FEAT: From Frequency-based Emotion Analysis to Transformers

AMS 560.01-02 / CSE 542.01 Final Report - Fall 2024

Priyanshu Jain, Omkar Shetty, and Jainam Rambhia

Introduction.

In today's digital era, where vast amounts of text data are generated incessantly, accurate sentiment analysis holds paramount importance across diverse domains.

This project is crucial for accurate sentiment analysis in today's data-driven world. It empowers businesses to make informed decisions and gain insights into customer satisfaction, public opinion, and socio-political dynamics. By leveraging advanced techniques and models, the project aims to improve the efficiency of sentiment analysis, enabling organizations to extract actionable insights from large textual datasets. Ultimately, its outcomes can enhance decision-making, drive business strategies, and provide valuable societal insights.

Through advanced techniques and state-of-the-art models, we aim to enhance the precision and efficacy of sentiment analysis. Our approach encompasses model architecture refinement, feature engineering, and ensemble learning techniques, utilizing baseline models like distilRoBERTa and distilBERT augmented with feedforward neural networks (FFNN) to elevate performance metrics such as F1 scores. Additionally, we explore hybrid methodologies, combining fine-tuned model embeddings with traditional machine-learning algorithms to extract richer feature sets and enhance predictive accuracy. By navigating ensemble models, feature fusion, and the integration of TF IDF with embeddings, we present a comprehensive approach to sentiment analysis with practical utility for businesses and researchers.

Background. Our work in sentiment analysis builds upon extensive research in natural language processing (NLP). Sentiment analysis, or opinion mining, finds applications from marketing to social media monitoring. Traditional methods like lexicon-based approaches had limitations. Recent advancements in deep learning, particularly transformer-based models like distilRoBERTa and distilBERT, revolutionize sentiment analysis. These models, derived from BERT, are fine-tuned on large datasets[4] for specific tasks, enhancing accuracy. Integration of feedforward neural networks further boosts performance.

Understanding transformer architectures and sentiment analysis methodologies is crucial. Transformer-based models like distilRoBERTa and distilBERT, derived from BERT, have significantly advanced sentiment analysis. Fine-tuning these models on domain-specific data enhances accuracy. Integration of feedforward neural networks further boosts performance. Familiarity with sentiment analysis techniques, including multi-class classification and feature fusion, provides essential context.

Data:

The dataset has been developed by Saravia et.al[1]. Tweets are extracted using appropriate hashtags and classified into one of the 6 tags: Sadness, Joy, Anger, Fear, Love, and Surprise. The dataset contains 416809 rows. [Fig1] displays the distribution of data. 80:20 split was performed for training and testing. We employed stratified sampling for equal training data of each class to prevent bias outcomes due to unequal distribution of data during training.

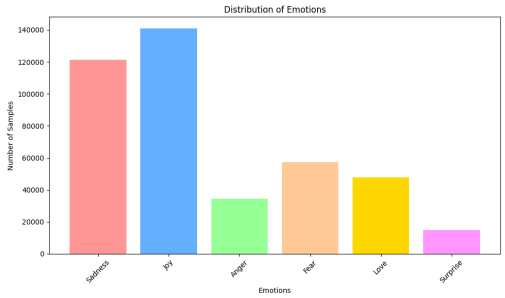

Fig1: Distribution of classes

Methods. In our endeavor to explore the efficacy of machine learning classifiers in emotion classification, we employed a meticulous methodology centered around fine-tuned DistilRoBERTa and DistilBERT embeddings and hybrid features incorporating TF-IDF representations. Initially, we trained two basic classifiers atop pre-trained DistilRoBERTa and DistilBERT models using a single-layer architecture. Leveraging a batch size of 16 and two epochs, we fine-tuned this model with a learning rate of 2e-5 and weight decay of 0.01. The dataset, sourced from English Twitter messages, comprised 416,809 rows, categorized into six basic emotions: anger, fear, joy, love, sadness, and surprise. Notably, we allocated 80% for training and 20% for testing, ensuring a robust evaluation of our models.[2]

