Politics of emotions or propaganda?

Matteo Bisotti
Department of Computer Science, University of Milan

July 17, 2025

1 Introduction

This report aims to present the work carried out for the final project of the Natural Language Processing course, taught by Professor Alfio Ferrara at the University of Milan.

Several studies in the field of communication have shown that emotions not only enhance the persuasive effectiveness of a message but also help shape public opinion, particularly in political contexts.

The main objective of this work is to analyze the strategic use of emotional language within political texts, such as speeches, social media posts, or parliamentary debates.

In recent years, the analysis of emotions in text has become increasingly significant in the field of Natural Language Processing, largely due to the advancements brought by Transformer-based architectures. Models such as Bert [4] and its variant RoBERTa [5] have proven particularly effective for tasks involving emotion classification.

The integration of these models with annotated datasets has made it possible to address the emotion classification task more accurately and in a way that is applicable to more complex contexts, such as political communication, where emotions can play a crucial role in shaping public consensus.

However, classification alone is not sufficient to understand the rhetorical meaning of emotions. For this reason, explainability techniques play a key role, aiming not only to identify which emotion is detected, but also to explain why the model produced a certain prediction. These techniques are divided into local methods, where the model explains why it made a decision on a particular input, and global methods, where the overall behavior of the model across the entire dataset is described. A particularly effective approach is SHAP (SHapley Additive exPlanations) [6], which assigns each feature a contribution value to the model prediction. Based on game theory, each feature is considered as a player contributing to the final outcome.

The code developed for this project is available in a public repository [1].

2 Research question and methodology

The aim of this project is to analyze the use of emotions in political texts, with particular attention to their rhetorical role and persuasive function. To this end, a pre-trained Transformer model was used, which is capable of identifying and classifying the emotions expressed within speeches, statements, and other political textual content.

The work was divided into two main phases.

The first phase involved the fine-tuning of a pre-trained model using the **GoEmotions** dataset [3], which includes annotations for 28 emotional categories. This allowed the model to be adapted to the task of multi-label emotion classification.

The second phase applied the fine-tuned model to a second dataset, **eu_debates** [2], containing real political speeches. This enabled several types of analysis, including:

- observing the distribution of emotions across different speeches;
- comparing the use of emotional language among different political parties and historical periods;
- detecting relevant rhetorical patterns through explainability techniques.

From a formal perspective, the project addresses a multi-label classification problem in a supervised learning setting, applied to political texts. To tackle this task, a pre-trained RoBERTa model, an optimized variant of BERT, was used along with its associated tokenizer to ensure consistent tokenization and input text representation compatible with the model architecture.

3 Experimental results

Two distinct datasets were used for this project.

3.1 Dataset GoEmotions

GoEmotions, developed by Google Research, is an annotated corpus consisting of approximately 58,000 Reddit comments with 27 emotional labels and a neutral class. It was used to fine-tune the model and evaluate classification performance. It includes the following features:

- `text`: the reddit comment;
- `id`: the unique id of the comment;
- `author`: the reddit username of the comment's author;
- `subreddit`: the subreddit that the comment belongs to;
- `link_id`: the link id;

- `parent_id`: the parent id;
- `created_utc`: the timestamp;
- `rater_id`: the unique id of the annotator;
- `example_very_unclear`: whether the annotator marked the example as being very unclear or difficult to label;
- `labels`: the associated label.

The emotional labels can be divided into four distinct groups: positive emotions (admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief), negative emotions (anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, sadness), ambiguous emotions (confusion, curiosity, realization, surprise) and neutral one.

3.2 Dataset eu_debates

This dataset contains a collection of transcripts from political debates in the European Parliament, available via Hugging Face. It was used during the inference phase to analyze the behavior of the model in real political contexts. It includes the following features:

- `text`: the original text;
- `translated_text`: the translated text;
- `speaker_party`: the speaker's party;
- `speaker_role`: the speaker's role;
- `speaker_name`: the speaker's name;
- `debate_title`: the title of the debate;
- `date`: the date of the speech;
- `year`: the year of the speech.

