SFWRENG 4NL3 Assignment 4: Pretrained Transformers

Sumanya Gulati

1 April 2025

Contents

List of Tables

List of Figures

1 Dataset

For this assignment, I am using the GoEmotions dataset extracted from the HuggingFace can be accessed here. The task at hand involves emotion classification, aiming to identify and categorize the emotions expressed in textual data. This is essential for applications such as sentiment analysis, mental health assessment, enhancing human-computer interactions and more.

The GoEmotions dataset comprises of about 58,000 English Reddit comments, each annotated for 27 distinct emotion categories or marked as neutral. The simplified version of the dataset with predefined train, val and test splits has been used for this assignment.

1.1 Data Collection

The dataset has been constructed by selecting English-language comments from Reddit by researchers at Amazon Alexa, Google Research and Stanford Linguistics. A complete list of authors can be found here. The comments were extracted from Reddit using a variety of automated methods along with data curation techniques such as reducing profanity, length filtering, sentiment and emotion balancing, masking and more. Further information about the data collection process can be found in section 3.1 of this paper.

1.2 Dataset Structure

Each instance of the dataset corresponds to a reddit comment with an ID and one or more emotion annotations including neutral. The simplified configuration of the dataset which has been used for this assignment, includes:

- `text`: the Reddit comment
- `labels`: the emotional annotations
- `comment_id`: a unique identifier for the comment

The input for the task is a Reddit comment in English and the corresponding output is a set of one or more labels corresponding to the 27 emotion categories or neutral, reflecting the emotional content of the comment.

The labels are stored as a list of integers ranging from 0 to 27 where each integer represents an emotion category or neutral. The label mapping is as follows:

Label Number	Label Category
0	admiration
1	amusement
2	anger
3	annoyance
4	approval
5	caring
6	confusion
7	curiosity
8	desire
9	disappointment
10	disapproval
11	disgust
12	embarrassment
13	excitement
14	fear
15	gratitude
16	grief
17	joy
18	love
19	nervousness
20	optimism
21	pride
22	realization
23	relief
24	remorse
25	sadness
26	surprise
27	neutral

Table 1: Label Mapping

1.3 Evaluation Metrics

The following evaluation metrics have been used to assess the performance of the BERT-based model:

- Model Performance:
  - Emotion-level Precision, Recall, F1: Measured per each emotion in the GoEmotions taxonomy.
  - Transfer Learning: F1 score on data transferred between domain X and GoEmotions.
- Deciison thresholds: No thresholds are used. The data is presented in full granularity.
- Uncertainty and variability: Repeated experiments have yielded results with similar taxonomical rankings.

The model has been evaluated on 10 publicly available datasets including 9 benchmark datasets presented in compilation by Bostan and Klinger and the GoEmotions eval set. Full details about the evaludation results can be found in the paper.

1.4 Data Splits

The dataset is divided into training, validating and test splits as follows:

Data Split	Number of Instances
Training	43,410
Validation	5,426
Test	5,427

Table 2: Data Split Overview

Class distributions for the training and test splits are shown in the figure 1.4.

	Emotion	Train Count	Test Count
0	admiration	4130	504
1	amusement	2328	264
2	anger	1567	198
3	annoyance	2470	320
4	approval	2939	351
5	caring	1087	135
6	confusion	1368	153
7	curiosity	2191	284
8	desire	641	83
9	disappointment	1269	151
10	disapproval	2022	267
11	disgust	793	123
12	embarrassment	303	37
13	excitement	853	103
14	fear	596	78
15	gratitude	2662	352
16	grief	77	6
17	joy	1452	161
18	love	2086	238
19	nervousness	164	23
20	optimism	1581	186
21	pride	111	16
22	realization	1110	145
23	relief	153	11
24	remorse	545	56
25	sadness	1326	156
26	surprise	1060	141
27	neutral	14219	1787

Figure 1: Class Distribution in Training and Test Splits

1.5 Additional Features

Although the comments vary in length, the maximum sequence length has been capped at 30 tokens in the training and evaludation datasets to ensure concise and focused emotional expressions. Furthermore, linguistic context consistency has been ensured by only choosing comments that are in English.

2 Methodology

2.1 Preprocessing

Since the `data_extract.py` file only extracts the data from the PDF and store it in a text file, the first step in preprocessing the dataset was to clean the text. This was accomplished using the following methods through

iterative testing:

- Replacing all new lines with a blank space.
- Removing all words that have a hyphen followed by a blank space. For example, long- term to longterm.
- Removing all leading and trailing spaces along with the replacement of multiple blank spaces with a single space.
- Removing all digits to streamline the frequency count and ranking of alphabetic tokens. This was implemented because with the digits in the corpus, the years and other random numbers referencing legal cases and such populated the top 10/top 25 counts corrupting the data analysis.
- Removing all punctuation for more accurate tokenization.

