XB_0085 | NERC, Sentiment and Topic Analysis by Group 43

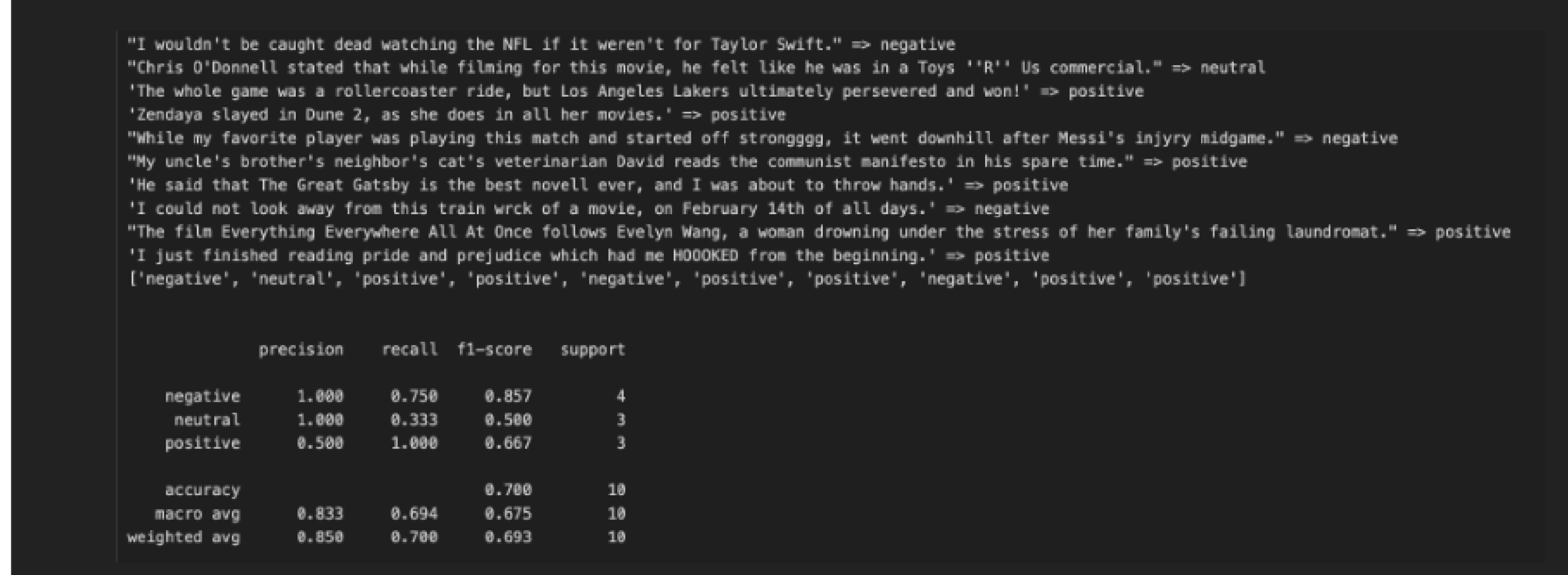
Sentiment Analysis

Approach

The main approach to conduct sentiment analysis was to train a Multinomial Naïve Bayes Classifier with an adequate training data annotated to have either a negative, a positive, or a neutral sentiment. Being one of the most common classifiers used for sentiment analysis, Multinomial Naïve Bayes Classifier utilizes discrete data where each feature represents the token frequency. With the flexibility to set a minimum threshold to disregard possibly noise making features with TfidfVectorizer, Naïve Bayes Classifier proves itself to be one of the easiest and most efficient approaches to implement. The working principle of the Multinomial Naïve Classifier is simple, it estimates the maximum likelihood for every feature to be belonging to a certain sentiment based on the previous words it has encountered within the files under that corresponding sentiment folder, and whichever sentiment outputs the maximum likelihood, the classifier classifies that word to have that corresponding sentiment.

Results

5 distinct experiments were conducted with 2 being bipolar classification (subjective/objective and positive/negative) and 3 being tripolar classification (negative/neutral/positive) utilizing CountVectorizer, TfidfTransformer, and MultinomialNB Classifier from sklearn library each within itself consisted of 20 distinct experiments with min_df values ranging from 1 to 20. One unique was also conducted utilizing VADER to predict the sentiments of test sentences, however, as the results obtained through VADER were incomparably accurate against Naïve Bayes Classifier, they were omitted from this section (for details, see the section “Relevant Links”). The requirement for 5 distinct experiment was to ensure the maximum accuracy is achieved, which is obtained to be 0.7 corresponding that 7 out of 10 sentiments were predicted correctly. The maximum accuracy was obtained through the collection of 30,000 tweets with equal distribution amongst the sentiments (10,000tweet/sentiment) with min_df value of 1 indicating that every word within the training data was integral in obtaining the optimal accuracy. The classification report obtained from this experiment is as below, where the sentences are the test sentences, and the arrows point towards the predicted sentiment:



```
"I wouldn't be caught dead watching the NFL if it weren't for Taylor Swift." => negative
"Chris O'Donnell stated that while filming for this movie, he felt like he was in a Toys ''R'' Us commercial." => neutral
'The whole game was a rollercoaster ride, but Los Angeles Lakers ultimately persevered and won!' => positive
'Zendaya slayed in Dune 2, as she does in all her movies.' => positive
"While my favorite player was playing this match and started off stronggggg, it went downhill after Messi's injyry midgame." => negative
"My uncle's brother's neighbor's cat's veterinarian David reads the communist manifesto in his spare time." => positive
'He said that The Great Gatsby is the best novell ever, and I was about to throw hands.' => positive
'I could not look away from this train wrck of a movie, on February 14th of all days.' => negative
"The film Everything Everywhere All At Once follows Evelyn Wang, a woman drowning under the stress of her family's failing laundromat." => positive
'I just finished reading pride and prejudice which had me HOOKED from the beginning.' => positive
['negative', 'neutral', 'positive', 'positive', 'negative', 'positive', 'positive', 'negative', 'positive', 'positive']

              precision    recall  f1-score   support

    negative      1.000     0.750     0.857         4
     neutral      1.000     0.333     0.500         3
    positive      0.500     1.000     0.667         3

    accuracy                          0.700        10
   macro avg      0.833     0.694     0.675        10
weighted avg      0.850     0.700     0.693        10
```

