tokenizing and embedding text more computationally feasible and the results more comparable, we decided to only scrape comments for posts rather than the entire post body written by the original poster. This made our data collection process more efficient given that scraping comments was faster and there are more comments than posts. Additionally, we filtered comments so that we only include comments that have 280 characters or less. We ended up with a dataset containing 9,000 values, 1,000 text values per each emotion.

2. Contribution 2:

The objective of this contribution is to evaluate the effectiveness of various vectorization techniques when used with classifiers in the task of emotion classification. This analysis will cover both Twitter tweets, which include six fundamental emotions, and Reddit comments, which incorporate three additional emotional categories. The specific aims of this contribution are to address the following key questions:

1. How do different vectorization techniques—namely TF-IDF, Word2Vec, BERT, DistilBERT, ELECTRA—perform when integrated with traditional classifiers such as Support Vector Machines (SVM) and Logistic Regression?
2. Upon determining the most effective architecture, what specific components or features of this architecture contribute to its superior performance over other vectorization and classification combinations?

5.2 Methods and Results

The first analysis will be conducted on vectorization methods and traditional classifiers. In this analysis, six feature extractor methods will be utilized:

Feature Extractor	Purpose

TF-IDF Vectorization	Measures word importance based on frequency in a document and rarity across all documents. Able to highlight distinctive words that may indicate specific emotions.
Word2Vec Vectorization	Represents words as vectors based on contextual similarity in a text corpus. Able to capture semantic relationships and potentially improve the detection of emotional subtext.
BERT Embeddings [7]	BERT is a pretrained model that uses two objectives: masked language modeling (MLM) and Next Sentence Prediction (NSP). It creates deep contextual representations by considering both the left and right context of each word in a sentence.
DistilBERT Embeddings [8]	DistilBERT was introduced as a “smaller, faster, cheaper, lighter” and distilled version of BERT. It is a Transformer model trained by distilling BERT base, which has 40% less parameters than BERT, runs 60% faster while preserving over 95% of BERT’s performance.
ELECTRA Embeddings [9]	Unlike BERT, ELECTRA uses a different pre-training objective called replaced token detection. This involves replacing a small fraction of input tokens with potential alternatives and training the model to distinguish the replaced tokens from the original ones. Thus, ELECTRA embeddings also capture contextual information but with a more efficient training process.

For classification tasks with Transformers models, such as text classification or sentiment analysis, it's common to use the pooled representation for making predictions. This is because the pooled representation is derived from the output of the last layer of the models. By using it, we effectively summarize the entire input sequence in a single vector, which can then be fed into a classifier as input to make predictions. The vectors produced by each of these extractors will then be inputted into each of the following two classification models, in order to create a combination between each feature extractor and each model.

Classifier	Purpose
Logistic Regression	An efficient baseline for binary and multiclass classification.
SVM	Able to handle high-dimensional data from text, making it robust in distinguishing between emotional categories, even when data is not linearly separable.

Each model will be trained on the training set, which will comprise 70% of the sample data. Hyperparameter tuning, such as the number of maximum iterations for logistic regression, will be conducted as needed. Accuracy and F-1 score will be used to evaluate the models. Our results are summarized below.

Feature Extractor	Classification Model	Accuracy	F1 Score
TF-IDF Vectorization	Logistic Regression	78.810	78.833
	SVM	78.571	78.440
Word2Vec Vectorization	Logistic Regression	67.571	67.602
	SVM	70.333	70.373
BERT Embeddings	Logistic Regression	85.333	85.585
	SVM	78.619	79.637

DistilBERT Embeddings	Logistic Regression	**86.333**	**86.370**
	SVM	**82.857**	**83.210**
ELECTRA Embeddings	Logistic Regression	85.095	85.209
	SVM	79.952	80.434

Our results show that the combination of DistilBERT embeddings and the Logistic Regression classifier outperforms all other models, achieving an accuracy of 86.33% and an F1 score of 86.37%. This can primarily be attributed to several components of DistilBERT's architecture. In general, transformer models like BERT, ELECTRA, and DistilBERT use Transformer layers to capture contextual and semantic nuances of words in a sentence. Unlike traditional methods like TF-IDF, which simply consider word frequency, or Word2Vec, which capture semantic similarities but lack deep contextual understanding, Transformer models process the entire sequence of words and contextualize the text's meaning. This allows them to differentiate between emotions that are contextually similar but semantically distinct, such as 'love' and 'joy' and 'fear' and 'surprise'.

Additionally, Transformer models leverage an attention mechanism that allows them to weigh the importance of different words in a sequence. This is especially useful for emotion detection tasks because it helps models to focus on contextually significant words and phrases that convey specific emotions [11].

Our results mostly align with this discussion: Transformer models typically performed better than Word2Vec and TF-IDF using both Logistic Regression and SVM as classifiers. While we hypothesized that BERT would outperform other Transformer models, we see that its scores fall slightly behind DistilBERT. BERT's relatively low performance, when compared to DistilBERT, may be likely due to the fact that its complex architecture requires extensive data to capture diverse language patterns and nuances effectively, and in our task, data was relatively limited.