3.3 Model's evalutation

GoEmotions dataset was divided into: *train*, *validation* and *test*.

Training set was used to fine-tune the RoBERTa model. Training was carried out for 4 epochs, using a batch size of 16 and the `BCEWithLogitsLoss` loss function, which is suitable for a multi-label classification problem. At the end of each epoch, the model was validated in the validation set with a batch size of 64, computing both accuracy and F1 score metrics. The model with the best validation F1 score was saved. The test set was used at the end of training to evaluate the model's performance on unseen data. During this project, throughout all its phases, an NVIDIA RTX 4060 GPU was used.

To evaluate the performance of the fine-tuned model on GoEmotion, accuracy, precisiom, recall and F1-score were calculated and reported as a classification report, shown in Table 1.

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
admiration	0.67	0.76	0.71	547
amusement	0.75	0.86	0.80	270
anger	0.56	0.45	0.50	192
annoyance	0.37	0.21	0.27	304
approval	0.45	0.29	0.35	368
caring	0.35	0.55	0.43	143
confusion	0.52	0.36	0.43	136
curiosity	0.46	0.51	0.48	252
desire	0.55	0.56	0.55	84
disappointment	0.33	0.24	0.28	152
disapproval	0.44	0.29	0.35	260
disgust	0.39	0.46	0.42	106
embarrassment	0.58	0.58	0.58	33
excitement	0.59	0.39	0.47	105
fear	0.50	0.65	0.57	69
gratitude	0.94	0.89	0.91	332
grief	0.00	0.00	0.00	18
joy	0.54	0.59	0.56	179
love	0.76	0.87	0.81	268
nervousness	0.56	0.26	0.36	19
optimism	0.54	0.51	0.53	207
pride	1.00	0.11	0.19	19
realization	0.41	0.14	0.21	123
relief	0.00	0.00	0.00	22
remorse	0.48	0.60	0.53	60
sadness	0.49	0.49	0.49	144
surprise	0.48	0.58	0.52	132
neutral	0.70	0.58	0.63	1818

Table 1: Classification report per emotion

The classification report shows good performances for the most frequent emotions, such as *gratitude*, *love* and *amusement*. On the other hand, emotions with low support like *grief*, *relief* and *pride*, perform badly. More ambiguous or nuanced emotions, such *realization* and *approval*, despite having sufficient support, show intermediate performance, likely due to the variability in how they are expressed in text, making them linguistically more complex.

3.4 Embeddings

To examine how fine-tuning altered the model's semantic representation, the embedding space was visualized both before and after training on GoEmotions. In

particular, embeddings associated with selected emotion-related keywords were extracted in order to show their position within the vector space.

Since the embeddings generated by RoBERTa have a dimensionality of 768, a dimensionality reduction technique was applied to make them suitable for visualization. Specifically, Principal Component Analysis (PCA) was used with the number of principal components set to 3, in order to obtain an interpretable visual representation.

In the 3d plot below 1, the embeddings of emotional words are shown before (blue points) and after (red points) fine-tuning the RoBERTa model. The arrows indicates the displacement vector for each embedding. Before fine-tuning, the embeddings were highly concentrated and did not exhibit a clear sematic separation between different emotions. After fine-tuning, the words became more widely distributed in space: words associated with similar emotions (for example *happy* and *joy* or *angry* and *hate* and *fear* or *terrified*) tend to be positioned closer together.