It can be observed from the classification report that precision scores for negative and neutral sentiments are 1.000, indicating that all the predicted negative and neutral sentiment were in fact the true sentiment of those sentences. However, the precision score for the positive sentiment is 0.500, indicating that half of the predicted positive sentiments did, in fact, contain another sentiment. Comparing our predictions to the true sentiments revealed that 3 positively predicted sentences had another sentiment as below:

5 My uncle's brother's neighbor's cat's veterinarian David reads the communist manifesto in his spare time.
6 He said that The Great Gatsby is the best novell ever, and I was about to throw hands. negative
8 The film Everything Everywhere All At Once follows Evelyn Wang, a woman drowning under the stress of her family's failing laundromat. neutral

As an addition, three word-clouds were generated as below consisting of colour-coded tokens prevalently used within the training data of their corresponding sentiment, where green represents positive sentiment, red represents negative sentiment and yellow represents neutral sentiment:

best say watching looking want awesome bed first think one fun hey yay know go new nice everyone oh day hope cool need time yes glad haha work wait good night days thanks really love sleep great morning yeah see today going people happy back home better thank tomorrow getting
funds presentation transcript stockmarket update million energy week investing news economy buy sahre trading fed trade financial results trump today spy top market ceo markets apple stocks says next earnings price group dividend https inc deal year reports call new billion nasdaq declares bank marketscreener business china conference
getting house week early twitter bed think school snow dont morning back time wish know day find hate im going sad bad class miss sick today feel tomorrow cant work one really sleep go last phone gome want need tonight long feeling night sorry cold hope tired though working

Discussion

Upon inspecting the training data, the reasoning behind the false predictions is revealed as follows:
For sentence 5, it can be observed that the sentence contains 4 distinct “’s” token which ranks as the 8th token amongst the tokens found in positive sentiment data but 10th amongst the tokens found in negative sentiment data and 12th amongst the token found in neutral sentiment data. This ranking combined with lack of other tokens within the neutral sentiment data promotes the sentence to be predicted as having a positive sentiment rather than a neutral sentiment.
For sentence 6, it can be observed that the tokens “great” and “best” rank 32nd and 88th amongst the tokens found in positive sentiment data whereas the tokens “throwing” or “hands” are lacking amongst the negative sentiment data. This imbalance, thus, promotes the sentence to be predicted as having a positive sentiment rather than a negative sentiment.
For sentence 8, it can be observed that as previously the token “’s” ranks 8th amongst the tokens of positive sentiment data whereas none of the tokens indicating the negative and neutral sentiment are found amongst negative and neutral sentiment data. This imbalance leads the sentence to be predicted as having a positive sentiment rather than a neutral sentiment.
The results allow us to gain a crucial insight on our training data, it is that the diversity and size of our training data is not adequate enough to exceed the 0.7 accuracy. This situation suggests that more training data, which must still be distributed evenly amongst the sentiments (previous experiments on imbalanced sentiment number failed) is needed to have a better accuracy. Thus, it can be deduced that our results are restricted with our training data, suggesting that the more we expand our training data while keeping the distribution amongst the sentiments balanced is the key to achieve a higher accuracy. The major challenge against this is the lack of training data annotated to have a “neutral” sentiment as the maximum number of tweets were found to be 10,000 whereas movie reviews to be 5,000 (for details, see the section “Relevant Links”).

Data Description

Various datasets were used during the experimentation process (see “Relevant Links” section for their access link). The initial training dataset was set to Cornell’s movie review dataset, however, as the accuracy based on that dataset was not exceeding 0.6, we have switched our approached and found a dataset containing 3 million tweets with each having sentiment label of either negative, positive, or neutral. Although the author has not disclosed the annotation process, upon closer inspection, the sentiment labels was uncovered to be reliable. One observation on that dataset was that the neutral sentiments were only 10,725 amongst the whole 3 million. Since we aimed to have a homogenous train data, and since storing all 3 million tweets would exceed our storage capacity, we have decided to parse only 10,000 tweets per sentiment. Thus, from the .csv file containing all the tweets, 10,000 per sentiment was extracted into separate files and put under their corresponding sentiment folders for the training to be done.

