IBM-322
PROJECT REPORT
Group 3

Submitted By:
Atharv Chhabra 21118025
Hrishit B P 21114042
Kirtankumar Vijaykumar Patel 21114051
Priyanshu Behera 21114077
Tanmay Bakshi 21122048

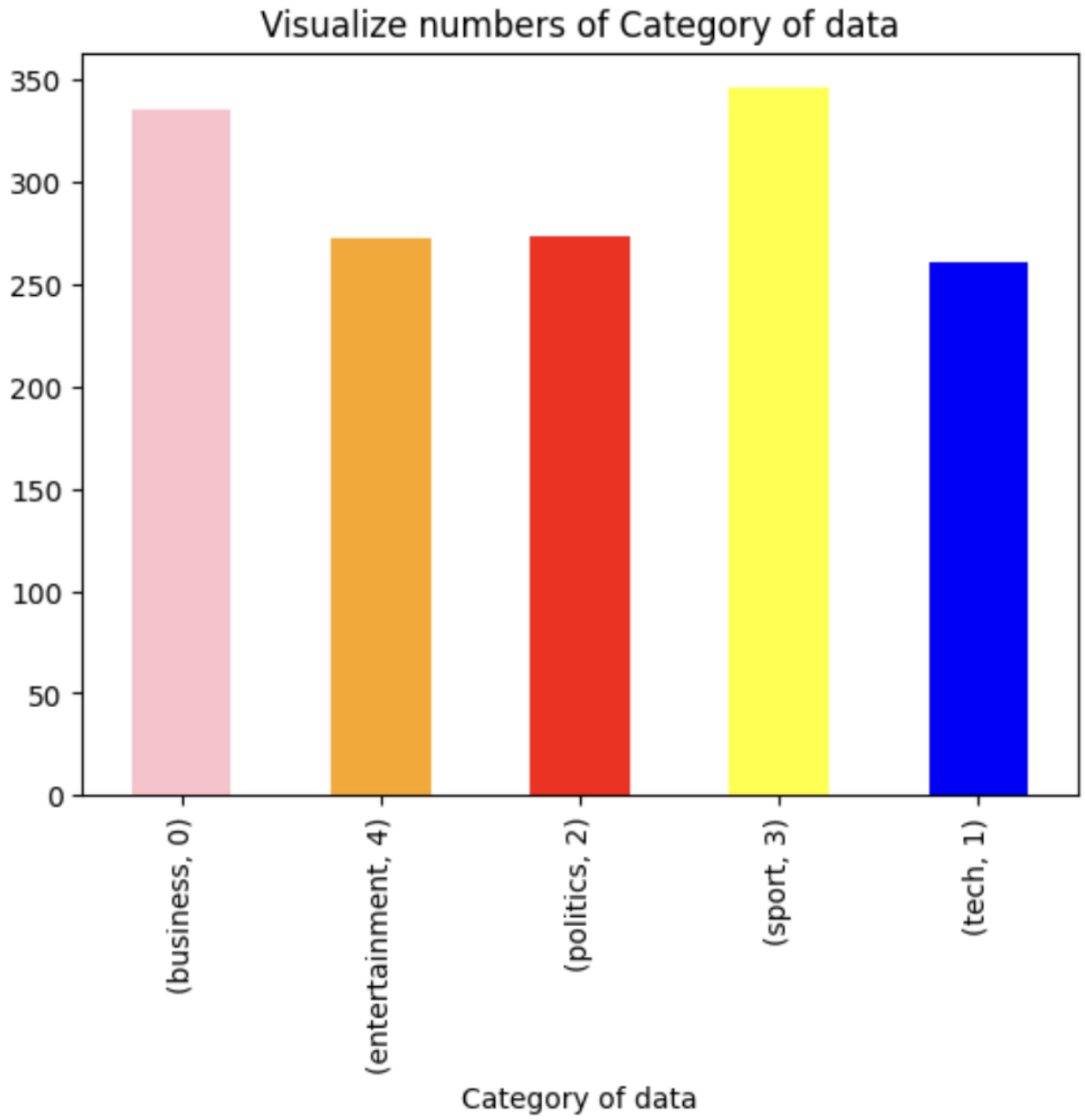
Motivation of the problem

Given a news article, we aim to predict which specific genre it belongs to. The test data is composed of articles from across 5 genres: Business, Entertainment, Politics, Sports and Tech.
We compare the performance of some Supervised learning algorithms, namely Softmax and Support Vector Classifier.

