TEXT MINING FOR AI: PROJECT POSTER

Group 23 - Arshana Ramautar, Federico Signorelli, Rafaëlla van Nee, Viktoriya Li

APPROACHES

TOPIC ANALYSIS

NASVM (WORD EMBEDDINGS)

The Support Vector Machine (SVM) algorithm is recognized for its versatility and efficiency in training across various applications. A notable benefit of employing SVM is its capacity to handle extensive feature sets without necessitating the kernel trick, thereby streamlining the training process. This characteristic renders it particularly efficacious for tasks such as text polarity detection, a domain where it enjoys widespread application. For Topic Analysis we use the multinomial naive bayes. Similar to its application in sentiment analysis, the Multinomial Naïve Bayes classifier demonstrates robust performance across a variety of text classification tasks, including topic analysis. It operates by predicting the topic of new documents through an analysis of the frequency with which words or features appear within the text. As with previous implementations, the absence of hyperparameters significantly simplifies the configuration process for this model, facilitating an efficient setup.

NERC

CONDITIONAL RANDOM FIELD (CRF)

Conditional Random Fields (CRFs) are particularly suited for the Named Entity Recognition and Classification (NERC) task due to their ability to model contextual dependencies and integrate several features for sequence tagging [1]. NERC involves identifying and classifying entities in text into predefined categories, which makes it a supervised model. As the task benefits from understanding the context in which entities appear, CRFs attempt to capture these contextual relationships by considering the sequence and correlation of labels and input samples. As CRFs excel particularly in NERC, we decided to use this technique for this task.

SENTIMENT ANALYSIS

DISTILBERT

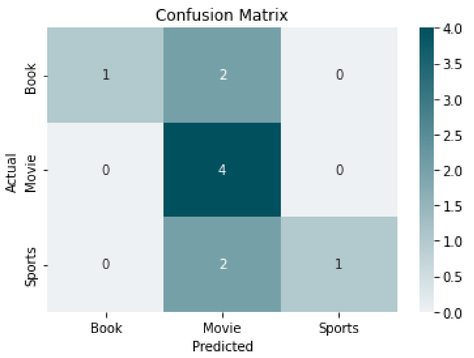
For sentiment analysis, we also use BERT (Bidirectional Encoder Representations from Transformers). This system considers not just individual words but also their surrounding context. By training on large amounts of text data, BERT learns to recognize subtle cues and linguistic nuances that convey sentiment [4]. This enables BERT to accurately determine whether a piece of text expresses positive, negative, or neutral sentiments towards a given subject. DistilBERT chosen as it is smaller and faster than BERT but retains most of its performance, it being a compact version of it.

MNB

We chose the Multinomial Naive Bayes (MNB) algorithm because of its simplicity and performance, particularly in the domain of text classification. A key advantage of using MNB lies in its probabilistic approach to handling large feature spaces typically associated with text data [5]. By calculating the conditional probability of each class based on the frequency of words, MNB can efficiently manage high-dimensional data. This efficiency, combined with its assumption of feature independence, allows for quick training times and robust performance, making it a popular choice for tasks like sentiment analysis and topic categorization where it is often applied.

VADER

In sentiment analysis, we use VADER (Valence Aware Dictionary and Sentiment Reasoner) which is a widely recognized tool for assessing emotional valence, due to its established utility in our previous lab sessions [3]. VADER operates by leveraging a lexicon where lexical features are correlated with specific emotion intensities, referred to as sentiment scores. This approach enables the precise evaluation of sentiment within text data.



DATA DESCRIPTION

NERC

To train the CRF on the NERC task we made use of the CoNLL 2003 NER labeled dataset, previously seen in lab session 4. It is a widely established set containing high quality annotations of the types Person (PER), Organization (ORG) and Location (LOC).The method of annotation follows the IOB (Inside, Outside, Beginning) tagging scheme that the test set also follows. As the dataset is already highly curated and our online research did not lead to any better alternative, we made unaltered use of the CoNLL 2003 for the training. Differently, we noticed a labeling error in the test data, when some NEs of the type person were classified using the PER tag, while the other ones used the PERSON tag. We corrected this by modifying the PERSON tag to the PER tag. To make use of the CoNLL dataset, we prepared features corresponding to the current and previous words, already structured and labeled in the dataset, to feed to our CRF model.

SENTIMENT ANALYSIS

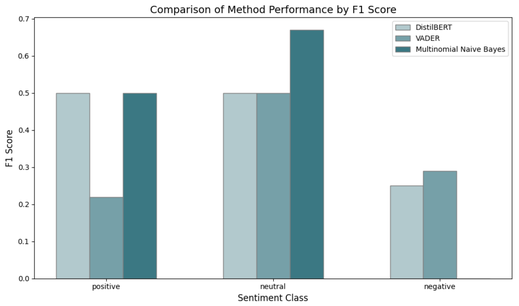
For sentiment analysis, we used the airline tweets dataset used during lab session 3. The dataset consists of tweets from airlines that have been classified as neutral (1515 tweets), positive (1490 tweets), or negative (1750 tweets) sentiment. Each category contains real-world customer interactions with various airlines, capturing a wide range of customer experiences and sentiments toward airline services.The high quality and dimension of the dataset forced us in our decision to perform the comparative analysis, using this as training and validation data, in the task of sentiment analysis. While VADER is designed to handle sentences that have the form of tweets, for DistilBERT we used its own encoding method to generate the needed word embeddings, and for MNB we prepared the data by use of a TF-IDF vectorizer, with the goal of effectively representing the set for training.