Topic Analysis with 2 Different Approach Comparison

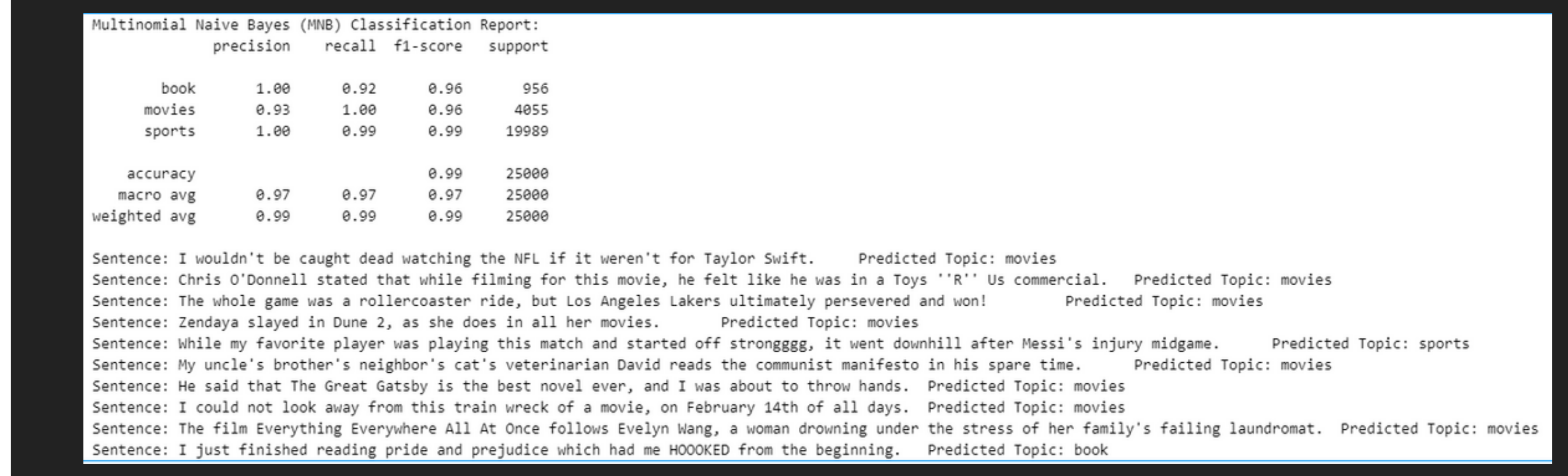
Approach

Two different algorithms has been used for topic modelling for comparison. The prefered platform is VSCode rather than Google Collab because of the inefficiency of data processing using the Web Browser. Multinomial Naive Bayes (MNB) has a probability based approach while assuming feature independence, and Support Vector Machine (SVM) is a margin based algorithm. Using them for the same task of topic analysis between 3 labels has showed the strengths and the shortcomings of these algorithms compared to analysis on similar models such as Bert and RoBERTa.

Both models have been trained on the same dataset, book and movie review data for their respective topics, and 2 sport tweets datasets for the topic of sports. Because of the difference in approach in the 2 algorithms, and the size difference of the data available, certain weight adjustments have been made especially for the data for movie reviews obtained through IMDB.