The observed performance of distilRoBERTa being superior, we fine tuned and extracted 768-dimensional embeddings from the encoder of the DistilRoBERTa model, serving as features for subsequent machine learning models. Our selection of classifiers—Support Vector Classifier (SVC), RandomForestClassifier, HistGradientBoostingClassifier, and AdaBoostClassifier[5]—was strategic, aiming for diverse algorithms known for their effectiveness across various tasks. SVC, in particular, demonstrated marginal superiority over a FeedForward Neural Network (FFNN) baseline in terms of both accuracy and F1 score, highlighting its aptness for our classification task.

To enrich the feature space and potentially enhance classification performance, we augmented the DistilRoBERTa embeddings with TF-IDF representations. This decision was grounded in the rationale that TF-IDF captures textual nuances relevant to emotion classification, complementing the semantic embeddings from DistilRoBERTa. By fusing these hybrid features, we repeated the entire process, observing improved performance compared to utilizing DistilRoBERTa embeddings alone. Notably, the ensemble of DistilBERT and DistilRoBERTa models was also investigated, yielding promising results, with the best-performing model being the fine-tuned ensemble embeddings augmented with TF-IDF features and classified using SVC.

Our methodology [Fig3] operated under several assumptions, primarily relying on the assumption that the labeled Twitter messages accurately reflected the underlying emotions. We also assumed that distant supervision techniques used in generating the dataset introduced minimal noise, allowing for reliable model training and evaluation. Furthermore, our approach assumed that the DistilRoBERTa model, after fine-tuning, effectively captured emotional nuances in the text, facilitating robust feature extraction. By meticulously adhering to these assumptions and leveraging a systematic approach, we were able to navigate the intricacies of emotion classification and delineate optimal strategies for leveraging machine learning techniques in this domain.

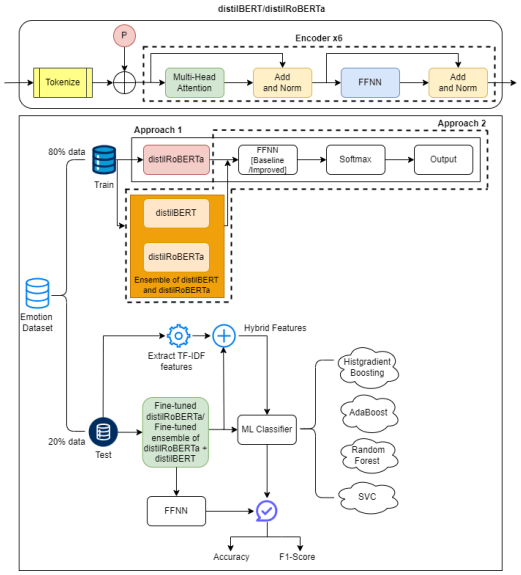

Fig3: Flowchart of the approach

Evaluation/Results.

We are using Backbone1 as distilRoBERTa and Backbone2 as Ensemble of distilRoBERTa and distilBERT for this section.

Model	Accuracy	F1
distilRoBERTa	0.9392	0.9032
distilBERT	0.9388	0.8991

Table1: Baseline Models

Model	Accuracy	F1
distilRoBERTa	0.9399	0.9081
distilBERT	0.9398	0.9035

Table2: Baseline Models with hybrid features

[Table1][Table2] DistilRoBERTa performs better than DistilBERT with a classifier head. This might be because of a more extensive pre-training process and architecture refinements in distilRoBERTa.