2.1.1 Naive-Bayes

Further preprocessing for Naive-Bayes included the following steps:

- Using the list of common stopwords from the *nltk* library to remove words that do not reflect the actual content.
- The list of stopwords was expanded to account for commonly found words in the dataset that did not convey the sentiment or the content of the sentence. This includes words such as `said, would, one, jan, feb`.
- To further refine our results, commonly used legal jargon was also added to the list of stopwords. This includes words such as `hearings, committee, senator, judge, case`.

2.1.2 Latent Dirichlet Allocation (LDA) Preprocessing

Further preprocessing for LDA included the following steps:

- Adding additional stop words including 'sgpohearings' and 'sgpohearingstxt' that are not actual words with defined meanings.
- Using NLTK's lemmatizer removing any tokens starting with 'sgpo'.

2.2 Bag-of-Words Representation

As shown in tutorials, the `CountVectorizer()` function was used to implement Bag-of-Words on the dataset. Probability calculations given in the instructions PDF for this assignment were then used to implement the `calculate_word_probabilities()` function. The output from this function was then to calculate the Log Likelihood Ratio.

2.3 Naive-Bayes Model

Given the fact that the dataset is made up of real transcripts of legal hearings, the used of names of the judges and/or the senators interviewing them are over-represented in the dataset. This skews the results of analyses performed on the processed dataset.

To combat this issue and to maintain the focus on the goal of this assignment, classifications were establish so the focus of the analyses would remain on words that align with the characteristics of the nominee. 4 different classifications - Legal Interpretation, Background Experience, Personal Characteristics and Competency Qualification were created.

Each of these classifications was populated with a list of related words that speak to that particular classification, such as - words like 'ruling', 'precedent', 'principle', 'rights' and more were added to the Legal Interpretation classification. By analyzing the frequency of these words in the vocabulary for the female and male nominees, any potential bias towards a nominee can be spotted. If the focus of the hearing is on 1-2 classifications as opposed to a general overview of all of them, this would highlight implicit bias.

2.4 Topic Modelling

The `gensim` and `pyLDSvis` libraries were used to implement LDA topic modelling as shown in the tutorial. The `get_topic_words()` and `get_document_topic()` functions were then used to get information on the topics and documents.

2.5 Experimentation

In addition to importing all the functions from the `processing.py` file, two new functions were added to implement text normalization variation and LDA variation. `PorterStemmer()` was used to implement stemming as a part of the preprocessing of the data. Another function was then implemented to create a TF-IDF matrix using the `TfidfVectorizer()` function. The code for this was based on the tutorials.

3 Results and Analysis

The results and insights gather from the analyses has been summarized in the following subsections.

3.1 Naive-Bayes

Gender	Category	Word	LLR Score
female	background experience	backgrounds	1.931906639972084
female	background experience	experienced	1.0156159080979297
female	background experience	experiences	1.9682742841429581
female	background experience	framework	2.52789007207838
female	background experience	services	1.4982706548972224
female	personal characteristics	biases	1.6645918703447364
female	personal characteristics	ethnicity	2.6250538205320293
female	personal characteristics	gender	1.0750667224516768
female	personal characteristics	genderbased	2.10180567676748
female	personal characteristics	grace	1.2751271035830118
female	personal characteristics	identity	1.1209764237557547
female	competency qualification	applicability	1.8941663119892365
female	competency qualification	availability	1.46928311802397
male	legal interpretation	judicially	1.1407866747180364
male	legal interpretation	overruling	1.2014112965344719
male	background experience	practiced	1.4245548478486825
male	background experience	worker	1.0762481535804653
male	background experience	workload	1.4817132616886308
male	personal characteristics	beliefs	1.3067718121922978
male	personal characteristics	embraced	1.7693953341404107
male	personal characteristics	unbiased	2.0570774065921924
male	competency qualification	disability	1.209779546204988
male	competency qualification	wellqualified	2.0570774065921924

Table 3: Results from the Naive-Bayes Analysis

As shown in table 3.1, the analysis shows that when interviewing female nominees are interviewed, greater focus is put on their personal characteristics with conversations about biases, gender, grace and identity taking up a significant amount of attention. In terms of their background experiences and competency qualifictaions, the focus is put on their availability and the top words from the category across documents are neutral in tone. None of the words added to the legal interpretation category were found in a significant frequency for the female nominees.

As for the male nominees, however, words like 'judicially' and 'overruling' are highlighted showing the focus on the conversation around the nominee's career in law and their judgments. The conversations around personal characteristics seem to be more focused on the individual as opposed to their gender or other aspects of their identity. The most starking difference can be seen in the tone of the competency qualifications since the word 'wellqualified' seems to have a very high probability showing the appreciative tone of the hearings.