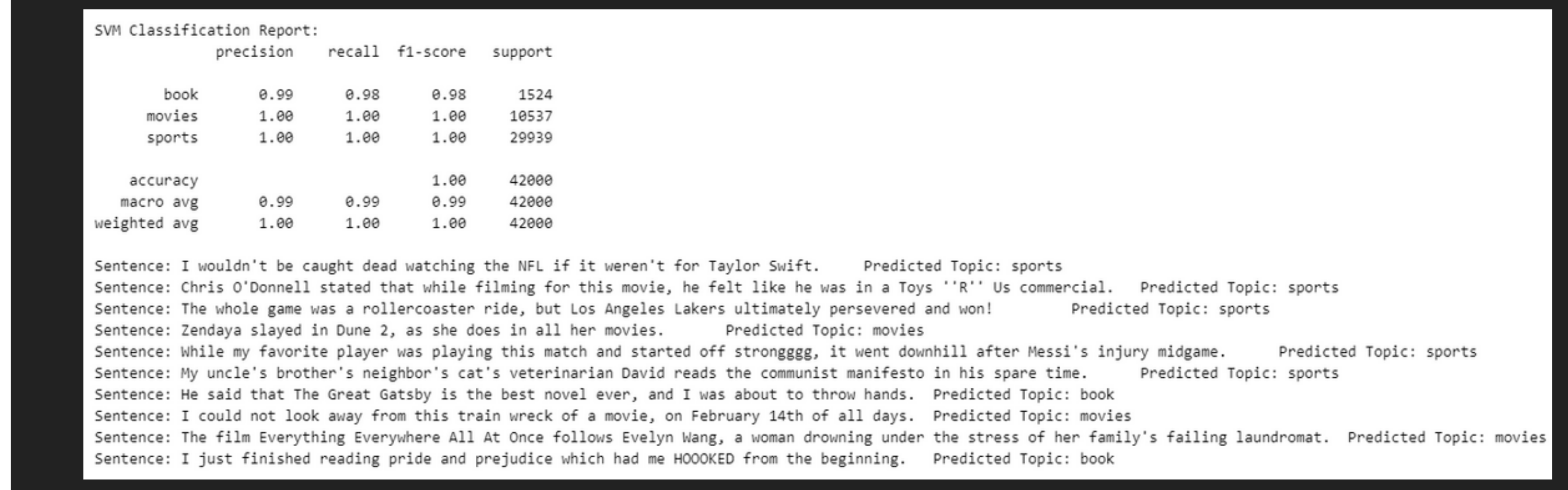
Results

Naive Bayes approach showed success in movies label more than others, and scored 6/10 out of the test sentences. The huge support for sports did not have the effect expected and the the model focused on the label of movies more. The average accuracy is %99 and f1-scores only drop to %96 for book and movies.



```
Multinomial Naive Bayes (MNB) Classification Report:
              precision    recall  f1-score   support

        book       1.00      0.92      0.96       956
      movies       0.93      1.00      0.96      4055
      sports       1.00      0.99      0.99     19989

    accuracy                           0.99     25000
   macro avg       0.97      0.97      0.97     25000
weighted avg       0.99      0.99      0.99     25000

Sentence: I wouldn't be caught dead watching the NFL if it weren't for Taylor Swift.     Predicted Topic: movies
Sentence: Chris O'Donnell stated that while filming for this movie, he felt like he was in a Toys ''R'' Us commercial.   Predicted Topic: movies
Sentence: The whole game was a rollercoaster ride, but Los Angeles Lakers ultimately persevered and won!        Predicted Topic: movies
Sentence: Zendaya slayed in Dune 2, as she does in all her movies.       Predicted Topic: movies
Sentence: While my favorite player was playing this match and started off strongggg, it went downhill after Messi's injury midgame.      Predicted Topic: sports
Sentence: My uncle's brother's neighbor's cat's veterinarian David reads the communist manifesto in his spare time.      Predicted Topic: movies
Sentence: He said that The Great Gatsby is the best novel ever, and I was about to throw hands.  Predicted Topic: movies
Sentence: I could not look away from this train wreck of a movie, on February 14th of all days.  Predicted Topic: movies
Sentence: The film Everything Everywhere All At Once follows Evelyn Wang, a woman drowning under the stress of her family's failing laundromat.  Predicted Topic: movies
Sentence: I just finished reading pride and prejudice which had me HOOOKED from the beginning.   Predicted Topic: book
```

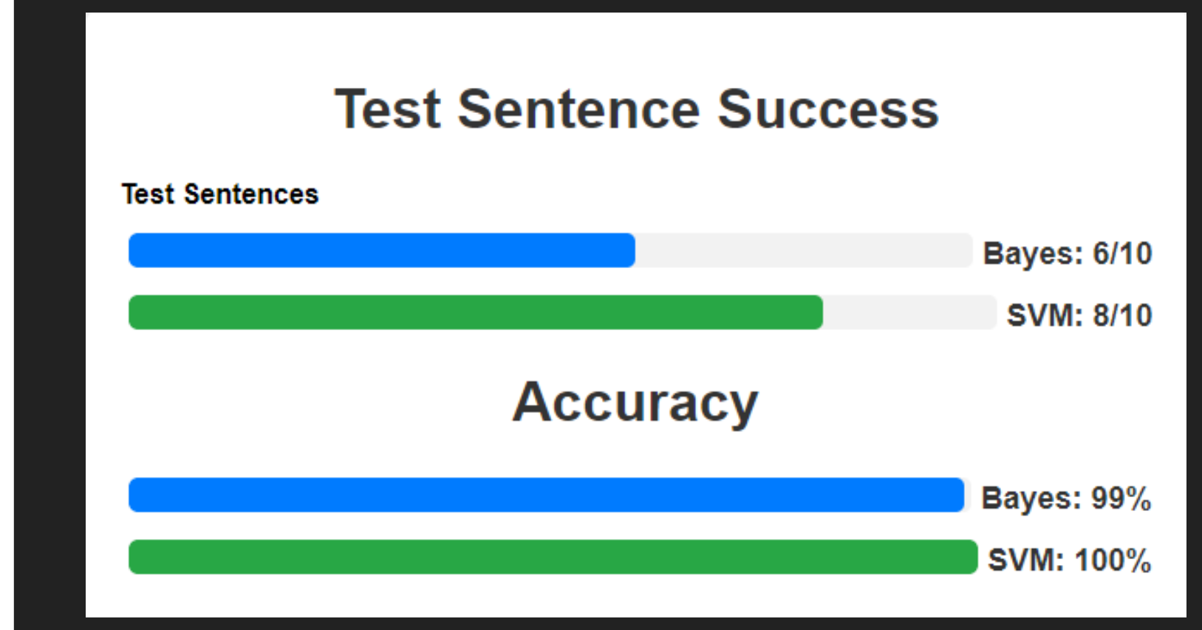
SVM had better results as it scored 8/10 out of the test sentences. The algorithm has been actually effected by the huge data present for the topic of sports. The algorithm generally had better scores than Naive Bayes approach, with a 1.00 accuracy and only dropping to %98 in f1-scores, which is for book label.



