**Topic Classification**

**Methods**

The methods used to classify the test texts were a BERT transformer and SVM. These models differ from each other. SVM is seen as a generic supervised learning algorithm that can classify texts by finding optimal divisions in the training set features (lab 2.2).

BERT(or transformers in general) differ by pre-training on a generic schema and then fine-tuning to the training set(Transformers: State-of-the-art natural language processing). Transformers use contextual embedding(lecture 4 - Sentiment Analysis) which takes the surrounding tokens into account the context, this assist in reducing ambiguity as there is more context in which a token is used. SVM does not take into account the context surrounding the token and only training using specific training data related to the topics.

Related to how small the test set is, comparing the 2 methods is to see if the basic nature of SVM is good enough to Classify accurately or if a large model technique such as transformers is necessary.

For the features used in SVM, the TF-IDF scores of the tokens will be used. This prevents generic words from seeming of the same importance as unique related to the topics (e.g. ‘was’ being as important as ‘restaurant’).

**Data**

For both methods, training data is needed, for SVM to train and for BERT to fine-tune. 3 datasets on Kaggle and processed them into the dataset names ‘train-dataset1.csv’

Dataset sources:

Restaurant reviews (topic classification)

<https://www.kaggle.com/datasets/fahadsyed97/restaurant-reviews?select=precovid_reviews.csv>

Book reviews

<https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews>

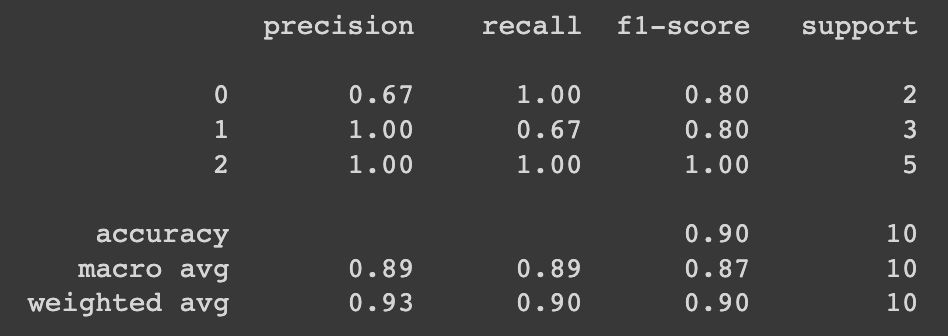
Movie reviews

<https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews>

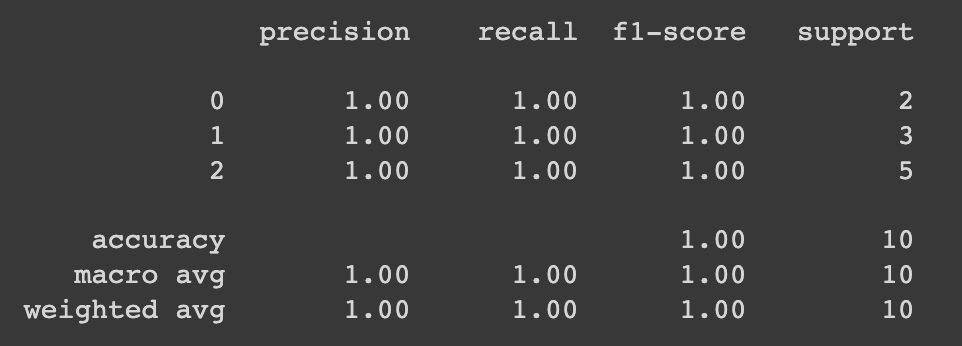
This data was imported, removed all columns except the text column, added the column ‘topic’ and input topic, select 1000 rows from each dataset, and then concated the datasets rows together to from ‘train-dataset1.csv’.

**Analyse and improvements**

SVM



BERT



The results show that BERT performed perfectly whilst SVM predicted the wrong label for one of the reviews which ultimately effected the recall of restaurant and the precision of book. BERT most likely got a perfect score due to its pre-training and contextual embedding allowing it to look at the context of the words. SVM did not look at the context of the words and connected certain tokens and features to certain classes. In the sentence that was a restaurant review but was classified as a book review, the token ‘diner’ (which was the only word associated to a restaurant review) was not located in the training set, so this word could not help classify the review and the rest of the feature were more related to book reviews.