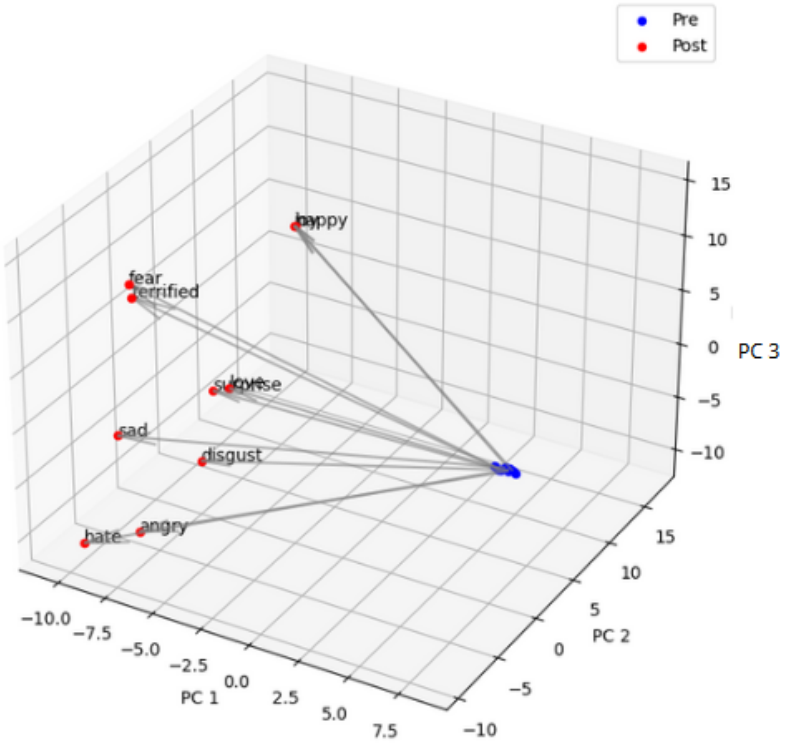

Figure 1: Embeddings space Pre vs Post fine-tuning

The heatmap in Figure 2 shows the cosine similarity between each pair of emotion-related sentences. Higher values indicate greater semantic proximity in the new vector space learned by the model. This representation allows for a direct observation of how the model differentiates or groups together similar, opposite, or ambiguous emotions.

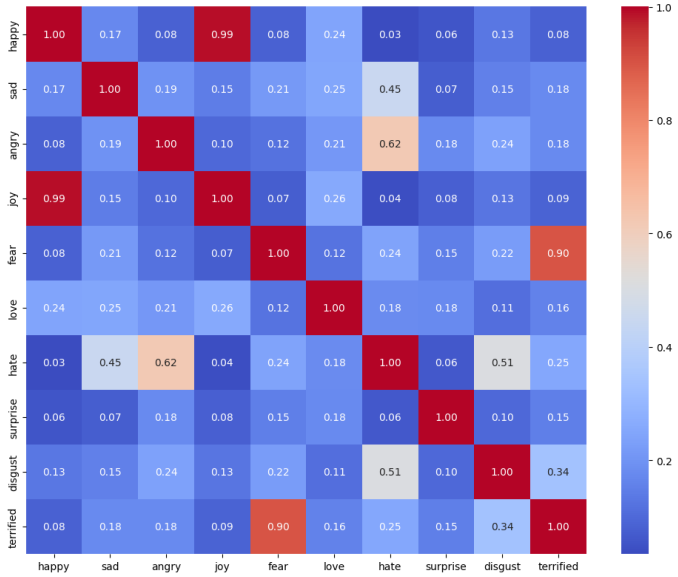

Figure 2: Heatmap of cosine similarity (Post fine-tuning)

3.5 Inference on political texts

One of the main objective of the project was to study the strategic use of emotions in political language, analyzing how they are distributed across parties and temporal contexts. To do this, the fine-tuned model was applied to the *eu_debates*.

From a practical perspective, the inference process was performed in batches using `PyTorch`, with a batch size of 64. For each batch, the model produced logits, which were converted into probabilities using a sigmoid function, as this is a multi-label classification problem. The predicted emotions were then mapped to their corresponding names using the model `id2label` dictionary.

Once inference was completed, the emotions associated with each speech were aggregated by year, shown in Table 2, and by political party, shown in Table 3, and visualized using heatmaps.

