CMPUT 461
Leen Alzebdeh (alzebdeh)
Sukhnoor Khehra (skhehra)

Naive Bayes Classifier for Relation Extraction - Report

Metrics

Accuracy table

Training accuracy	Test accuracy
78.60% with std of 0.19	89.50%

Confusion matrix

\ Actual Predicted \	Director	Characters	Performers	Publishers
Director	82	6	3	0
Characters	5	87	7	2
Performers	3	4	92	1
Publishers	4	6	1	97

Micro and macro metrics

	Micro (pooled)	Macro
Precision	89.50%	89.50%
Recall	89.51%	89.52%

Our Choices and Methodology

Tokenization and Model Parameters

Tokenization: We tokenized the sentences by splitting them according to spaces.

Handling unknown words: Words not encountered in the training set were ignored during prediction, aligning with the typical Naive Bayes approach.
Smoothing: Laplace smoothing was employed to handle zero probabilities, ensuring that unseen words in the test data do not completely nullify the probability of a label.

Data Preprocessing and Feature Selection

The data preprocessing included parsing CSV data into lists of tuples and then splitting the data read from the CSV file into 3 folds for cross-validation.

Error Analysis

Misclassification Trends

High Misclassification in 'Characters' Class: The confusion matrix indicates a notable trend of misclassification involving the 'characters' class. Specifically, a considerable number of instances that were true 'performers' were incorrectly classified as 'characters' (n=7). This suggests a potential overlap in the language or context used to describe these two categories in the dataset. In addition, a significant number of ‘characters’ were misclassified as ‘director’ (n=6)
Director and Publisher Confusion: Another observed trend was the misclassification between 'director' and 'publisher' classes. Although less pronounced than the characters-performers confusion, it indicates a likely similarity in the feature words between these two categories.

Common Attributes in Misclassified Texts

Shared Contextual Words: Upon reviewing the misclassified texts, we noticed that they often contained similar contextual words or phrases. For example, terms related to artistic or creative works were present in both 'performers' and 'characters' classes, leading to confusion.

> Example: Line 47, where a “character” label was misclassified with a “performer” label. Mentioning “Simpsons” and “Groundskeeper Willie” may have created ambiguity, as these terms may be more relevant to performers instead of characters.
>
> - *1593,"The phrase was first popularized in the "" Simpsons "" episode "" ' Round Springfield "" (season 6 , 1995).Sound recording of Groundskeeper Willie 's line "" About : Political humour "" .",characters,8,23 24*
>
> Example: Line 12, where a “director” label was misclassified with a “publisher” label. This specific misclassification does not occur very frequently. With further examination of the text, we can see that the mention of “Sony” may have interfered with classification as it is a publisher.
>
> - *2359,"Following the release of "" Spider - Man 3 "" , Sony Pictures Entertainment had announced a May 5 , 2011 , release date for Sam Raimi 's next film in the earlier series .",director,5 6 7 8,25 26*

Lack of Distinguishing Features: Many misclassified texts lacked strong distinguishing features or keywords that could be decisively linked to one specific class. This was particularly evident in sentences that were shorter or less descriptive.

Identified Reasons Behind Errors

Insufficient Contextual Understanding: The initial model lacked mechanisms to effectively capture and utilize the context in which entities (head and tail) appeared. This led to difficulties in accurately distinguishing between closely related classes without the needed extra context.
Feature Overlap and Noise: The initial feature set likely contained overlapping and noisy features that contributed to the misclassifications. The use of a simple bag-of-words model without feature selection allowed common, non-discriminative words to influence the classification.

Grad Extension

Accuracy table

Training accuracy	Test accuracy
91.18% with std of 0.06	90.00%

Confusion matrix

\ Actual Predicted \	Director	Characters	Performers	Publishers
Director	85	4	4	1
Characters	6	89	5	3
Performers	3	3	93	3
Publishers	0	7	1	93

Micro and macro metrics

	Micro (pooled)	Macro
Precision	90.00%	90.00%
Recall	90.03%	90.02%

The updated model shows an improvement in accuracy and maintains balanced performance across different classes.

Our Choices and Methodology

Tokenization and Model Parameters

Tokenization: We tokenized the sentences by splitting them according to spaces.
Model parameters: ??
Handling unknown words: Words not encountered in the training set were ignored during prediction, aligning with the typical Naive Bayes approach.
Smoothing: Laplace smoothing was employed to handle zero probabilities, ensuring that unseen words in the test data do not completely nullify the probability of a label.

