UNDERSTANDING TEXT FROM DIFFERENT TYPES OF CONSUMER REVIEWS

Text Mining for AI Final Project Report

B.Yalcin, E.Erel, A.Ilbay, D.Gülal

INTRODUCTION

In this project, NLP techniques such as Named Entity Recognition (NERC), Sentiment Analysis and Topic Analysis are applied to a set of restaurant, book, movie reviews with the goal of gaining a deeper understanding of customer opinions and preferences. In this way, we aim to detect sentiments and topics of the product reviews.

METHODOLOGY

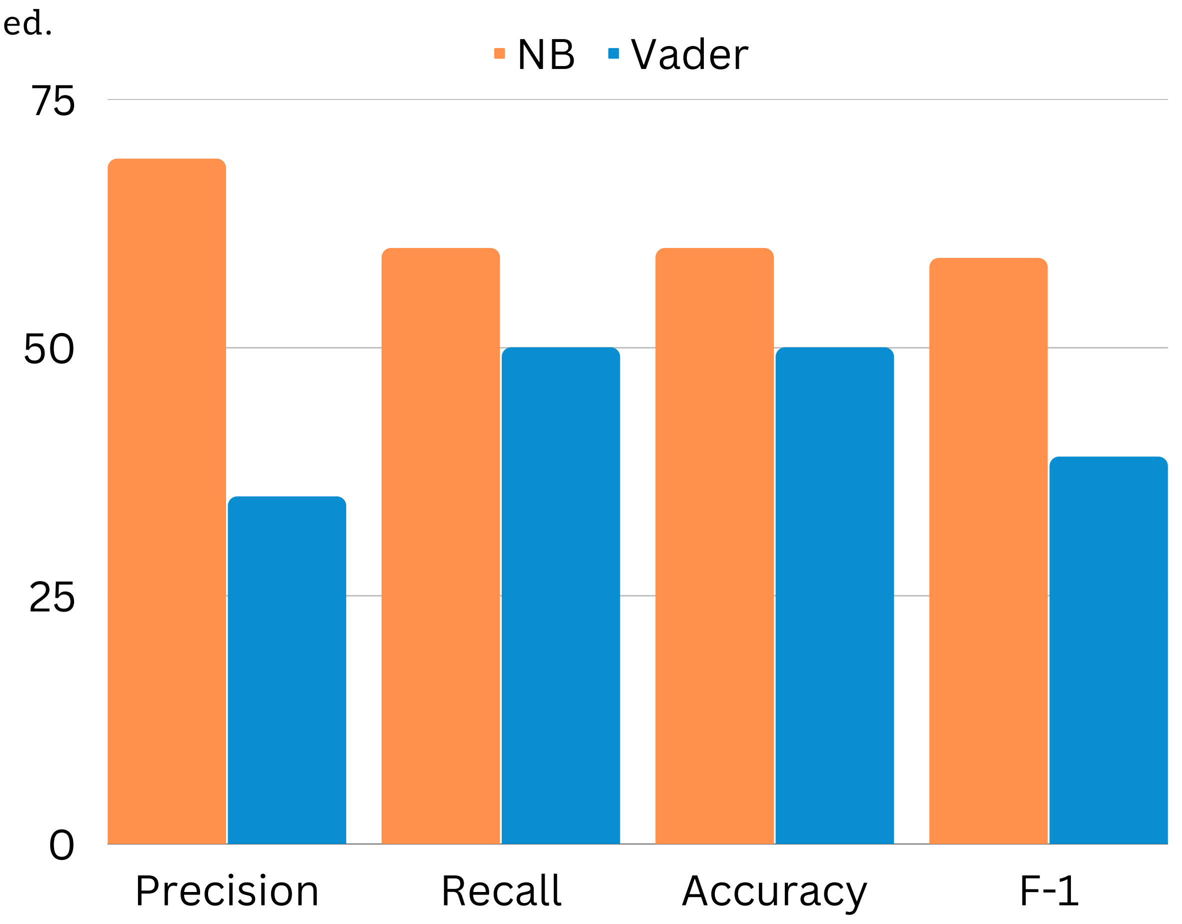
Data for Sentiment analysis and Topic analysis is collected from book reviews, movie reviews, and restaurant reviews datasets, merged into one. The dataset consists of 4 columns: sentence id, text, sentiment, and topic. For NERC analysis Annotated Corpus dataset is used.
Feature extraction with:

- NERC analysis: dataset trained with SVM. The motivation for using SVM with Annotated Corpus Dataset is because Support Vector Machines are commonly used in NLP tasks for classification and regression.
- Sentiment analysis
  - VADER: rule-based sentiment analysis system using a lexicon of word-sentiment score pairs. The main motivation for selecting this approach is to make a clear distinction between the performances of rule-based and machine-learning systems during sentiment analysis.
  - Naive Bayes: Statistical classification method based on Bayes theorem. The motivation was to use a machine-learning algorithm with simplicity, speed and effectiveness.
- Topic analysis
  - We used LDA to test how well it could predict our labels and used regression algorithms because they are simple, efficient and easy to use.
  - Gensim, LDA, Tf-idf, SVM