TOPIC ANALYSIS

For topic analysis, we combined three datasets, where one is the movie_database downloaded from the NLTK library, and the other two originate from Kaggle. The movie_reviews dataset available through the NLTK (Natural Language Toolkit) library is a popular collection used for text processing and sentiment analysis tasks. It's part of NLTK's corpus of datasets designed to support natural language processing (NLP) research and development. The dataset consists of movie reviews sourced from the Internet Movie Database (IMDb) that are classified as either positive or negative. This binary classification makes the dataset a commonly used benchmark for sentiment analysis models, where the goal is to predict the sentiment of a text as positive or negative based on its content.

LIMITATIONS

While the dataset we used possessed quality annotations and organized structures, further methods were utilized to make the encodings of sentences or tokens more effective. Systems like MNB might have benefitted from further data preprocessing, such as the removal of stopwords and lemmatization. More importantly, the effectiveness of the NERC's training set and its associated results were notably impacted by the pronounced shortage of 'Work of Art' type labels. Similarly, the limited size of the book reviews dataset may have adversely affected the topic classification outcomes. Analysing the test set while performing NERC and Topic Classification also convinced us of how important and effective integration of the two techniques can be in text mining pipelines, particularly the results of NERC feeding into the topic classifier to guide its focus.



RESULT ANALYSIS AND DESCRIPTION

COMPARISON OF MNB VS VADER VS BERT FOR SENTIMENT ANALYSIS

MNB had an accuracy of 0.4. As we observe the results, it appears that none of the sentences was classified as positive. By inspecting the dataset, it appears that it could be because of contemporary terms and references appearing in the test data (e.g. slayed, HOOKED). Furthermore, the model is probably not too robust to grammatical errors and variations in spelling. Still, it shows how even such a traditional, quick-to-train analyser can have a decent performance. VADER had an accuracy of 0.3. As it was noticed that sentence classification in test data appeared quite shifted toward the positive side, we also adjusted the thresholds for positive, neutral and negative to fit performance here. The results are not impressive and maybe while VADER should be equipped to handle colloquial expressions and variations in capitals, it could not work so well at it. In sentences such as 8, context appears to be important to also figure out the emotion.DistilBERT also has an accuracy of 0.4. Its precision appears to be better than its recall and performance appears to be weaker when classifying the negative. It could be that the transformer model is not best equipped to handle short sentences such as those in the test set and that the airline tweets while having the same proportion of labels for the sentiments, have a different way of handling sentences that did not properly train DistilBERT. Comparing results for the three shows a bigger difference in the performance against each specific sentiment target than between the three models itself. This is revealing as it shows that the test data contains indeed tricky information to classify, and that even with powerful model, when the fine tuning is limited they will also suffer from it. In comparison easy and quick MNB proves itself to be not much worse and therefore to the point. Finally, it is interesting to notice how ‘positive’ was the class which appeared hardest to classify for the three, with complete miss for MNB. A possible interpretation could be that a lot of sentences would contain words traditionally negative but that in context might have a different more subtle meaning.

NERC

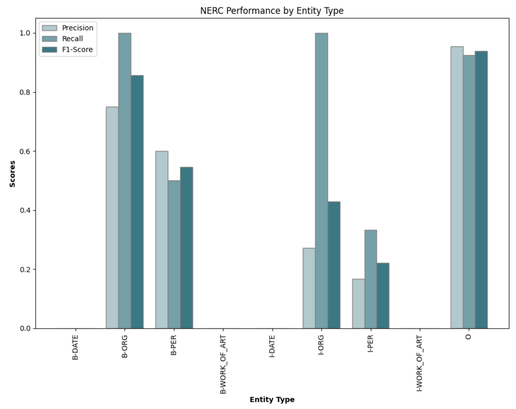
The micro average is heavily influenced by all of the Os (the words that are not part of an NE), it is not surprising that it is relatively high. Looking at the macro averages, which take the average of all of the different classes, we see this going very low. By analysing the shape of the results, we see that this is due in particular to two types of named entities: Work of Art and Date. As these two entities were not in the training test, they are unrecognisable for the CRF and their presence will probably confuse the classifier and make it classify them as ORG or PER, reducing not only the macro average of precision and recall but also the precision in classifying ORG. If we look at the other categories, it appears that the classifier was better at detecting the beginning and insides of the Organization than People, but had a decent performance of both. As we researched to verify whether we could integrate our dataset, it became clear that Work of Art and Date were not traditional Named Entity Labels. Future research, as suggested in the paper [6], could also focus on developing more on other named entity types.