Backbone	Classification head	Accuracy	F1
Backbone 1	Improved FFNN	0.9393	0.9039
	SVC	0.9395	0.9041
	Random Forest Classifier	0.9288	0.8883
	HistGradient BoostingClassifier	0.9312	0.8944
	AdaBoostClassifier	0.6921	0.4811
Backbone 2	Improved FFNN	0.9406	0.9088
	SVC	0.9421	0.9102
	Random Forest Classifier	0.9295	0.8919
	HistGradient BoostingClassifier	0.9313	0.8945
	AdaBoostClassifier	0.6944	0.4905

Table3: Results of models without hybrid features

Backbone	Classification head	Accuracy	F1
Backbone 1	Improved FFNN	0.9409	0.9103
	SVC	0.9454	0.9125
	Random Forest Classifier	0.9295	0.8919
	HistGradient BoostingClassifier	0.9313	0.8945
	AdaBoostClassifier	0.6787	0.4143
Backbone 2	Improved FFNN	0.9468	0.9133
	SVC	0.9501	0.9174
	Random Forest Classifier	0.9296	0.8928
	HistGradient BoostingClassifier	0.9344	0.9004
	AdaBoostClassifier	0.689	0.4151

Table4: Results of models with hybrid features

From [Table3] and [Table4], it is evident that adding fused features help the models learn better. The only exception is the case of AdaBoost Classifier as it relies on sequentially fitting weak learners to instances, with subsequent learners focusing on instances misclassified by previous ones. However, TF IDF[6] fused features introduce a higher-dimensional, potentially more complex feature space compared to the original features. This complexity might make it harder for AdaBoostClassifier to effectively adjust its weighting of instances, leading to suboptimal performance as it struggles to adapt to the new feature representations.

[Fig4] shows that Backbone2 + SVC shows the most superior results in terms of Accuracy, Precision, Recall as well as F1 Score. This can be attributed to its ability to effectively delineate complex decision boundaries in the feature space created by the hybrid embeddings and TF IDF features. SVC excels in scenarios where data points may not be linearly separable, leveraging kernel functions to map them into higher-dimensional spaces where separation is feasible. This capability proves beneficial when dealing with the amalgamated features generated from the distilRoBERTa and distilBERT embeddings hybrid with TF IDF, as they might exhibit intricate relationships that require

non-linear decision boundaries for accurate classification.

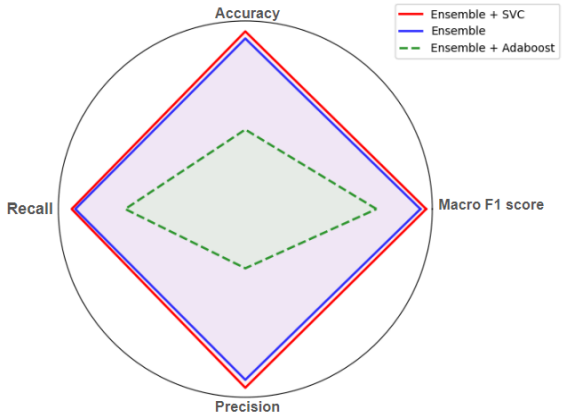

Fig4: Comparison of classification metrics for Backbone 2 and machine learning models