Data Preprocessing and Feature Selection

Along with the same CSV data reading and 3-fold cross-validation splitting found in the first part of this report, we then added these choices:
Head and Tail Entities:

- Based on the errors of the initial model, we determined that the model required more contextual clues to further accurately classify the text data.
- After improving the model: to integrate the contextual information provided by the positions of head and tail entities, we modified the input data by appending "HEAD_" and "TAIL_" tags to the corresponding words. This method effectively turns positional information into distinguishable features for our Naive Bayes model.

Feature selection/ handling Stop Words:

- Based on the errors of the initial model, we determined that the model could benefit from reduced ambiguity. To reduce ambiguity, we implemented feature selection mechanisms, as well as handling stop words.
- Our approach to feature selection was based on mutual information scores, selecting the top 80% of words. This method inherently filtered out some of the stop words, as they

typically exhibit low mutual information due to their uniform distribution across classes. Filtering out the most common stop words alone was not found to be helpful.

Error Analysis

Misclassification Trends

Many of the trends found in the initial analysis were found once again:

***High Misclassification in 'Characters' Class**: The confusion matrix indicates a notable trend of misclassification involving the 'characters' class. Specifically, a considerable number of instances that were true 'performers' were incorrectly classified as 'characters' (n=7). This suggests a potential overlap in the language or context used to describe these two categories in the dataset. In addition, a significant number of ‘characters’ were misclassified as ‘director’ (n=6)*

***Director and Publisher Confusion**: Another observed trend was the misclassification between 'director' and 'publisher' classes. Although less pronounced than the characters-performers confusion, it indicates a likely similarity in the feature words between these two categories.*

However, there were improvements in reducing misclassifications involving ‘characters’ being labelled as ‘directors’.

Example: In line 233, after implementing additional preprocessing, the text was correctly classified. Initially, the text was classified as ‘directors’. Because “Ridley Scott”, is mentioned, the original classifier may have determined that the text is more director-aligned. It seems that with our improvements, the classifier can classify these ‘characters’ texts with more context, and can make more accurate classifications as a result.

- *844,"In Ridley Scott 's 1979 horror film "" Alien "" , Ellen Ripley and her crew attempt to capture an escaped Xenomorph with the help of a cattle prod .",characters,8,21*

Common Attributes in Misclassified Texts

Many of the misclassification attributes remained the same, even if certain misclassifications were reduced with our improvements:

Shared Contextual Words: Upon reviewing the misclassified texts, we noticed that they often contained similar contextual words or phrases. For example, terms related to artistic or creative works were present in both 'performers' and 'characters' classes, leading to confusion.

Example: Line 47, where a “character” label was misclassified with a “performer” label. Mentioning “Simpsons” and “Groundskeeper Willie” may have created *ambiguity, as these terms may be more relevant to performers instead of characters.*

- *1593,"The phrase was first popularized in the "" Simpsons "" episode "" ' Round Springfield "" (season 6 , 1995).Sound recording of Groundskeeper Willie 's line "" About : Political humour "" .",characters,8,23 24*

Example: Line 12, where a "director" label was misclassified with a "publisher" label. This specific misclassification does not occur very frequently. With further examination of the text, we can see that the mention of "Sony" may have interfered with classification as it is a publisher.

- *2359,"Following the release of "" Spider - Man 3 "" , Sony Pictures Entertainment had announced a May 5 , 2011 , release date for Sam Raimi 's next film in the earlier series .",director,5 6 7 8,25 26*

Lack of Distinguishing Features: Many misclassified texts lacked strong distinguishing features or keywords that could be decisively linked to one specific class. This was particularly evident in sentences that were shorter or less descriptive.

Identified Reasons Behind Errors

Insufficient Contextual Understanding: Although contextual understanding improved via head and tail entity positional context, the classifier still struggled with correctly classifying characters and performer classes. This may be due to the dataset itself, as the text for both of these classes contains very similar words. To further improve the model, more contextually advising mechanisms would need to be put in place.

Conclusion

Our Naive Bayes classifier effectively demonstrates the capability of classifying text into predefined relation categories. Integrating head and tail positional information and focusing on feature selection based on mutual information improved the model's performance. Future work may explore alternative methods for feature selection and further refinement of preprocessing strategies to enhance model accuracy and robustness.