NLP+CLASSIFICATION PROJECT

The aim of this project is to analyze a collection about comments on purchases about drinkable and edible products.

The collection is stored in the file “products.csv” made up of the following the fields:

- Id: number
- ProductId: String
- UserID: String
- ProfileName: String
- HelpfulnessNumerator: int
- HelpfulnessDenominator: int
- Score: int in the range [1,5]
- Time: int
- Summary: String
- Text: String

Using the field “Summary+Text”, it is necessary to follow the next steps:

1. **Preprocessing**

Mandatory preprocessing steps

- Remove useless characters: ! "_ $% & / ( ) = _ ˆ* ¡@
- Remove all capital letters
- Lemmatize all terms

Optional preprocessing steps:

- Remove contractions: don’t → do not[1]
- Remove repeated words: great great show
- Remove or replace emoticons
- Correct wrong words: spelling corrector (if it takes long, do not use it)

2. **Vectorization**

Vectorize every opinion by following different configurations:

- TFIDF
- TFIDF + N-grams
- TFIDF + N-grams + POS tagging

[1] https://englishstudypage.com/grammar/list-of-contractions-in-english/

- TFIDF + N-grams + POS tagging + other features: number of words, number of sentences, etc.

3. Feature selection

Select the best features using the selectKBest and removing 70% of the features used per configuration.

4. Classification algorithm

Using the selected best features, use 2 classification algorithms to classify the opinions according to the field “score” (1, 2, 3, 4 or 5 stars). To do so, 70% of the dataset will be used for training and 30% for testing. Tune the different parameters, if possible, by a cross validation.

Write a report describing the process followed (use a latex template):

- Problem description
- Methods and materials: Classification algorithm and dataset for testing
- Experiments and results: containing the following evaluations metrics: precision, recall, F-measure and confusion matrices
- Conclusions