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

        book       0.99      0.98      0.98      1524
      movies       1.00      1.00      1.00     10537
      sports       1.00      1.00      1.00     29939

    accuracy                           1.00     42000
   macro avg       0.99      0.99      0.99     42000
weighted avg       1.00      1.00      1.00     42000

Sentence: I wouldn't be caught dead watching the NFL if it weren't for Taylor Swift.     Predicted Topic: sports
Sentence: Chris O'Donnell stated that while filming for this movie, he felt like he was in a Toys ''R'' Us commercial.   Predicted Topic: sports
Sentence: The whole game was a rollercoaster ride, but Los Angeles Lakers ultimately persevered and won!        Predicted Topic: sports
Sentence: Zendaya slayed in Dune 2, as she does in all her movies.       Predicted Topic: movies
Sentence: While my favorite player was playing this match and started off strongggg, it went downhill after Messi's injury midgame.      Predicted Topic: sports
Sentence: My uncle's brother's neighbor's cat's veterinarian David reads the communist manifesto in his spare time.      Predicted Topic: sports
Sentence: He said that The Great Gatsby is the best novel ever, and I was about to throw hands.  Predicted Topic: book
Sentence: I could not look away from this train wreck of a movie, on February 14th of all days.  Predicted Topic: movies
Sentence: The film Everything Everywhere All At Once follows Evelyn Wang, a woman drowning under the stress of her family's failing laundromat.  Predicted Topic: movies
Sentence: I just finished reading pride and prejudice which had me HOOOKED from the beginning.   Predicted Topic: book
```

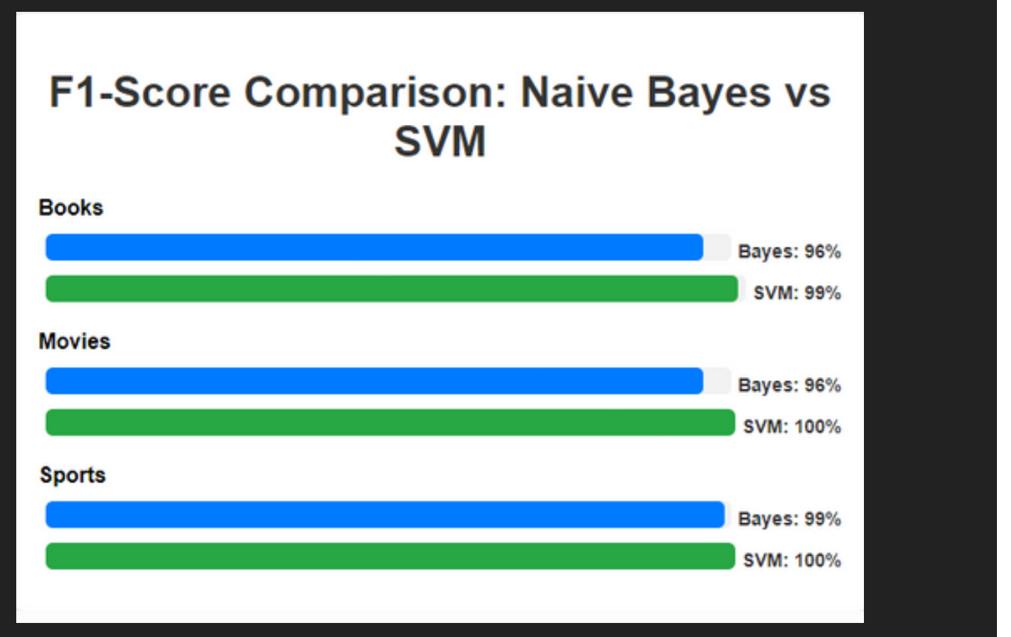
Multinomial Naive Bayes shortcoming of assuming feature indepence is displayed here as it sometimes cannot handle real-world scenarios, especially complex text data. Compared to MNB, the approach of SVM has shown better results because it manages to handle bigger data better. The sports dataset had a better effect on SVM and the model managed to recognize some of the test sentences labeled as sports. The SVM model also outperformed Naive Bayes approach in every f1-score and reported perfection in accuracy and 2 of the topic scores. These results should also be interpreted as the present data for the model is still relatively short of achieving these numbers as it can be seen from the test sentence success, the model also evaulates its performance with the same input, and these makes the results to be relevant only for inputs that are from, or linguistically really similar to the data presented for the model.

Discussion



Less represented Book data became problematic throughout the process as movie and sports data was more easy to find and fit to the job at hand. Adding more movie or sports related datasets made the model more focused on them and the lack of Book data halted some chances for further data improvement and gathering. Adding more book review or dialogue data can improve the model greatly and create chances for further data processing in other labels.
The general future improvements for both models would be increasing the data size for both labels and experimenting with different weights to adress the shortcomings of both algorithms. The quality of the data can also be a point to improve as Naive Bayes model had problems recognizing complex text that did not have clear patterns for it to analyze. These shortcomings could be handled with larger and targeted data.

