Natural Language – Project Paper

Group 20: Afonso Ribeiro (ist1102763), Pedro Curto (ist1103091), Pedro Ribeiro (ist1102663)

Models

For this project, we have selected a handful of models that we believed to have nice concepts while being relatively simple to train and test, given our hardware limitations. Along the way, we explored some approaches that ended up being omitted, since they didn't produce results relevant to our goal.

Pre-Processing

For the next few models, except BERT, we manually selected and removed stop words from the dataset. We found that alternative methods, such as lemmatization, undersampling and oversampling, negatively impacted the performance of the models. This probably happened because they reduced the vocabulary size and made it harder to see unique terms. For BERT, we simply removed extra spaces, as in Transformer-based models we need the full context for accurate embeddings.

Multinomial Naïve-Bayes

To evaluate the importance of each word in the plots, we employed TF-IDF and generated a vector representation for each document.

Using these vectors, we trained a multinomial Naïve Bayes model with the parameter 'alpha' set to 0.1. A higher value of alpha increases the smoothing effect, but we observed that increasing and inflating probabilities for unseen words, ultimately led to a poorer model performance.

Support Vector Classification

We tried two different approaches for generating word embeddings for the model:

- Using GLOVE: we got a pretrained GLOVE model and used it to get an embedding for each word. The embeddings for the plots were then calculated by averaging each of their words' embeddings.
- Document based TF-IDF: the same approach as the Naïve Bayes model. We used TF-IDF to score the words at document level, capturing terms importance within each document.

We then fed these embeddings to a basic Support Vector Classifier model with a linear kernel and settled on a regularization parameter of value 1, to see if a more geometrical approach could improve the previous results. We also tested the model with different kernels, but the results were worse.

DistilBERT

Lastly, we fine-tuned a pre-trained DistilBERT model, a lighter version of BERT, along three epochs. Given that DistilBERT is optimized for efficiency and works well with smaller datasets like ours, we chose it over the full model. However, it is important to note that it has a maximum token length of 512, which posed a challenge for processing our plot data, since a lot of plots exceeded this limit. To address this issue, we had to truncate the plots to fit within the 512-token range, shortening plots that exceeded the limit and adding padding to the smaller ones. We tested an idea in which we divided the plots that exceeded the token limit in chunks, labelling them with the same category as the original plot. However, we decided not to proceed with this approach due to its inconsistent behaviour.

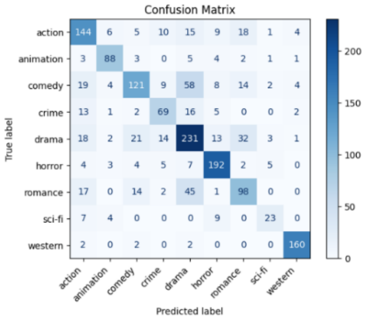
Extra models – some short experiments

To see if sentiment analysis could help improving the SVC model, we tried classifying each plot based on the sentiment polarity, scoring them from 0 to 1. We then added this result to the vector of the TF-IDF as an extra feature and ran the SVC model to see if it had any effect on the space division. It slightly improved it, but not anything extraordinary. We also tried to get the embeddings from a pre-trained ELMo model, like we did with GLOVE, but using it consumed a lot of memory. We tried fixing this by dividing the plots in batches and reducing their size gradually, but it still took too long and crashed due to the memory usage.

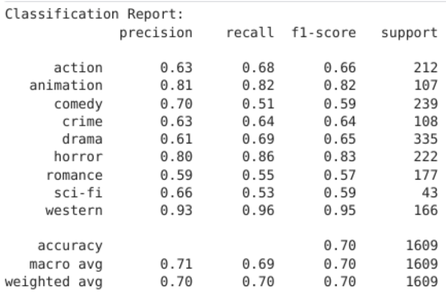
Experimental Setup and Results

Each model was trained and tested using an 80/20 train-test split. We evaluated performance based on both accuracy and F1-score. To visually interpret the results, we used confusion matrices. Furthermore, we extracted the misclassified plots for each model into separate files and reviewed these misclassified plots ourselves to better understand the challenges our models faced, identifying any patterns in their misclassifications.

Below is the confusion matrix for the BERT model, which showed the best performance, followed by its classification report. We will be using these results for comparison with the remaining models, to see if we can identify why it performed better. All the classification reports and confusion matrices for the remaining models can be analysed through the notebooks.



	action	animation	comedy	crime	drama	horror	romance	sci-fi	western
action	144	6	5	10	15	9	18	1	4
animation	3	88	3	0	5	4	2	1	1
comedy	19	4	121	9	58	8	14	2	4
crime	13	1	2	69	16	5	0	0	2
drama	18	2	21	14	231	13	32	3	1
horror	4	3	4	5	7	192	2	5	0
romance	17	0	14	2	45	1	98	0	0
sci-fi	7	4	0	0	0	9	0	23	0
western	2	0	2	0	2	0	0	0	160



```
Classification Report:
              precision    recall  f1-score   support

      action       0.63      0.68      0.66       212
   animation       0.81      0.82      0.82       107
      comedy       0.70      0.51      0.59       239
       crime       0.63      0.64      0.64       108
       drama       0.61      0.69      0.65       335
      horror       0.80      0.86      0.83       222
     romance       0.59      0.55      0.57       177
      sci-fi       0.66      0.53      0.59        43
     western       0.93      0.96      0.95       166

    accuracy                           0.70      1609
   macro avg       0.71      0.69      0.70      1609
weighted avg       0.70      0.70      0.70      1609
```

Discussion

Dataset Analysis

Before starting to experiment with models, we analysed the genre distribution, the most frequent and unique words for each genre, the global word frequencies, plot lengths and conducted sentiment analysis.