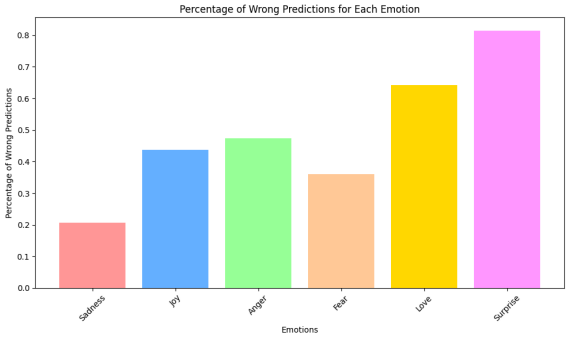

Fig5: Percentage Errors per emotion

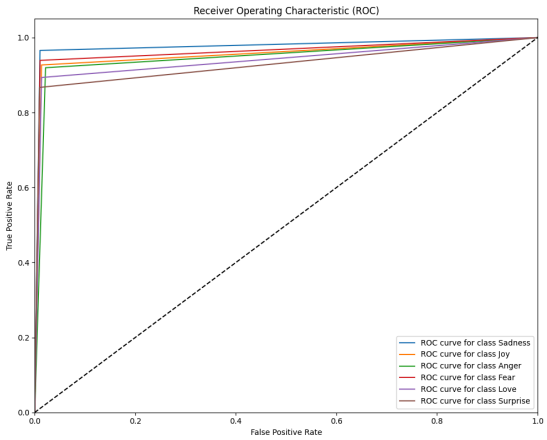

Fig6: ROC_AUC curve for all emotions for Backbone 2 + SVC

From [Fig5] and [Fig6], it is evident that it is comparatively difficult to classify the label “Surprise”. This can be attributed to the fact that “Surprise” labels are the minimum in number.

Detailed Comparison to Existing Tools & Traditional Methods:

- **Lexicon-Based Models:** Traditional lexicon-based sentiment analyzers, such as AFINN or VADER, are highly interpretable but limited in their ability to capture nuanced and context-sensitive sentiments. These models work well for simple positive/negative sentiment detection but fall short when analyzing complex emotions like “Surprise” or “Fear,” often misinterpreting context due to rigid reliance on pre-defined word lists.

- **Static Embeddings:** Older machine-learning methods using Word2Vec or GloVe embeddings fail to account for polysemy (words with multiple meanings based on context). For instance, a tweet with "I’m over the moon" may convey excitement (Joy), but traditional embeddings would not distinguish this contextual sentiment effectively.

Modern Transformer-Based Approaches:

- **Baseline Transformers:** DistilRoBERTa and DistilBERT are lightweight variants of BERT, optimized for speed and memory usage. These models excel in capturing contextual meaning from text and significantly outperform traditional methods in both accuracy and F1 score (baseline ~93.9% for distilRoBERTa, ~93.8% for distilBERT). DistilRoBERTa’s improved pre training objectives and architecture refinement give it an edge over DistilBERT, particularly in handling subtle emotional nuances.

- **Hybrid Feature Models:** By combining transformer embeddings with TF-IDF features, the project innovatively bridges deep-learning advancements with traditional statistical representations. This hybridization enriches feature spaces with additional layers of textual information, resulting in improved accuracy and F1 scores (e.g., Backbone2 + SVC achieves ~95% accuracy, outperforming standalone transformer models).

Model-Specific Observations:

- **Support Vector Classifier (SVC):** SVC, with its ability to model non-linear decision boundaries through kernel functions, emerges as

the most robust classifier. It effectively handles the complexity of hybrid embeddings, offering superior accuracy and interpretability.

● **Comparative Strengths:**

While Random Forests and HistGradientBoostingClassifiers deliver respectable performance, their results (accuracy ~93%) lag behind the hybrid SVC.

FFNN, despite being computationally heavier, only marginally outperforms simpler models, highlighting the project's focus on cost-effective solutions. In summary, this project introduces a paradigm shift by augmenting high-performance transformer models with interpretable, feature-rich hybrid embeddings, surpassing existing tools in accuracy, adaptability, and resource efficiency.

Explainability:Tackling the Black-Box Criticism:

Transformer-based models, despite their power, are often criticized for their opaque decision-making processes. This project addresses this limitation by embedding explainability at multiple levels:

● Hybrid Feature Design: Incorporating TF-IDF representations alongside embeddings adds interpretability, allowing practitioners to trace which textual features contribute to sentiment classification.

For example, if a tweet includes rare or domain-specific terms like "foreclosure" or "euphoria," TF-IDF helps capture their importance, complementing the contextual understanding from transformer embeddings.

● Error Analysis: Visual tools like percentage error charts (e.g., Figure 5 in the report) and ROC_AUC curves illuminate performance gaps, such as the difficulty in classifying "Surprise" due to data sparsity. These insights not only pinpoint model weaknesses but also guide targeted data augmentation strategies.