Less represented Book data became problematic throughout the process as movie and sports data was more easy to find and fit to the job at hand. Adding more movie or sports related datasets made the model more focused on them and the lack of Book data halted some chances for further data improvement and gathering. Adding more book review or dialogue data can improve the model greatly and create chances for further data processing in other labels.
The general future improvements for both models would be increasing the data size for both labels and experimenting with different weights to adress the shortcomings of both algorithms. The quality of the data can also be a point to improve as Naive Bayes model had problems recognizing complex text that did not have clear patterns for it to analyze. These shortcomings could be handled with larger and targeted data.



Data Description

The data for the models have been collected mostly through NLP forums and Topic Analysis guides. All of the data is stored in a single column in a .csv file that the model can go through and label with the respective topic. All of the data are stored in .csv files and presented in a single column of “ReviewContent” that has been made so that processing them is easier and creates similar code.

For the label of book, Book Review Dataset from Kaggle has been used. The data is valuable as the text is rich in context and has many book names appearing repeatedly. And reviews has a different linguistic tone compared to test sentences given.

Same approach has been followed for movies, relatively big IMDB reviews dataset has been obtained through Kaggle, and the movies dataset has the same linguistic features as the book reviews. The dataset is relatively large compared to the book reviews, and because of this both models utilizes the ideal portion of it. Both models have been successful in terms of recognizing the label movies.

The limitation on book data limited the sports data similar to movies. Two different datasets of World Cup and Football-Soccer Twitter data has been used for labeling sports topic. Although the data is large in size, the negative effects has been less on the book sentences compared to movies.

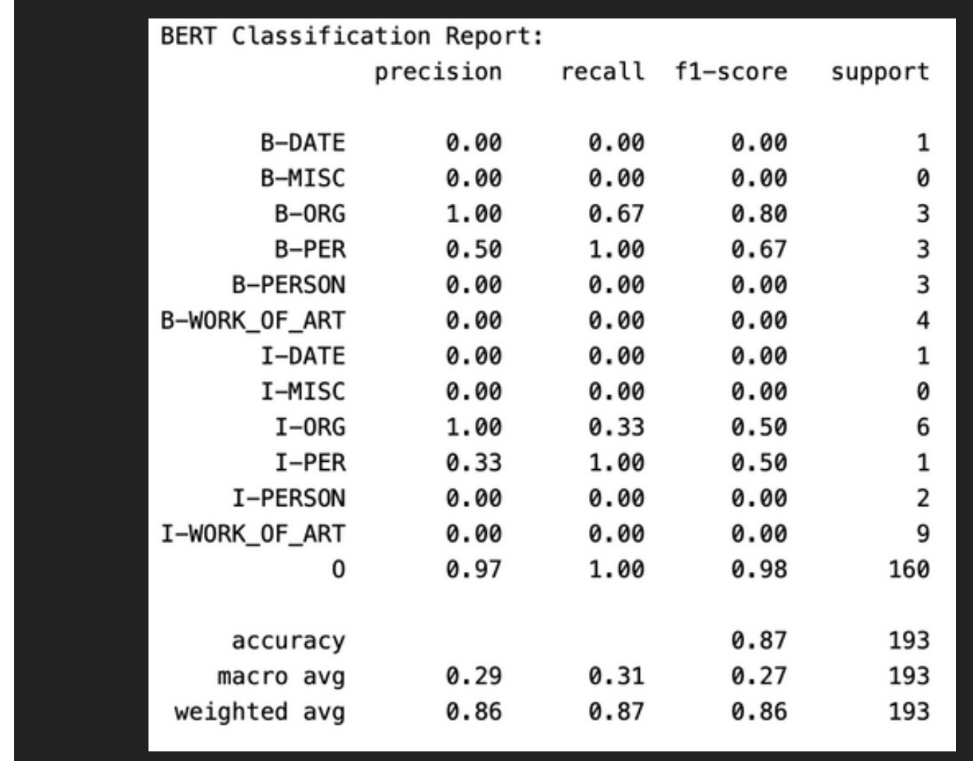
NERC

Approach

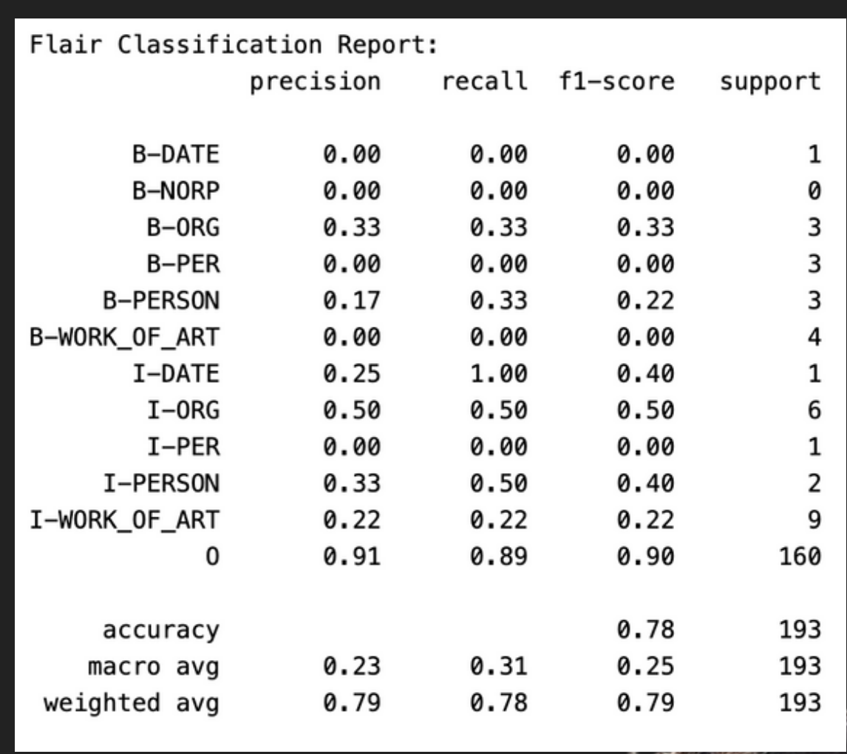
- For NERC, two different algorithms have been used as experiments: BERT and Flair. Both of the models used were fine-tuned transformer models for NERC from Hugging Face.
- The BERT model was chosen because of its fast performance and for its utilization of the bidirectional self-attention mechanism which can allow it to capture more contextual information. The Flair model was chosen for its wide range of entity categories.
- The BERT model was pre-trained on the CoNLL-2003 database and it has four types of entities: location, person, organization, and miscellaneous while the Flair model was pre-trained on the OntoNotes 5.0 corpus and it has 18 types of entities which include other entity types from the test data like date and work of art.

Results

- The BERT model returns a nested structure of token-label pairs. Flair model shows the entity’s boundaries as token indexes with the category of the entity.
- On the test data, the BERT model (0.87) showed better accuracy compared to the Flair model (0.78).

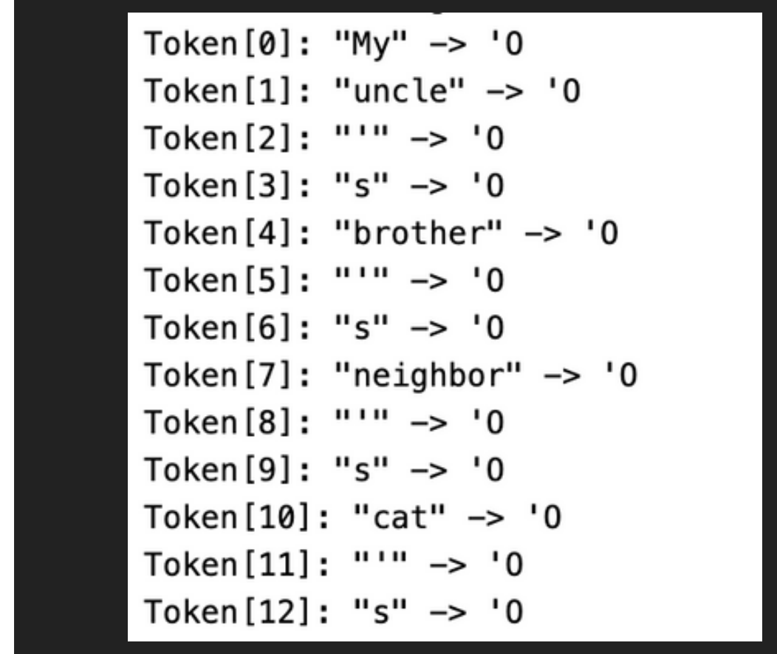