For this reason, we decided to explore DistilBERT, a distilled version of BERT that can handle lower amounts of data more effectively. As we hypothesized, DistilBERT's performance was higher than that of any other model. DistilBERT's architecture provides a balance between performance and computational efficiency. It leverages the concept of knowledge distillation, which is a compression technique in which a small model is trained to reproduce the behavior of a large model, in this case BERT [8]. By training on the same corpus as BERT, DistilBERT inherits the robust language representations learned by the larger model but in a more efficient form, which is crucial when dealing with limited data sets.

Similarly to BERT, ELECTRA, a non-distilled model, also performed relatively well, with a Logistic Regression accuracy of 85.09% and an SVM accuracy of 79.95%. This is likely due to ELECTRA's unique training objective. The model is trained as a discriminator that predicts whether each token in a sequence was replaced by a generation model, leading to more efficient learning representations. This "replaced token detection" task allows ELECTRA to leverage more input during training, which can lead to better performance when compared to models like BERT [9].

It's also interesting to note that despite its simple architecture, TF-IDF's performance was only slightly lower than that of ELECTRA, BERT, and DistilBERT, with a 78.81% Logistic Regression accuracy and a 78.83% SVM accuracy. TF-IDF's relatively strong performance can be attributed to its robustness in capturing term frequency and document specificity. Its vectors are sparse and interpretable, making them less prone to overfitting compared to dense embeddings like those used in Transformers [12]. While TF-IDF lacks the deep contextual understanding of transformer models, it clearly serves as a competitive baseline, particularly when data and computational resources are more limited.

During our initial run of the Transformer models, we utilized the base models for BERT, DistilBERT, and ELECTRA. This yielded unsatisfactory results which we identified resulted from two issues: 1) The text data from Reddit posts being too large. 2) The Transformer models not being fine-tuned on relevant data for our specific task. Addressing issue 1) was done during the data pre-processing stage, while issue 2) was tackled by leveraging the Hugging Face's Transformer repository where we identified the aforementioned models but pre-trained on an emotion classification task. This allowed us to transfer the knowledge that these models learned during pre-training onto our task and improve the generalization of our models since the models have already learned the nuances of this specific text data. The pre-trained models were the following:

- BERT: bhadresh-savani/bert-base-uncased-emotion [13]
- DistilBERT: Rahmat82/DistilBERT-finetuned-on-emotion [14]
- ELECTRA: mudogruer/electra-emotion [15]

The dataset used to pretrain these models was the same as the one we used as the base, before adding the extra Reddit data.

6) Compute/Other Resources Used

For this project we found that the use of the GPU and smaller batch size was essential due to the computational demands of processing Transformer embeddings with pre-trained models. Transformer models require a lot of computational power, especially when dealing with high dimensional data such as text data. Thus, we needed to leverage parallel processing from the GPU to speed up our processes. Processing data in batches was also needed because we needed to fit the data into the GPU memory as without this we were exceeding the available GPU memory.

7) Conclusions

Our project demonstrates the significant advantages of using advanced Transformer models, particularly DistilBERT, for emotion detection in social media text. By combining DistilBERT embeddings with a Logistic Regression classifier, we achieved an accuracy of 86.33% and F1 score of 86.37%, which highlights DistilBERT's ability to capture nuanced contextual and semantic information, essential for distinguishing between closely related emotions. The model's smaller size, using roughly 40% of the parameters of BERT, makes it significantly more efficient in terms of computational resources while maintaining high performance. This balance between efficiency and effectiveness is crucial, particularly when working with limited data and computational power. Interestingly, our findings also show that baseline vectorizations like TF-IDF also perform relatively well, achieving a Logistic Regression accuracy of 78.81% and an SVM accuracy of 78.83%. This suggests that TF-IDF serves as a practical and competitive baseline that doesn't require the extensive preprocessing, fine tuning, and even pretrained embedding integration necessary for Transformer models. While TF-IDF can be surprisingly effective and offers a simpler more resource-efficient approach, to achieve the highest possible performance, using Transformer models like DistilBERT is essential.

A potential direction of future improvement we could take is exploring ensemble methods and observing the effects of combining predictions from multiple models. In the future, we could also increase the sampling size of the data collected and perhaps use other sources for our data collection. For example, instead of Twitter and Reddit, we could also use Whatsapp or Facebook. Increasing our emotional spectrum with emotions that are more complicated or intertwined, as well as using more transformer models are some other steps we could take to increase the robustness of our experimental analysis.

Identification of emotions and text sentiment is vital in various spheres of life. Hence, later our models could be used by companies who can analyze their customer feedback with ease, or help content creators gauge their audience's reactions or even help virtual assistants be able to understand and respond appropriately to users’ emotional cues.