● Classifier Comparisons: The systematic evaluation of different classifiers (e.g., SVC vs. FFNN) enables transparent decision-making. It showcases how and why simpler models, like SVC, achieve better trade-offs between accuracy and computational overhead.

● Model Architecture Transparency: By combining interpretable machine learning models (e.g., SVC) with advanced neural embeddings, this project ensures stakeholders understand how decisions are made—a vital factor in sensitive applications like healthcare or social campaigns. Explainability isn't just a checkbox; it's a cornerstone of this project's methodology, ensuring that results are both actionable and credible.

Scalability Analysis Computational Efficiency:

● Lightweight Transformers:

DistilRoBERTa and DistilBERT, being 40% smaller and 60% faster than BERT, make this project scalable for real-time applications. These models deliver state-of-the-art results without the prohibitive resource costs associated with full-scale transformers.

● Optimized Training:

The use of a learning rate of 2e-5, weight decay of 0.01, and limited epochs (2) reflects a well-calibrated approach that balances performance with training efficiency. This setup ensures scalability for larger datasets without requiring excessive computational power.

Generalizability Across Domains:

The hybrid approach—fusing context-sensitive embeddings with traditional features—adapts easily to various datasets and languages. For instance:

Business sentiment analysis could integrate domain-specific terms using TF-IDF.

Multilingual applications could fine-tune the base transformers on localized data, broadening usability.

Resource Optimization:

SVC Over Neural Networks:

SVC's ability to deliver superior results with hybrid features proves ideal for scalability. It circumvents the computational demands of deeper networks, ensuring feasibility for

environments with constrained resources (e.g., edge devices or non-GPU systems).

Extensibility:

The ensemble methodology is future-proof. By integrating additional transformers or fine-tuning on specific domains, the project can scale to larger datasets, more complex sentiment taxonomies, or multimodal inputs (e.g., combining text with images or audio for sentiment analysis).

Real-World Implications

This project has broad societal and practical relevance. Its scalability, explainability, and innovative features can power solutions for world hunger by analyzing public sentiment on social campaigns, detecting critical feedback, and prioritizing resource allocation. For example:

- Campaign Optimization: Real-time sentiment tracking for global hunger initiatives could refine messaging strategies, increasing donor engagement and funding.
- Crisis Response: Identifying emotions like fear or sadness in disaster-related tweets could guide resource deployment more effectively, ensuring timely interventions.
- Policy Advocacy: Aggregated sentiment insights could highlight pressing issues for policymakers, leading to more focused and impactful hunger alleviation programs.

Conclusion

In conclusion, our study highlights several key takeaways in emotion classification. Firstly, even in state-of-the-art techniques, incremental improvements are attainable through the integration of relevant traditional methods, as demonstrated by our fusion of SOTA word embeddings with TF-IDF features yielding a modest enhancement over sole word embeddings. Secondly, the necessity to balance performance and cost is evident, exemplified by our comparison of baseline FFNN with ML models, where SVC emerged as a superior classifier despite the potential of more complex neural networks, which incur substantial time, space, and computational overheads. Lastly, the efficacy of combining multiple powerful models for enhanced results is underscored, exemplified by our ensemble of DistilBERT and DistilRoBERTa on hybrid features, which yielded superior classification outcomes. These insights collectively underscore the nuanced decision-making involved in optimizing emotion classification models.

Division of Work

The division of work among the group members is as follows:

1. Omkar Shetty:

Research and Background Study: Conducted a detailed literature review on sentiment analysis techniques, including traditional lexicon-based methods and transformer-based models such as DistilRoBERTa and DistilBERT.

Model Fine-Tuning: Focused on fine-tuning the DistilRoBERTa model, extracting embeddings, and integrating feedforward neural networks (FFNN) for performance enhancement.

2. Priyanshu Jain:

Data Preprocessing and Feature Engineering: Managed the dataset preprocessing, including stratified sampling to balance class distributions. Developed hybrid feature representations by combining TF-IDF with embeddings from DistilRoBERTa.