The heatmap in Figure 3 shows the distribution of emotions across years. There are differences in the overall number of texts, with a peak between 2010 and 2012, which should be taken into account. We can observe that most speeches are associated with positive emotions—often recurring in political language—such as *approval*, *gratitude*, and *optimism*, or with neutral ones, while negative emotions appear less frequently. This may be consistent with the institutional context of the texts, as speeches from the European Parliament tend to maintain a formal and controlled tone compared to those delivered at rallies or outside institutional settings. So the analysis highlights, despite quantitative differences between years, a consistent trend over time toward neutral or positive language.

Year	Value
2009	1876
2010	7285
2011	10602
2012	6901
2013	458
2014	400
2015	746
2016	644
2017	508
2018	458
2019	1364
2020	1414
2021	1836
2022	2474
2023	1804

Table 2: Distribution per year

Party	Value
ALDE (*Centre*)	3993
ECR (*Right-wing*)	2127
GUE/NGL (*Left-wing*)	3068
Greens/EFA (*Centre-left*)	2987
ID (*Right-wing*)	2222
NI (*Independents*)	2314
PPE (*Centre-right*)	13733
S&D (*Centre-left*)	8326

Table 3: Distribution per party

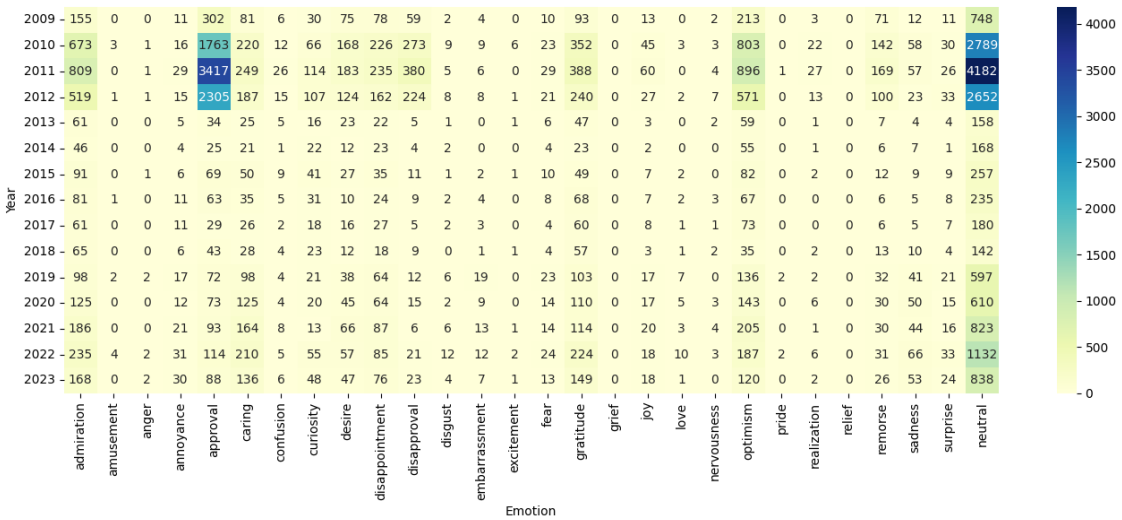

Figure 3: Heatmap on years

The heatmap in Figure 4 shows the distribution of emotions across parties. Here too, there are differences in the overall number of texts among the parties. However, a predominance of positive or neutral emotions emerges. No party shows a high use of negatively charged speeches, confirming that the language used in the European Parliament tends to maintain an institutional and measured tone, regardless of party ideology.