BERT does not need improve as it got a perfect score. A method I would use to find improvements in SVM would be to make a function that would try different features to determine the best pick of feature to get the best performing SVM in this context. If this wouldn’t work the training set would be expanded to see if this improves the performance.

Luca Rosic: 1) coded topic classification, 2) analysied topic classification results 3) topic classification poster part

student 2: 1) coding, 2) analysis, 3) poster preparation

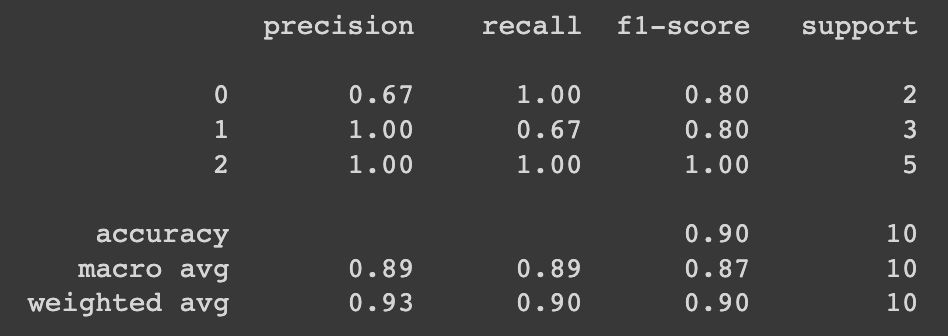
student 3: 1) coding, 2) analysis, 3) poster preparation

## Shortened version

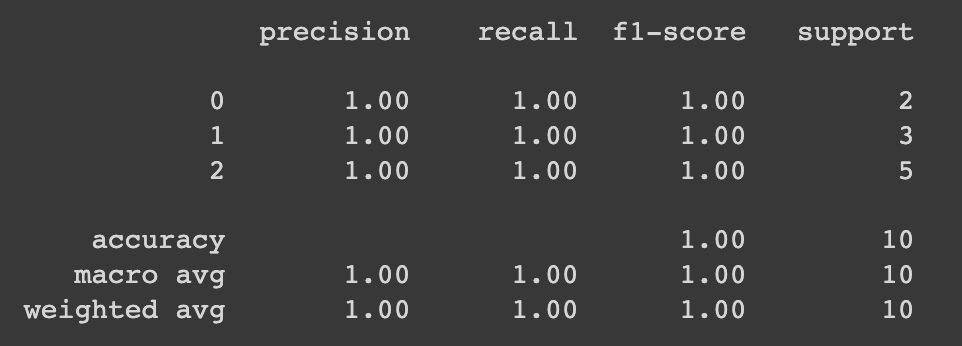
Methods used for topic classification include SVM and BERT transformers. SVM is seen as a generic supervised learning algorithm that can classify texts by finding optimal divisions in the training set features (lab 2.2)(Multi-category news classification using Support Vector Machine based classifiers). The features used in the SVM was the TF-IDF scores of tokens for classification, while BERT pre-trains on a generic schema and then fine-tunes to the training set(Transformers: State-of-the-art natural language processing). Transformers use contextual embedding, which reduces ambiguity by taking surrounding tokens into account(lecture 4 - Sentiment Analysis)

For the test, three datasets on Kaggle were used and processed into a single training set with 3000 date points(1000 for each review topic).

SVM



BERT



SVM predicted one review incorrectly, resulting in a lower precision score for book and lower recall score for restaurant. BERT achieved a perfect score due to its pre-training and contextual embedding. To improve SVM, a function could be used to find the best feature by training different features to train the model or the training set could be expanded.

Luca Rosic: 1) coded topic classification, 2) analysied topic classification results 3) topic classification poster part

student 2: 1) coding, 2) analysis, 3) poster preparation

student 3: 1) coding, 2) analysis, 3) poster preparation