Snapshot of the data

	ArticleId	Text	Category
0	1833	worldcom ex-boss launches defence lawyers defe...	business
1	154	german business confidence slides german busin...	business
2	1101	bbc poll indicates economic gloom citizens in ...	business
3	1976	lifestyle governs mobile choice faster bett...	tech
4	917	enron bosses in $168m payout eighteen former e...	business
...	...	...	...
1485	857	double eviction from big brother model caprice...	entertainment
1486	325	dj double act revamp chart show dj duo jk and ...	entertainment
1487	1590	weak dollar hits reuters revenues at media gro...	business
1488	1587	apple ipod family expands market apple has exp...	tech
1489	538	santy worm makes unwelcome visit thousands of ...	tech

1490 rows × 3 columns


Visualize numbers of Category of data
350
300
250
200
150
100
50
0
(business, 0)
(entertainment, 4)
(politics, 2)
(sport, 3)
(tech, 1)
Category of data

Methodology

We've made use of majorly two models for prediction: Softmax and Support Vector Classifier.
First, we cleaned the given data making use of various techniques learned in the course such as:

- Conversion to Lowercase
- Tokenization
- Lemmatization
- Removal of Stopwords
- Taking care of "not" in sentences

To observe the major words used across the various articles, we constructed a word cloud of the cleaned text, again taking care to remove any peculiar stopwords, so as to obtain more relevant results.

Once the cleaning of the data was complete, we made use of TF-IDF in order to construct a term frequency matrix of the given articles. We didn't stop at using just DTM because it doesn't take into account the significance of a given word's usage across documents. For example, if a given word is observed in fewer documents, then it's more likely that the word holds more significance than some other word which is made use of in many more documents.

We then split the given data, 70% of it being test data and the remaining 30% being train data. We then trained a softmax model on the test data which would predict the genre of a given article. The accuracy of the model obtained was about 97.3%.

In order to be able to compare various models, we also made use of the Support Vector Classifier to do the same job i.e. predict the genre given an article. We trained it on the same test data that we obtained through splitting earlier and ended up observing an accuracy score of about 97.7%.

Analysis

It can be observed that the removal of stopwords significantly affects the model that we'd try to use later on. If we don't remove the stopwords, it would end up biasing the results and reduce the significance of our interpretation of the results as they would be based simply upon words which are made use of across documents, for e.g. articles such as "a", "an" and other stopwords that appear in a majority of articles such as "say", "however", etc. Thus their removal is essential to increasing the relevance of the results of the trained models.

The usage of TF-IDF is essential in order to account for the significance of words across articles. TF-IDF automatically assigns weight to terms based on their importance in the corpus. Terms that are frequent in a particular document but rare across the entire corpus receive higher weights. This can be beneficial because common words are often less informative and just made use of for conveying the meaning of the sentence, not the idea or topic being talked about in a given article. TF-IDF also accounts for the length of articles/documents through normalization, making it more robust to variations in document lengths.

We made use of two models to make predictions of the genre given an article: Softmax and Support Vector Classifier.

Softmax is a generalization of Logistic Regression for classification across more than 2 classes. In our case, each genre would be assigned a probability and the genre with the highest probability would be predicted as the output for a given article. In our test data, it's implicitly assumed that a given article is more probable to belong to only one genre, rather than belonging to multiple genres simultaneously.
Support Vector Classifiers can also be used for multiclass classification. It ends up computing a clear decision boundary which can be used to classify the articles across genres. The decision boundary is determined by the support vectors, which are instances that lie closest to the hyperplane.

Results

Classification Report for Softmax:

```
classification report
               precision    recall  f1-score   support

     business       0.97      0.98      0.98       103
         tech       0.99      0.95      0.97        77
     politics       0.97      0.94      0.96        81
        sport       0.97      1.00      0.98        97
entertainment       0.97      0.99      0.98        89

     accuracy                           0.97       447
    macro avg       0.97      0.97      0.97       447
 weighted avg       0.97      0.97      0.97       447
```

Classification Report for SVC:

```
 classification report
               precision    recall  f1-score   support

     business       0.96      0.99      0.98       103
         tech       0.99      0.97      0.98        77
     politics       0.99      0.93      0.96        81
        sport       0.99      1.00      0.99        97
 entertainment      0.97      0.99      0.98        89

     accuracy                           0.98       447
    macro avg       0.98      0.98      0.98       447
  weighted avg      0.98      0.98      0.98       447
```

Interesting findings/insights/observations

To reiterate, the usage of TF-IDF over just DTM increases the accuracy and relevance of the obtained results, since we've actually taken into account not just the frequency, but also the uniqueness of a word to a particular document.

In our case, a given article is more probable to belong to only one genre and since SVC is most effective when the goal is to find a clear margin of separation between classes, it ends up having a bit higher accuracy than Softmax. Thus we learn that the distinction boundary between classes is more defined in the case of SVC than Softmax.

Data source

BBC News Train.csv from BBC News Classification (Kaggle).
Link :
https://www.kaggle.com/competitions/learn-ai-bbc/data?select=BBC+News+Train.csv