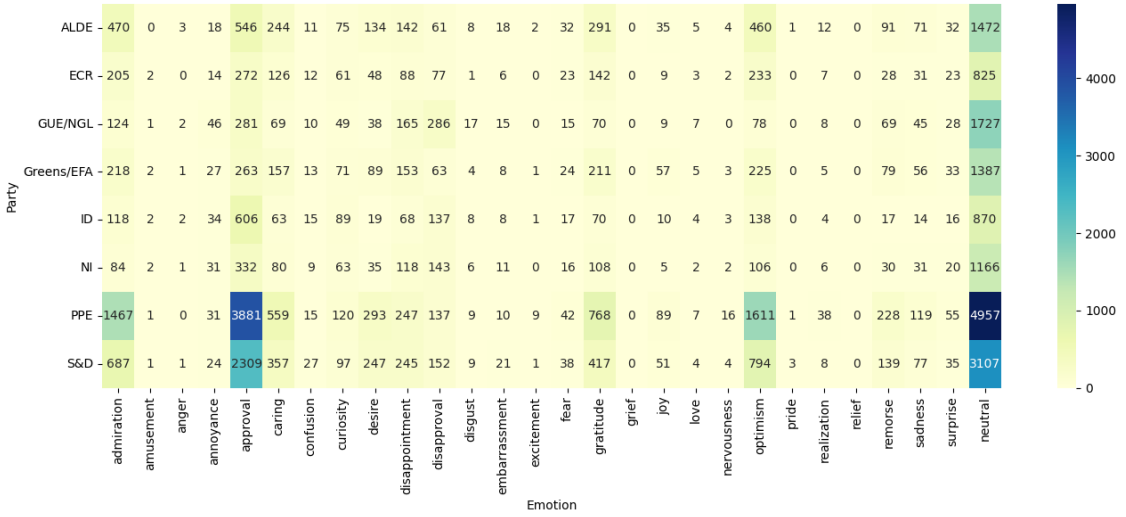

Figure 4: Heatmap on parties

3.6 Explainability methods

To interpret the model's behavior and understand which words most influence emotion classification, a global explainability technique based on SHAP was applied. Average SHAP values were calculated for each word with respect to the emotion classes, and a heatmap was used to visualize the 10 words with the highest predictive impact for each emotion, both on the GoEmotions dataset and on the dataset of political speeches.

As shown in Figure 5, the terms associated with the emotions are largely consistent with the labels: for example, words such as *thank*, *thanks*, and *welcome* appear in the row for *gratitude*, while terms like *scared*, *haunting*, and *nightmare* are associated with the emotion *fear*. Furthermore, vulgar or aggressive terms such as *fuck* or *pissed* are linked to the *anger* label.