As mentioned throughout the paper, the largest setback is the computational power required to fully train Transformer models, which is why we decided to use pre-trained models from Hugging Face. Had we trained the models on data, we would have needed a larger corpus, which we would not be able to run efficiently given our limited resources. There are also environmental considerations linked to GPU use, which we would face if we were to train Transformer models with large amounts of data. Ethically, we understand that social media data, where users are free to express themselves and their opinions however they want, may be sensitive and not appropriate to use to train models that AI models can leverage to use for better machine-human interactions. Yet, we also believe that understanding certain emotions may be important for the purpose of enhancing and improving communication between people, where machines could aid in providing guidance on how to address certain tweets or texts based on the emotion that a user is trying to express.

References

[1] N. Elgiriyewithana, “Emotions,” Kaggle.com, 2024. https://www.kaggle.com/datasets/nelgiriyewithana/emotions (accessed Apr. 22, 2024).

[2] J. Park, “Finding Good Representations of Emotions for Text Classification,” 2018. Accessed: Apr. 22, 2024. [Online]. Available: https://arxiv.org/pdf/1808.07235.pdf

[3] P. Shaver, J. Schwartz, D. Kirson, & C. O'Connor, “Emotion knowledge: Further exploration of a prototype approach.Journal of Personality and Social Psychology”, 1987. Accessed: Apr. 22, 2024. Available: https://doi.org/10.1037/0022-3514.52.6.1061

[4] W. Wang, L. Chen, K. Thirunarayan and A. P. Sheth, "Harnessing Twitter "Big Data" for Automatic Emotion Identification," 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, Amsterdam, Netherlands, 2012, pp. 587-592, doi: 10.1109/SocialCom-PASSAT.2012.119.

[5] R. Allyn, “The Important Difference Between Emotions and Feelings,” Psychology Today, 2022. https://www.psychologytoday.com/us/blog/the-pleasure-is-all-yours/202202/the-important-difference-between-emotions-and-feelings (accessed Apr. 22, 2024).

[6] “Counting characters,” Twitter.com, 2024. https://developer.twitter.com/en/docs/counting-characters (accessed May 14, 2024).

[7] J. Devlin, M.-W. Chang, K. Lee, K. Google, and A. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” Available: https://arxiv.org/pdf/1810.04805

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *arXiv.org*, Oct. 02, 2019. https://arxiv.org/abs/1910.01108

[9] K. Clark, M.-T. Luong, G. Brain, Q. Le Google Brain, and C. Manning, “ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS.” Available: https://openreview.net/pdf?id=r1xMH1BtvB

[11] S. Patil, “Attention Mechanism in the Transformers - Sagar Patil - Medium,” Medium, Sep. 25, 2023. https://medium.com/@sagarpatiler/attention-mechanism-in-the-transformers-fd067df25ea (accessed May 14, 2024).

[12] Pradeep, “Understanding TF-IDF in NLP: A Comprehensive Guide - Pradeep - Medium,” Medium, Mar. 21, 2023. https://medium.com/@er.iit.pradeep09/understanding-tf-idf-in-nlp-a-comprehensive-guide-26707db0cec5 (accessed May 14, 2024).

[13] “bhadresh-savani/bert-base-uncased-emotion · Hugging Face,” Huggingface.co, Apr. 20, 2023. https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion (accessed May 14, 2024).
[14] “Rahmat82/DistilBERT-finetuned-on-emotion · Hugging Face,” Huggingface.co, Apr. 20, 2023. https://huggingface.co/Rahmat82/DistilBERT-finetuned-on-emotion (accessed May 14, 2024).
[15] “mudogruer/electra-emotion · Hugging Face,” Huggingface.co, Apr. 20, 2023. https://huggingface.co/mudogruer/electra-emotion (accessed May 14, 2024).

Broader Dissemination Information:

Your report title and the list of team members will be published on the class website. Would you also like your pdf report to be published?
No.

(Exempted from page limit) Attach your midway report here, as a series of screenshots from Gradescope, starting with a screenshot of your main evaluation tab, and then screenshots of each page, including pdf comments. This is similar to how you were required to attach screenshots of the proposal in your midway report.

Project Check In

Contributors	Task	APRIL										MAY														
		S 21	M 22	T 23	W 24	Th 25	F 26	S 27	S 28	M 29	T 30	W 1	Th 2	F 3	S 4	S 5	M 6	T 7	W 8	Th 9	F 10	S 11	S 12	M 13	T 14	W 15
Andres, Leah, Ana	Run DistilBERT and evaluate						🟩	🟩	🟩																	
Andres, Leah, Ana	(Potentially) reduce max. sequence length of sample									🟩	🟩	🟩														
Andres, Leah, Ana	Run layer-based analysis of best performing transformer model												🟩	🟩	🟩											
Andres, Leah, Ana	Make results visualizations															🟩	🟩	🟩	🟩							
Andres, Leah, Ana	Paper and Video																		🟩	🟩	🟩	🟩				