3.2 Topic Modelling

The complete list of topic words along with their labels can be found in the `topic_words.csv` file. The averages for the top 20 of these topics split by the gender can also be found in the `category_topics.csv` file.

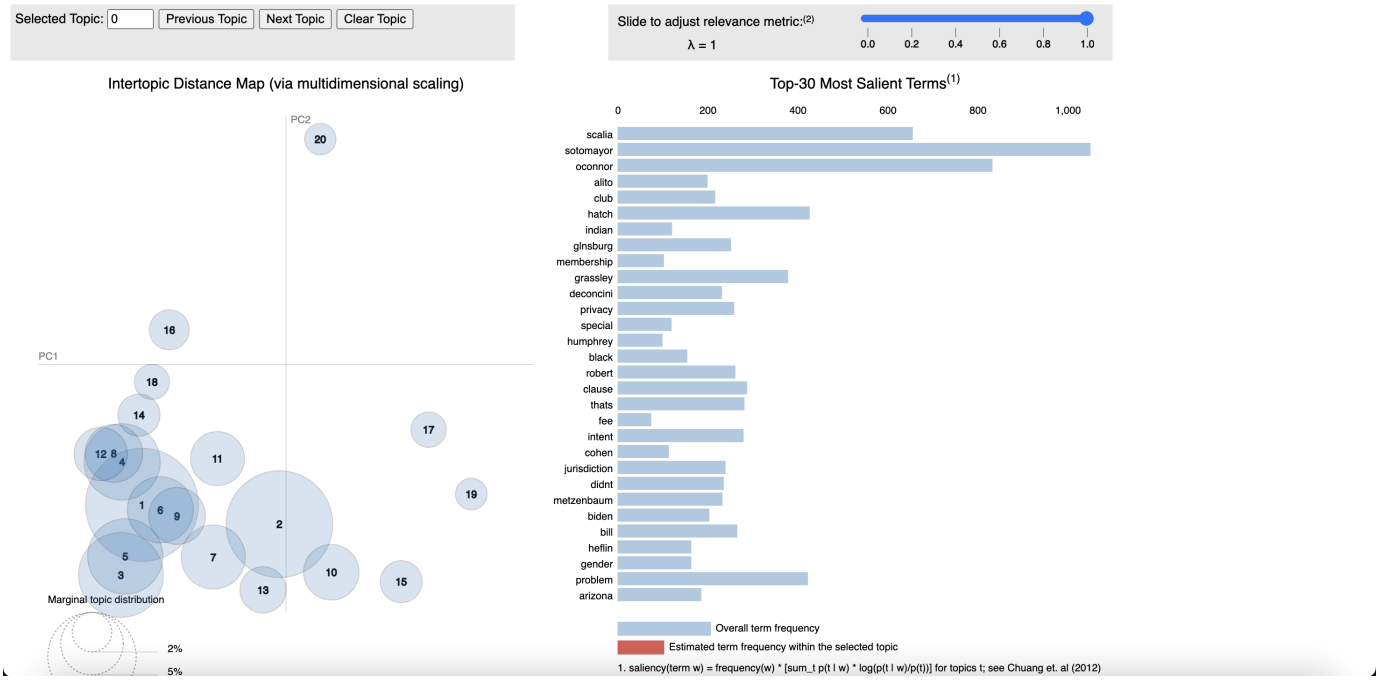

Figure 2: Results from LDA Topic Modelling

As shown in figure 3.2, the intertopic distance map visualizes the relationship between the different topics where each bubble represents a topic with the bubble's size indicating the topic's overall prevalence in the dataset. Topics 12, 8, 4, 1, 6, 9, 5, 3 and more are overlapping and clustered together suggesting similarity.

The 30 most important words are either names, identity-related terms (gender, black) or general legal jargon. A few interesting insights have been listed here:

- Topic 12 that mentions Justice Ginsburg and legislation is much more popular along female nominees given the work she has done championing women's rights and gender equality.
- The use of words such as 'business' and 'agency' are significantly more popular with male nominees showing the focus on the nominee's competencies and judicial philosophies. This also includes top 15 where talks of legal jurisdictions with words such as 'privacy', 'clause' and 'national' are way more popular with male nominees. This is supported by the arrangement of the thematic clusters for topics 5, 15, 8 and 11.
- The thematic clusters show topics 3, 12, 2 and 13 that focus on Justice Ginsburg, Justice O'Connor,gender and voting legislation are more heavily focused on privacy and reproductive rights, often regarded as 'women''s issues'.