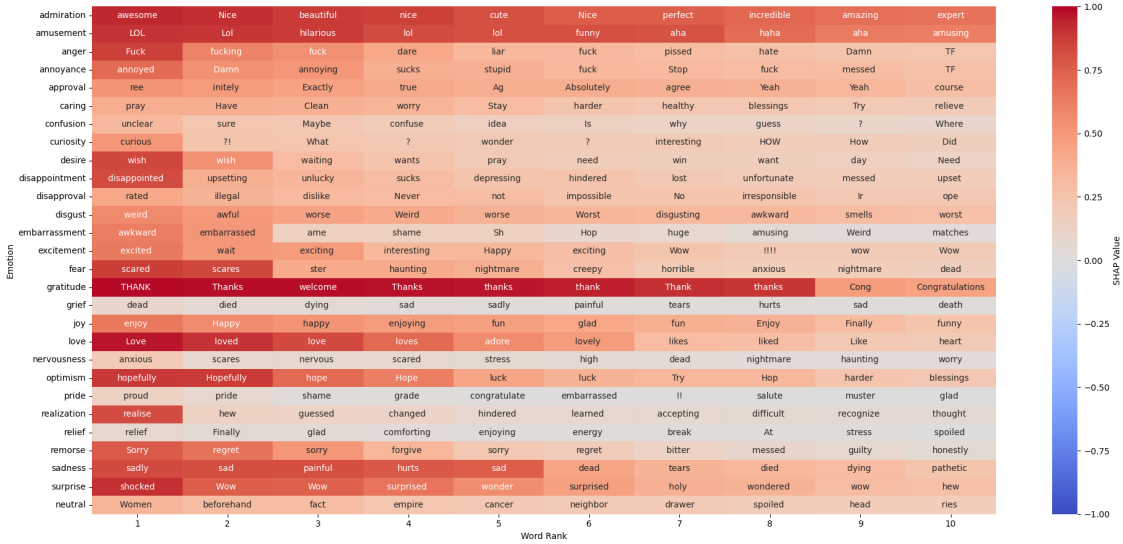

Figure 5: Top 10 words on GoEmotions

Unlike the GoEmotions dataset, which contains short phrases and comments, the language of political speeches is more elaborate, and this is reflected in the observed SHAP values: as suggested by the lighter coloring of the heatmap, words tend to have a more diluted predictive impact, likely because the emotional content is spread over a larger context, making the individual contributions of specific words less pronounced.

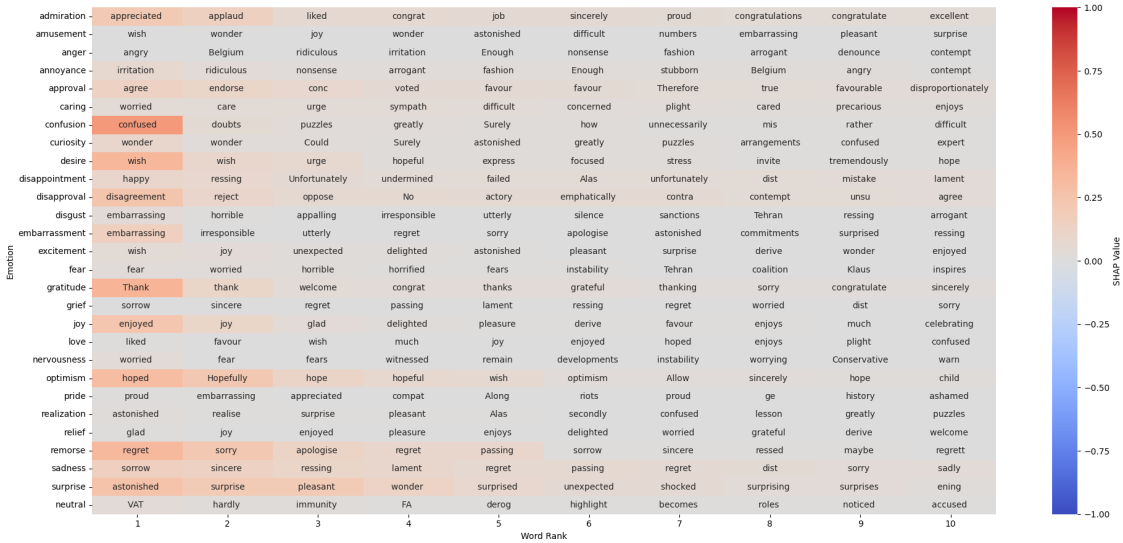

Figure 6: Top 10 words on eu_debates

4 Concluding remarks

This work shows how the combined use of Transformer models, in this case the pretrained RoBERTa model, and explainability techniques can provide effective tools for emotion analysis and classification in political texts.

The results highlight several trends. First of all, the political speeches analyzed are characterized by a predominant use of positive or neutral emotions, while strongly negative ones are less frequent. This is consistent with the formal and institutional nature of the context in which they originate.

However, the use of explainability techniques has also revealed some limitations. In long and complex political speeches, the predictive contribution of individual words tends to be diluted, as shown by the generally lower SHAP values compared to those observed in the GoEmotions dataset.

In this project, the generative model GPT-4 was used for generating code related to plotting graphs, debugging portions of code involved in model training, improving the report's form and syntax in English. The ideas behind the experiments conducted for embedding visualization and explainability technique, however, were entirely conceived by the author of this report.

References

[1] Matteo Bisotti. Repository github. `https://github.com/MatteoBisotti/nlp_project`, 2025.

[2] Ilias Chalkidis and Stephanie Brandl. Llama meets eu: Investigating the european political spectrum through the lens of llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

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

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[6] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4765–4774, 2017.