Our analysis revealed that the dataset is heavily skewed, with drama being significantly overrepresented compared to genres like sci-fi. This imbalance could limit the performance of models like Naïve Bayes, as we'll see, which rely on probability estimates that may become biased towards the dominant genre.

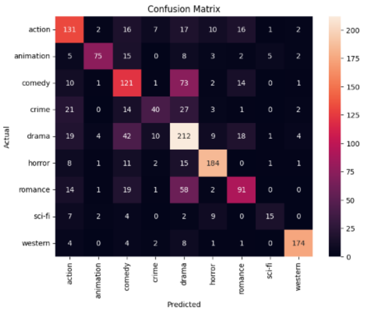
While we identified genre-specific keywords, we also observed that context plays a crucial role in distinguishing genres. The word "killed" frequently appears in both horror and action movies, but in one genre is more related to supernatural creatures and the other to human homicides and investigation. Without capturing this context, simple keyword analysis might not be enough for accurate classification. Additionally, sentiment analysis provided limited value, as the sentiment polarity across genres was relatively similar for the model we applied. This suggests that sentiment alone may not be a strong differentiator for genre classification in this dataset. We concluded similarly for plot length, which didn't add particular value as a feature.

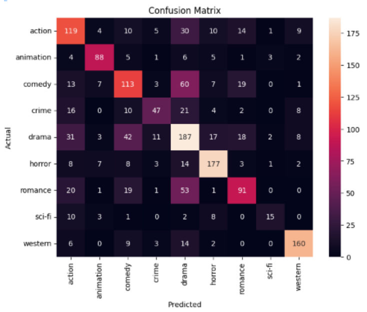
Results Analysis

Our best-performing model was the fine-tuned DistilBERT, primarily due to its ability to understand and interpret context. BERT uses self-attention to capture relationships between words, considering both the preceding and the following terms in a sentence. This makes it great at understanding context, which, as we've mentioned before, is critical for accurate classification in our dataset.

It outperformed models such as the TF-IDF with SVC, which assigns words importance based solely on their frequency in documents, without understanding their context. Consider this example of a misclassified clean text from the TF-IDF model: "shiva middle class medical student chennai naga local mafia goon fall love mahalakshmi leads lot problems." The true label for this text was action, but the model incorrectly classified it as romance, likely due to the presence of the word "love." In contrast, the BERT model correctly identified the genre as action.

Analysing the following confusion matrices for the models using TF-IDF (left image) and GLOVE (right image) along with an SVC, and comparing its predictions for the drama genre with those from the BERT model (showed in the previous section), we reached a similar conclusion. The TF-IDF + SVC model struggled to differentiate drama from other genres. This is likely due to drama having a wide range of plots, many of which share keywords with other genres, making it hard for the model to distinguish them accurately. Similarly, the GLOVE + SVC model faced the same issue. As GLOVE produces static embeddings where each word has a fixed representation, it lacks the ability to capture the varying meanings of a word. In contrast, BERT generates contextualized embeddings, where a word's meaning shifts based on the sentence or genre it appears in. The larger variety of drama plots actually helped BERT, as its ability to detect context allowed it to better capture the nuances between drama and other genres.





The Naïve Bayes model had comparable performance to the SVC model and handled class imbalances surprisingly well. However, in some cases, it tended to favour genres with more frequent occurrences in the dataset. For instance, in the sentence: "Joe Puddlefoot becomes involved with criminals trying to steal valuable jade pieces belonging to the distinguished Sir Charles Goode," the model incorrectly classified it as comedy rather than crime, likely because comedy appeared 1,193 times in the dataset compared to just 541 occurrences of crime.

In the end, we found that the dataset lacked high quality, and this led to classification challenges. Many sentences were ambiguous, making it difficult for even human classifiers to correctly identify the genre based solely on the plots. For example, the sentence: "The movie is a triangular love story of an orphan boy Kanteerava played by Vijay," suggests a romance, but the actual genre is action. This highlights that many films may belong to multiple genres, further complicating classification.

Additionally, the dataset was inconsistent – some plots were long and detailed, covering the entire story, while others were just one-liners, like the previous one. All these inconsistencies likely contributed to the difficulty models faced, including BERT, in accurately distinguishing between genres like romance, action, and drama.

Future Work

Moving forward, we would like to implement a Longformer model to address the 512-token limit of BERT, since, by shortening long plots, we are losing a lot of information. Additionally, we aim to explore a model that combines ELMo embeddings with an SVC classifier, as ELMo also provides contextualized embeddings. We are interested, as well, in testing BERT as a feature extractor, because it would allow us to draw deeper conclusions.

Helpful Kaggle Notebooks

Text Classification with DistilBERT

Beginner to Intermediate NLP Tutorial