```
BERT Classification Report:
              precision    recall  f1-score   support

      B-DATE       0.00      0.00      0.00         1
      B-MISC       0.00      0.00      0.00         0
       B-ORG       1.00      0.67      0.80         3
       B-PER       0.50      1.00      0.67         3
    B-PERSON       0.00      0.00      0.00         3
B-WORK_OF_ART      0.00      0.00      0.00         4
      I-DATE       0.00      0.00      0.00         1
      I-MISC       0.00      0.00      0.00         0
       I-ORG       1.00      0.33      0.50         6
       I-PER       0.33      1.00      0.50         1
    I-PERSON       0.00      0.00      0.00         2
I-WORK_OF_ART      0.00      0.00      0.00         9
           O       0.97      1.00      0.98       160

    accuracy                           0.87       193
   macro avg       0.29      0.31      0.27       193
weighted avg       0.86      0.87      0.86       193
```



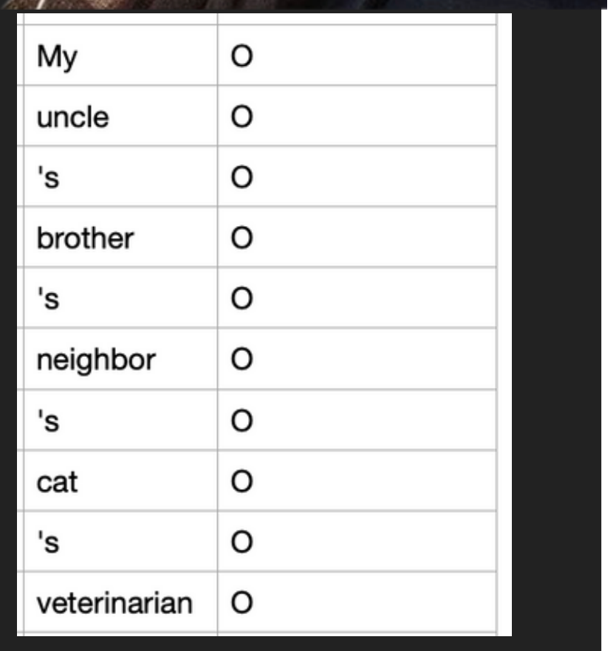
```
Flair Classification Report:
              precision    recall  f1-score   support