3.3 Experimentation

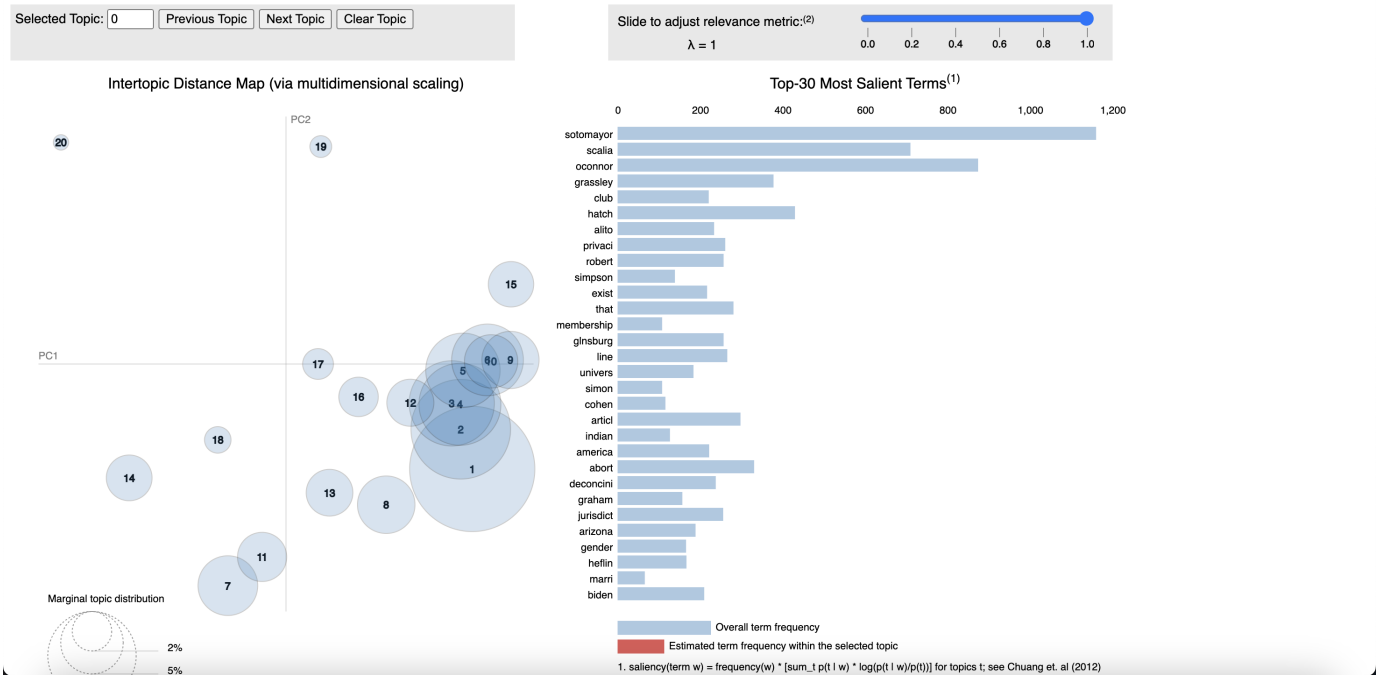

Figure 3: Results from LDA Topic Modelling Post-Stemming and with TF-DIF

As shown in figure 3.3, it is obvious that stemming has been used as terms like 'privacy' have been incorrectly spelt as 'privaci'. The analysis result is pretty similar to the previous section's result excluding the misspelt words.

4 Discussion

This section outlines common findings and insights dervided from the program.

4.1 Findings

As outlined in the description of the dataset, the hypothesis remains proven true. The Naive-Bayes analysis and the LDA analysis support the theory that when interviewing women nominees, more focus is placed on their personal characteristics and questioning their qualifications as opposed to the male nominees who are questioned over their judicial philosophy and appreciated for their competency and qualifications. For instance, topics involving privacy rights, gender, and specific legal clauses appeared more frequently in female discourse, while topics involving procedural aspects or judiciary figures had a higher presence in male discourse.

4.2 Reflection

Although I was aware of the importance of text normalization and preprocessing, their importance became a lot more evident during the course of this assignment. Even small changes in preprocessing steps altered the topic distributions, which highlighted the sensitivity of topic modeling to text cleaning procedure. It was also surprising (although should have been expected) that the transcripts are dominated by names and judicial jargon.

It has also become clear to me that topic modelling has limitations such as the fact that it provides a useful high-level overview of themes, it lacks the ability to capture nuanced contextual meanings. Some topics were too broad or contained words that seemed unrelated at first glance. Overall, this assignment deepened my understanding of text analysis and the complexities involved in extracting meaningful insights from unstructured textual data. It also reinforced the importance of critical interpretation when working with computational models - numbers alone do not tell the full story and human judgment is essential in deriving meaningful conclusions.