TOPIC ANALYSIS

The accuracy and f1-score of the SVM system are both around 0.6. It appears that the precision is overall higher than recall, both in the macro and weighted averages. The confusion matrix shows us that all the mistakes made by the classifier were mistakenly assigning the topic movie to the other two topics, which have perfect precision. Since the proportion of topics was equalised in the training data, it is indeed remarkable and interesting that the class movie attracted this much attention.

TEST SET & CONCLUSION

Through both manual analysis and the use of the techniques illustrated in this research, we have gathered information on the text contained in the test data and its relevance in showcasing NLP’s potential and difficulties. The test set contained 9 separate sentences, written in colloquial form and appearing similar to tweets. Every word in each sentence possessed NER tagging in the IOB format, the NE types corresponded to person, location, organization, work of art and date. The lack of these last two types in most task-suited datasets underlined the open nature of NER as a task and its need for a somewhat dynamic approach. Every sentence in the test set also had a sentiment (negative, positive, neutral) and topic (movie, sports, book) categorization. The sentiments proved to be somewhat challenging to classify due to their colloquial nature which made them rich in grammatical misspellings and contemporary slang or subtle variations of language and the layered cultural referencing present in even a few words. The topics came to show once again how important domain knowledge can be in a classification task and how integration and compatibility between different data sources are important.



WORK DIVISION

Arshana Ramautar: NERC coding, NERC analysis, general layout design of poster
Federico Signorelli: Sentiment Analysis (MNB & DistilBERT) coding, Sentiment Analysis (MNB & DistilBERT) analysis, organizing text of poster
Rafaëlla van Nee: Sentiment Analysis (VADER) coding, Sentiment Analysis (VADER) analysis, made graphs for poster
Viktoriya Li: Topic Analysis coding, Topic Analysis analysis, general layout design of poster

CRF

	precision	recall	f1-score	support
B-DATE	0.000	0.000	0.000	1
B-ORG	0.750	1.000	0.857	3
B-PER	0.600	0.500	0.545	6
B-WORK_OF_ART	0.000	0.000	0.000	4
I-DATE	0.000	0.000	0.000	1
I-ORG	0.273	1.000	0.429	6
I-PER	0.167	0.333	0.222	3
I-WORK_OF_ART	0.000	0.000	0.000	9
O	0.955	0.925	0.940	160
micro avg	0.839	0.834	0.836	193
macro avg	0.305	0.418	0.333	193
weighted avg	0.833	0.834	0.826	193

VADER

	precision	recall	f1-score	support
negative	0.20	0.25	0.22	4
neutral	1.00	0.33	0.50	3
positive	0.25	0.33	0.29	3
accuracy			0.30	10
macro avg	0.48	0.31	0.34	10
weighted avg	0.45	0.30	0.32	10

DISTILBERT

	precision	recall	f1-score	support
positive	0.50	0.50	0.50	4
neutral	1.00	0.33	0.50	3
negative	0.20	0.33	0.25	3
accuracy			0.40	10
macro avg	0.57	0.39	0.42	10
weighted avg	0.56	0.40	0.42	10

MULTINOMIAL NAIVE BAYES

	precision	recall	f1-score	support
negative	0.50	0.50	0.50	4
neutral	0.67	0.67	0.67	3
positive	0.00	0.00	0.00	3
accuracy			0.40	10
macro avg	0.39	0.39	0.39	10
weighted avg	0.40	0.40	0.40	10

SENTENCE EMBEDDINGS & SVM

	precision	recall	f1-score	support
book	1.00	0.33	0.50	3
movie	0.50	1.00	0.67	4
sports	1.00	0.33	0.50	3
accuracy			0.60	10
macro avg	0.83	0.56	0.56	10
weighted avg	0.80	0.60	0.57	10

REFERENCES:

1. Charles Suttonand Andrew McCallum(2012),"An Introduction to Conditional Random Fields", Foundations and Trends® in Machine Learning: Vol. 4: No. 4, pp 267-373. http://dx.doi.org/10.1561/2200000013
2. Maynard,D.,Bontcheva,K.,&Augenstein,I.(2016).Natural language processing for the Semantic Web. Morgan & Claypool Publishers. https://doi.org/10.2200/S00741ED1V01Y201611WBE015
3. Hutto,C.J.&Gilbert,Eric.(2015).VADER:AParsimoniousRule-basedModel for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014. https://doi.org/10.1609/icwsm.v8i1.14550
4. Alaparthi,S.,Mishra,M.BERT:a sentiment analysis odyssey.JMarketAnal 9, 118–126 (2021). https://doi.org/10.1057/s41270-021-00109-8
5. Kibriya,A.M.,Frank,E.,Pfahringer,B.,Holmes,G.(2004).Multinomial Naive Bayes for Text Categorization Revisited. In: Webb, G.I., Yu, X. (eds) AI 2004: Advances in Artificial Intelligence. AI 2004. Lecture Notes in Computer Science(), vol 3339. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30549-1_43
6. Jain, N., Sierra-Múnera, A., Ehmueller, J., & Krestel, R. (2023). Cultural Heritage and Semantic Web. Semantic Web, 14(2), 239-260. https://doi.org/10.3233/SW-223177

LINK: https://github.com/arshanadevi/text_mining_report/tree/main