      B-DATE       0.00      0.00      0.00         1
      B-NORP       0.00      0.00      0.00         0
       B-ORG       0.33      0.33      0.33         3
       B-PER       0.00      0.00      0.00         3
    B-PERSON       0.17      0.33      0.22         3
B-WORK_OF_ART      0.00      0.00      0.00         4
      I-DATE       0.25      1.00      0.40         1
       I-ORG       0.50      0.50      0.50         6
       I-PER       0.00      0.00      0.00         1
    I-PERSON       0.33      0.50      0.40         2
I-WORK_OF_ART      0.22      0.22      0.22         9
           O       0.91      0.89      0.90       160

    accuracy                           0.78       193
   macro avg       0.23      0.31      0.25       193
weighted avg       0.79      0.78      0.79       193
```

- Since BERT had 4 types of entities it failed to tag the other types included in test dataset such as: B-DATE, I-DATE, B-WORK_OF_ART, and I-WORK_OF_ART. Instead, it labeled these categories as miscellaneous.
- The BERT model showed better performance on the entities: B-ORG, I-ORG, and O.
- Since Flair’s pretraining included the extra entities, it performed better than BERT at predicting the labels: I-DATE, I-WORK_OF_ART.
- The reason for Flair’s lower performance could be tokenizing issues as it was not successful in tokenizing the punctations at the end of the sentences and also combinations that included both word and punctations such as the “’s” in the phrase “cat’s” .



```
Token[0]: "My" -> 'O
Token[1]: "uncle" -> 'O
Token[2]: "'" -> 'O
Token[3]: "s" -> 'O
Token[4]: "brother" -> 'O
Token[5]: "'" -> 'O
Token[6]: "s" -> 'O
Token[7]: "neighbor" -> 'O
Token[8]: "'" -> 'O
Token[9]: "s" -> 'O
Token[10]: "cat" -> 'O
Token[11]: "'" -> 'O
Token[12]: "s" -> 'O
```

- The image on the left shows the results of Flair model labeling and the image on the right shows the ground truth labels of the testing dataset.
- Because of its mislabeling of the punctation as extra ‘O’ entities, the model loses accuracy.



My	O
uncle	O
's	O
brother	O
's	O
neighbor	O
's	O
cat	O
's	O
veterinarian	O

Discussion

- The BERT model was easier to use and performed better even though it had missing types in its labels. It is obvious that the BERT model will perform much better than the Flair model if it is trained with a train dataset containing extra label categories in the test dataset. However, finding such a new dataset and extra fine-tuning of the model adds extra stages to the program and reduces the ease of use of the model.
- Another shortcoming of the Flair model was that it did not include the IOB format. Therefore, we wrote a method to manually add IOB format to entity labels in the Flair model. Although this method worked properly, a completely accurate result could not be achieved due to the incorrect tokenizing procedures described above, and as a result, the accuracy scores were lower compared to the BERT model. However, Flair, which comes pre-trained with 18 entities, offers an accurate and ready-to-use model based on labels that are not in BERT’s model.

Data Description

- No datasets were used for further training as both of the models were pre-trained.

Work Distribution

As the assignment required 3 distinct tasks with one having at least 2 different approaches and considering that we are a group of 4, each member was mainly responsible of the coding of one task to ensure equal division of labour. After the coding is complete, each member commented on the results of others. After every member has approved on the final results, each member has written their own analysis and discussion, after which every member has commented on the analysis of other members. Having done the necessary alternations based on the peer review, we, as a group, gathered all of our work together in a single poster. Therefore, the work distribution is as follows:
Arda Cem Çakmak: Sentiment Analysis (Naïve Bayes and VADER), analysis, poster preparation.
Sinemis Toktaş: NERC (BERT, Flair), analysis, poster preparation.
Emre Akça: Topic Analysis (Multinomial Naïve Bayes), analysis, poster preparation.
Berk Yavaş: Topic Analysis (SVM), analysis, poster preparation.

Relevant Links

TOPIC ANALYSIS - MNB vs SVM
Sports Data:
https://www.kaggle.com/datasets/kumari2000/fifa-world-cup-twitter-dataset-2022
https://www.kaggle.com/datasets/eliasdabbas/european-football-soccer-clubs-tweets
Movies Data:
https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
Book Reviews Data:
https://www.kaggle.com/datasets/shrutimehta/amazon-book-reviews-webscraped

SENTIMENT ANALYSIS - NAIVE BAYES CLASSIFIER & SVM
Movie Reviews Data:
https://www.cs.cornell.edu/people/pabo/movie-review-data/
3 Million Tweets Data:
https://www.kaggle.com/datasets/prkhrawsthi/twitter-sentiment-dataset-3-million-labelled-rows

NERC - BERT, Flair
- BERT model: https://huggingface.co/flair/ner-english-ontonotes-large
- Fleir model: https://huggingface.co/flair/ner-english-ontonotes-large

Source Code:
https://github.com/acemcakmak/text-mining-project.git

Team 43 Members:
Arda Cem Çakmak
Sinemis Toktaş
Emre Akça
Berk Yavaş