RESULTS

The sentiment analysis done with the Naive Bayes classifier represents an accuracy of 0.6. For negative class, the precision is 1.0, which means that all predicted negative instances are correctly classified. For neutral class, the precision is 0.5. For positive class, the precision is 0.6 that is 60% of the predicted positive instances are correct.For negative class, only 33% of the actual negative instances are correctly predicted. For positive class, the recall is 0.75, which means that 75% of the actual positive instances are correctly predicted.For negative class, the F1-score is 0.5. For neutral class, the F1-score is 0.571. For positive class, the F1-score is 0.667.The macro-average scores are shown in the figure [1].
Overall, the performance of the Naive Bayes classifier is not great as the F1-scores for all classes are relatively low, and the accuracy is only 60%. There is room for improvement in terms of selecting better algorithms such as Deep Learning approaches, improving the quality and diversity of the data, or fine-tuning the model's hyperparameters.

The sentiment analysis conducted using the VADER system yields overall 50% accuracy. For the positive category, system managed to identify all positive sentences with recall score of 1, however its precision is 0.5 meaning that it classifies that many sentences from other categories as positive. Recall and precision scores of 0.33 and 0.5 respectively for negative class are incompetent relative to the positive category. Evaluation metrics about the neutral category could be considered biased as the system failed to identify any sentences for this category. This presents ambiguity in terms of precision and F-1 score as well as macro-level scores. To improve the credibility of the experiment, test support could be increased as this would increase the chance of getting neutrally identified sentences, as well as the support of each category could be equalized.



[1]

In topic analysis, we trained and compared a few different models. Sklearn's LDA algorithm managed to match the topics with the labels, and performed with f1 scores higher than 0.95. On the contrary, gensim's LDA could not seperate the topics and had too many overlaps between them. Significant portion of the common words in the training data were not topic specific, for example "good", and this resulted in LDA algorithms putting them in all topics, for both gensim and sklearn. We used the default parameters when training and we conclude sklearn's defaults were fit for our training data, whereas gensim's was not. Then we trained supervised models with TF-IDF using Naive-Bayes and SVM, which performed very similarly, both with f1 scores higher than 0.98. Using supervised learning was shown to be superior to LDA in this context. Presence of labels on top of TF-IDF reduced the impacts of non topic specific words, and this explains why these methods were more successful than LDA.

In NERC task, the results show that for most of the named entity classes precision, recall, and f1 scores are 0, meaning that the system failed to identify them. As the support is very low, this indicates that there are very few instances of these entities in the dataset. This failure could be because of existence of the similar entities in the dataset (lack of divereseness). The system successfully identified B-ORG and B-PER tags as they both had a precision of 1.0. Overall accuracy in the NERC model is 0.96, which is a good score. It is because O tokens were mostly common in the dataset, and its precision, recall and f1 scores are very high.

CONCLUSIONS

Naïve Bayes classifier and VADER system are used for sentiment analysis. Main difference regarding the dynamics of between these two is that the former requires a training stage with various sentence-label pairs whereas the latter is a rule-based system making use of a lexicon of words and their sentiments scores. In our results, it is observed that Naïve Bayes Classifier outperformed VADER under each evaluation metric.[1] The fact that the sentence topics in the training set used by the NB Classifier fits with that of the test set might have played a role in this overperformance since NB retains context to a degree. On top of that, we observe perfect recall scores for both systems under positive category however, this is not the case for the other categories. This raises questions regarding the lack of quality of the test data.

We propose two improvements regarding the VADER system. The system makes use of a lexicon that disregards homophony and context. The addition of context identification feature along with homophone words to the lexicon would improve precision & recall metrics and might potentially help irony detection.Also, for the sentiment analysis using Naive Bayes method there is room for improvement in terms of using more advanced Naive Bayes variants, improving the quality and diversity of the data, or fine-tuning the model's hyperparameters.

In topic analysis, our models have high accuracy but they could be improved by removing non topic-specific words or using only some parts of speeches that exclude such words. Changing default parameters also could improve the system.

NERC model identified correctly most of the non-entity tokens in the dataset. However, the model performed poorly on named entity classes except for B-ORG and B-PER. To improve the results, a larger and diverse dataset could be used.

DIVISION OF WORK

1) coding, 2) analysis, 3) poster preparation
Begüm: 1) NERC and dataset preparation, 2) NERC analysis, 3) Contributed equally to the poster
Ata: 1) Topic analysis, 2)Topic analysis, 3) Contributed equally to the poster
Emirhan: 1) Sentiment analysis with VADER, 2)Sentiment analysis, 3) Contributed equally to the poster
Deniz:1) Sentiment analysis with Naive Bayes, 2)Sentiment analysis, 3) Contributed equally to the poster

Code zip folder link: https://drive.google.com/file/d/1llcuYYSXyz-_KKGcNAVKT_v9RI1MPbP6/view?usp=sharing