Machine Learning Implementation: Implemented machine learning classifiers like SVC, Random Forest, and AdaBoost to evaluate their performance with hybrid features.

3. Jainam Rambhia:

Ensemble Learning and Model Evaluation: Worked on integrating DistilRoBERTa and DistilBERT embeddings into ensemble models for improved classification accuracy. Conducted evaluation using metrics such as accuracy, F1 score, and ROC-AUC curves.

Result Analysis and Visualization: Analyzed model performance across different emotions and created visualizations like error percentages and classification metric comparisons.

4. Collaborative Efforts:

All members contributed to drafting the report, discussing methodologies, and interpreting results to ensure a cohesive presentation of findings.

The team collectively decided on the assumptions and strategies for optimizing emotion classification models.

Link to Executable Code:

https://drive.google.com/file/d/1fM_A4tKwhMTgmAOy-6hQ2AKZRir-jtHH/view?usp=drive_link

https://drive.google.com/file/d/1z7bPp7V7NNJ4ImDVE_4nfjGaEqk5BjWU/view?usp=drive_link

This code implements an emotion classification system using deep learning and machine learning techniques. It starts by importing necessary libraries and defining helper functions for data preprocessing and model training. Two neural network models, EmotionClassifier and EmotionClassifierImproved, are created using pre-trained transformers. The code then loads and preprocesses an emotion dataset, trains the models, and evaluates their performance. Additionally, it extracts features from the trained models and uses them to train various machine learning classifiers like SVM, Random Forest, and AdaBoost. The system also incorporates TF-IDF features and explores feature fusion to potentially improve classification accuracy. Throughout the process, the code prints performance metrics for both the deep learning models and the machine learning classifiers.

References.

1. Saravia, Elvis, et al. "CARER: Contextualized affect representations for emotion recognition." Proceedings of the 2018 conference on empirical methods in natural language processing. 2018.

2. Batra, H., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2021). BERT-Based Sentiment Analysis: A Software Engineering Perspective. ArXiv.

3. Joshy, A., & Sundar, S. (2022, December). Analyzing the performance of sentiment analysis using Bert, Distilbert, and Roberta. In 2022 IEEE International Power and Renewable Energy Conference (IPRECON) (pp. 1-6). IEEE.

4. Thiengburanathum, Pree, and Phasit Charoenkwan. "SETAR: Stacking Ensemble Learning for Thai Sentiment Analysis using RoBERTa and Hybrid Feature Representation." IEEE Access (2023).

5. Talaat, Amira Samy. "Sentiment analysis classification system using hybrid BERT models." Journal of Big Data 10.1 (2023): 110.

6. Kokab, Sayyida Tabinda, Sohail Asghar, and Shehneela Naz. "Transformer-based deep learning models for the sentiment analysis of social media data." Array 14 (2022): 100157.

Model Card

Model Details:

- **Incorporates hybrid features from DistilRoBERTa, DistilBERT models and TF-IDF features, followed by Support Vector Classifier for multi-label prediction using their respective final layers.**

Intended Use:

- **To be used for multi-emotions classification tasks, leveraging hybrid features from DistilRoBERTa and DistilBERT models. Suitable for applications requiring sentiment analysis, emotion recognition, or any other task involving multi-label classification from textual data.**

Factors:

- **Model performance across different emotional expressions aims to understand how the model generalizes to various populations.**

Metrics:

- **Accuracy and macro F1 score as performance measures, selected for their relevance to multi-label classification tasks.**

Ethical Considerations:

- **Privacy Protection measures have been taken during model development to mitigate risks of biased outcomes.**

Training Data:

- **CARER [1], train split (80%)**

Evaluation Data:

- **CARER [1], test split (20%)**

Caveats and Recommendations:

- **Further testing is warranted to assess model robustness across diverse demographic groups and linguistic contexts.**

Quantitative Analysis:

- **Accuracy achieved is 0.9501 and Macro F1 Score is 0.9174**