TEXT MINING PROJECT

Group 37: Elif | Ashilla | Elisa | Yuki

1 Motivation

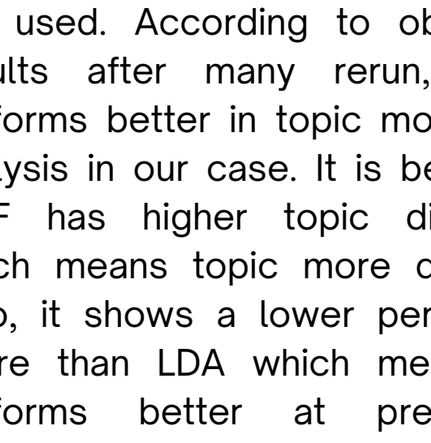
Sentiment Analysis: We compare the performance of two rule-based pre-trained models (TextBlob and VADER) and one supervised machine learning model (Naive-Bayes). Our motivation for this study is to provide a comprehensive analysis of the strengths and weaknesses of each approach, as well as to identify the most suitable method for sentiment analysis in a given context.

Topic Analysis: Our motivation for conducting topic analysis is to uncover underlying topics in the datasets. Additionally, the training datasets we are working with do not have explicit labels for each domain. As stated in the lecture, when labelled data is not available or missing, an unsupervised model is the appropriate solution. We will be using LDA and NMF for the topic modeling, which are suitable for our aim and relatively large datasets.

NERC analysis: NER is used in various applications, it's particularly useful for quickly extracting key information from large amounts of data because it automates the extraction process. The method applied for our NERC analysis consists of the follow steps: the test set is cleaned by removing unused columns. Then the names of the columns are changed to match with the nltk.corpus of the training set. SpaCy is used to predict the unknown labels for this dataset. The train and test features are concatenated into one list after the label prediction for the test set and then split back into training and test. A SVM is trained afterwards using the train features and labels and the test data is evaluated. The classification report shows an accuracy of 0.93. However, the test set is relatively small and it consists mainly of O's, this label is well predicted with a high recall, precision and f1-score. The other labels are not well predicted, they have lower recall, precision and f1-scores. This results in a low macro average. So we should keep in mind that although the accuracy of this model is high, the performance is not necessarily good. From the obtained results, we could conclude that this model is good at detecting "O" labels but not as good as predicting the other labels.

The NERC analysis dataset:

The dataset for training is from Crossweigh this training set contains 203621 instances where the majority of the training set consists of O-tags. The distribution of the training labels is plotted in a pie chart shown down below. The other golden labels (the labels that the NERC model should learn predict) from are: B-LOC (3.5%), B-PER (3.2%), B-ORG (3.1%), I-PER (2.2%), I-ORG (1.8%), B-MISC (1,7%), I-LOC (0.6%), I-MISC (0.6%). This shows that the training set is not balanced, there are far more O-labels.



2 Data

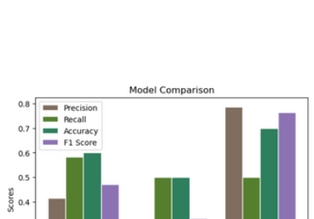
The sentiment analysis dataset comprises tweet texts, each labeled with a sentiment category of either "Positive," "Negative," or "Neutral." This dataset was chosen because the sentiment labels are necessary to train a model for sentiment analysis. Additionally, the dataset size is sufficiently large to ensure effective training of the model. By utilizing this dataset, we aim to develop a model that can provide us good result on the test set

The topic analysis dataset was obtained from three medium-sized training datasets from Kaggle: restaurant reviews, book reviews, and movie reviews. These datasets were chosen because they cover relevant examples that can be used to train the model, and they contain a large number of data points which provide a good representation of the topics and diverse perspectives related to each topic. Increasing the number of features in the dataset can improve the performance of the model, as demonstrated by Mikolov et al, which was one of our goals. Additionally, Jurgens et al. showed that a diverse training dataset can also improve the performance of an analysis model. The test data contains texts that reflect topics of interest in reviews about restaurants, books, and movies.

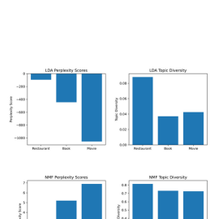
3 Discussion & Analysis

After conducting an analysis of our models, we created confusion matrices to examine the true positive and false positive results. This allowed us to make a more thorough comparison of each model's performance. We also compared each model's precision, accuracy, recall, and F1 scores to determine which model performed the best. After reviewing the results, we concluded that with the training set we provided, the VADER model outperformed the other models. Its superior performance in these metrics indicates that VADER is the most effective model for our purposes.

To compare the performance of the model for topic analysis, a perplexity score and topic diversity are used. According to obtained results after many rerun, NMF performs better in topic modelling analysis in our case. It is because NMF has higher topic diversity which means topic more diverse. Also, it shows a lower perplexity score than LDA which means, it performs better at predicting unseen data.



Error analysis for sentiment analysis



Error analysis for Topic analysis

4 Limitations

To improve the accuracy of **sentiment analysis**, we could explore more complex machine learning models and use more relevant and diverse datasets. However, due to time constraints, we had limited ability to do so. Additionally, while TextBlob, Vader, and Naive Bayes are useful, they have limitations in capturing the complexity of language and assumptions about feature independence, which may lead to suboptimal results.

To improve the perplexity of the **topic analysis,** prior knowledge and constraints can guide topic modelling toward more relevant topics. Testing different hyper parameters and techniques for preprocessing data can also improve performance, the use of larger. To address data spareness and ambiguous topic, regularization technique can be used. For **NERC analysis** the limitation lays in the prediction of I-labels. One way to improve this could be improving data preprocessing: The quality of the input data can affect NERC performance significantly. Therefore, cleaning and normalizing the data can improve NERC accuracy for predicting I-labels. Another way to improve the performance could be providing contextual information which can include information about the surrounding words and their relationship to the named entity.

Division of work: *We divided the work based on the model. Elif and Elisa were responsible for sentiment analysis, Ashilla worked on topic analysis, and Yuki focused on NERC analysis. As a team, we collaborated on the poster design.*

Citations & Code: