INTRODUCTION

For the final project, three Natural Language Processing (NLP) techniques have been be applied to a pre-provided text consisting of ten pre-labeled sentences of movie, book and restaurant reviews. These techniques consist of Named Entity Recognition and Classification (NERC), sentiment analysis, and topic analysis. For each technique a system is chosen to perform the NLP-technique, after which the results are analyzed. For topic analysis, RoBERTA and SVM have been applied and compared to each other. We hypothesize that RoBERTa will perform better than the SVM model, since it is a powerful language model that also considers contextual representation of the words. Concerning NERC, Conditional Random Field (CRF) has been applied and we hypothesize a strong performance with solid training data and proper feature extraction. For sentiment analysis, VADER has been applied. Our hypothesis states that its performance will be undesirable due to the misinterpretations that VADER makes based on its lexicon.

DATA PREPROCESSING

The CONLL2003 dataset is used as a training set for Named Entity Recognition and Classification (NERC). Concerning the training set for topic analysis and sentiment analysis, we first collected three different datasets on movie, book and restaurant reviews [1] [2] [3]. We then select only the relevant columns and add labels manually to each dataset. Since the restaurant reviews dataset contains only 1,000 instances and the two other dataset have over 10,000 instances, we decided to take 1,000 samples from all the datasets and combine them into a single dataset of 3,000 lines. This also ensures that the distribution of labels is balanced for all the topics in the final combined dataset.

While observing the distribution of the amount of NERC labels for the CONLL2003 training dataset, it can be noted that the data is fairly equally balanced except for the I-MISC and I-LOC, where there are significantly less instances.

METHODOLOGY

NERC

For Named Entity Recognition and Classification we made use of a CRF as a model. It allows for modeling dependencies between neighboring tokens [4]. As parameters we passed 'l2sgd' - Stochastic Gradient Descent with L2 regularization term as the algorithm, we set 'max iterations' to 100, and set 'all_possible_transitions' to False. Both the CONLL2003 dataset and the pre-provided test set are preprocessed by extracting features on a per sentence basis. Extracted features include for every word and the direct neighbors of every word the Part-of-Speech (POS) tag (NLTK was used to get POS tags for the test set), lowercased version and a Boolean value for if it is capitalized or a digit. The CRF model is trained on these features and then predicts IOB-labels for the provided test set.

Sentiment Analysis

For sentiment analysis, we decided to use Valence Aware Dictionary for Sentiment Reasoning (VADER) as a model. We decided to use VADER for sentiment analysis as it can assess sentiment to sentences without any training. VADER returns sentiment scores between 0 and 1 representing the categories negative, positive and neutral using the VADER lexicon. Furthermore, a compound score between -1 and 1 is provided that describes the overall sentiment of the text. To test our model, polarity scores were assigned to each sentence in the test set. This returns a dictionary of sentiment proportions for each of the categories [5]. The compound is then used as a feature to VADER to output the sentiment label of a sentence. If the compound is smaller than 0, a negative label will be asserted to the sentence. If the compound is greater than 0, a positive label will be asserted to the sentence. If the compound is 0, a neutral label will be asserted to the sentence.

Topic Analysis

For topic analysis, two separate supervised methods by using the RoBERTa transformer model and a support vector machine model (SVM) are employed. For the SVM, we pre-processed all words for each sentence and take those as a feature to feed into the DictVectorizer. The SVM is then trained on these vectors with the combined dataset that we mentioned in the data preprocessing section. The RoBERTa transformer uses the default settings of the roberta-base model, we only adjust the epochs to 5 due to time limitations and the learning rate to 1-e4 for standardization. The reason for using RoBERTa is because this language model is trained on a large dataset and takes into account the contextual representation of the words in a sentence compared to the BERT model [6]. The reason for choosing a SVM model is because we have labeled data and SVMs can be used for supervised classification approaches with fewer classes (three classes in our case) [7], and are desirable for small datasets.

RESULTS + ANALYSIS

Named Entity Recognition and Classification

As can be observed in the classification report below, for most of the IOB-labels the CRF model has done correct labeling, resulting in an f1-score of 1.00. As an exception the model has a decrease in performance for the IOB-labels ORG and PER, both for B and I. When closely examining the predicted labels the decrease in performance can be attributed to three separate cases of mislabeling. The first is labeling "Cuba Gooding Jr." as ORG instead of PER. The second is labeling "Blauwbrug" as PER instead of ORG. Finally, not recognizing "Dame" in "Dame Maggie Smith" as part of a PER label. A concrete explanation for confusing a person for an organization and vice versa is hard to identify, perhaps the CONNL2003 dataset could be improved upon by extending on data concerning these two labels. The CONNL2003 dataset is built up out of Reuters news stories from '96 and '97. One possibility is that PER labeling such as 'Dame' has not occured during this period.

	precision	recall	f1-score	support
B-LOC	1.00	1.00	1.00	4
B-MISC	1.00	1.00	1.00	3
B-ORG	0.75	0.75	0.75	4
B-PER	0.67	0.67	0.67	6
I-LOC	1.00	1.00	1.00	2
I-MISC	1.00	1.00	1.00	1
I-ORG	0.60	1.00	0.75	3
I-PER	1.00	0.62	0.77	8
O	0.99	1.00	1.00	183
accuracy			0.97	214
macro avg	0.89	0.89	0.88	214
weighted avg	0.98	0.97	0.97	214

Sentiment Analysis

VADER makes use of a lexicon-based approach in order to assign sentiment labels to sentences. As we can see from the final results that can be observed in the classification report below, the VADER model has an overall accuracy of 0.60. This is not a desirable score, as it only accurately assigned sentiment labels to 6 out of 10 sentences of the test set. However, we can analyse the output and see why VADER assigned the wrong sentiment to some of these phrases (see source code for the full output). In the fifth sentence, VADER assigned a positive sentiment to a sentence that has a neutral gold label. This can be explained by the fact that the word 'played' has a positive score in the lexicon, which thus makes the sentence positive to VADER. However, in this particular sentence the word 'played' refers to an actor playing in a movie, as opposed to the verb 'played' that refers to having fun. In the final sentence, VADER assigned a positive sentiment to a sentence that has a negative gold label. This can be explained by the fact that the word 'loved' has a positive score in the lexicon, which makes the sentence positive. VADER makes sure that all sentiment-bearing words before word "but" have their valence reduced to 50% of their values, while those after the "but" increase to 150% of their values. However, there are no sentiment-bearing words after the word 'but' in this sentence, so even though it is a negative sentence, VADER will assign a positive sentiment.

	precision	recall	f1-score	support
negative	1.00	0.33	0.50	3
neutral	1.00	0.33	0.50	3
positive	0.50	1.00	0.67	4
accuracy			0.60	10
macro avg	0.83	0.56	0.56	10
weighted avg	0.80	0.60	0.57	10

Topic Analysis: Method 1

The RoBERTa model is able to correctly identify the topic of all of the sentences that belong to each of the categories; namely, restaurants (labeled with 2 in the classification report), books (labeled with 0 in the clasisifcation report) and movies (labeled with 1 in the classification report). As can be observed from the classification report below, the RoBERTa model has a precision, recall and f1-score of 1.00 for each of the topics, resulting in a overall weighted average, macro average and accuracy of 1.00. As noted before, RoBERTa generates contextualized word representations [6]. This can make a model more robust as it can understand nuances in language by recognizing the different meanings of certain words in different contexts. Furthermore, RoBERTa has been pre-trained on a large dataset and is also trained for a long time (500k pretraining steps) [8]. Additionally, it uses effective training techniques such as dynamic masking and large mini batches which improves its performance [8]. Its perfect performance can also be explained by the fact that the test set is very small, containing only 10 sentences, while it has been trained on a much larger dataset. Overall, the RoBERTa model is extremely robust.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2
1	1.00	1.00	1.00	5
2	1.00	1.00	1.00	3
accuracy			1.00	10
macro avg	1.00	1.00	1.00	10
weighted avg	1.00	1.00	1.00	10

Topic Analysis: Method 2

The SVM model has an accuracy of 90%. By looking further in depth at the precision and recall of the model for each of the topics, we observe that the movie reviews have all been correctly classified with precision=100% and recall=100%. The model also correctly classifies all the book reviews that it recognizes with precision=100%, but fails to recognize all the book reviews in the test set, which results in a recall of 50%. It is also important to note that size of the book reviews in our dataset is only two instances. For the restaurant reviews, the SVM model has recall=100% which suggest that the model could recognize all the restaurant instances in the dataset. The precision for this topic is 75%, meaning that only two instances were correctly classified out of the three restaurant instances in the test set. Overall, we can observe that the model performs notably and is able to classify most of the reviews in their correct topics by looking at the macro average of all the results [9]. Since SVM models generally perform best for binary classification problems, the obtained results are expected for our classification task with three classes [10]. The final results can be observed in the classification report below.

	precision	recall	f1-score	support
book	1.000	0.500	0.667	2
movie	1.000	1.000	1.000	5
restaurant	0.750	1.000	0.857	3
accuracy			0.900	10
macro avg	0.917	0.833	0.841	10
weighted avg	0.925	0.900	0.890	10

TEXT MINING PROJECT

GROUP 31

Zahra Moradi (2690281)
Katrina Slebos Perez (2714445)
Mara Spadon (2688689)
Alexander van der Linden (2508637)

TASK DIVISION

Zahra	Katrina	Mara	Alexander
Coding: SVM	Coding: RoBERTa	Coding: VADER	Coding: NERC
Analysis: SVM	Analysis: RoBERTa	Analysis: VADER	Analysis: NERC
Poster: Introduction, Data Preprocessing, Methodology and Future Work	Poster: Introduction, Methodology, Conclusion and Future Work	Poster: Overall poster layout, Introduction, Conclusion and Future Work	Poster: Introduction, Methodology, Conclusion and Future Work

CONCLUSION

This paper provided an in-depth analysis of the performance of various techniques for NERC, sentiment analysis, and topic analysis. For NERC, better performance was reached than expected. Despite a few small labeling errors, the CRF model was able to correctly label most IOB-labels. For sentiment analysis, VADER did indeed have an undesirable performance, as it often misinterprets the meaning of verbs by not taking context into account. As expected, RoBERTa performed better than SVM regarding the topic analysis. This can be explained by the robustness of RoBERTa and the observation that SVM performs less optimally for non-binary classification problems [10]. This implies that feature engineering and contextual representation play a crucial role in the performance of text analysis models. To conclude, our hypotheses that were stated in the introduction turn out to be correct.

FUTURE WORK

Concerning the CRF model for NERC, additional research could be carried out for the analysis and evaluation of the model through parameter tuning and identifying overfitting and possibly combatting it with regularization. Some initial experimentation did not seem to affect the end performance of IOB-tagging. For sentiment analysis, the wrongly classified sentiments were mostly a result of VADER's inefficiency to understand the context and naunces of language. Hence, sentiment analysis could be improved by employing a language model instead to consider the context. This would prevent the first mistake that was described in the analysis. The performance of SVM is heavily correlated with the quality of the feature engineering method for topic analysis. Our framework currently only takes a set of preprocessed words of the topics to train the SVM model. However, with the addition of the tf-idf of these words we can further enhance the classification results [11]. Lastly, since RoBERTa had a perfect performance we could test it on a bigger test set to make sure the predictions are accurate on a larger scale as well.

REFERENCES

Link to source code /experiments:
https://drive.google.com/drive/folders/1v_PKNwkI-HaIzz9TEWtdF0BaYvjkee7?usp=share_link

[1] https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset

[2] https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

[3] https://www.kaggle.com/datasets/vigneshwarsofficial/reviews

[4] Markov et al. (2023). Lecture 5: Named entity detection and classification [Slide 21]. Vrije Universiteit Text Mining For AI Canvas.

[5] Markov et al. (2023). Lecture 4: Subjectivity mining [Slide 36]. Vrije Universiteit Text Mining For AI Canvas.

[6] "Overview of ROBERTa model," GeeksforGeeks, 10-Jan-2023. [Online]. Available: https://www.geeksforgeeks.org/overview-of-roberta-model/. [Accessed: 01-Apr-2023].

[7] "Learn How to Use Support Vector Machines (SVM) for Data Science," Ray, S., 2017. [Online]. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/. [Accessed: 01-Apr-2023].

[8] Y. Lui et al., Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[9] Markov et al. (2023). Lecture 3: Machine Learning for NLP [Slide 9]. Vrije Universiteit Text Mining For AI Canvas.

[10] "Support Vector Machines," Scikit-learn Developers, 2007. [Online]. Available: scikit-learn. https://scikit-learn.org/stable/modules/svm.html. [Accessed: 01-Apr-2023].

[11] Saigal, P., Khanna, V. Multi-category news classification using Support Vector Machine based classifiers. SN Appl. Sci. 2, 458 (2020).