

AnTeDe Lab 3

1 Notebook B

The goal of this first Notebook is to create a baseline against which to evaluate the subsequent Sentiment Analysis Models. This baseline approach consists of simply counting the positive and negative words in a given document. If more positive words are present than negative words the document is classified as positive. The lists of positive and negative words are taken from "<https://ptrckpry.com>".

This approach is very simple and has some clear drawbacks. By using an off-the-shelf list of words the model only knows words that are in the static list (Not learned from the corpus). Further, some words may have a more positive or negative association depending on the domain. For example, the word "addicting" is in the list of negative words, but in the context of films and tv shows an "addicting movie" could be a very positive review. Finally, the model ignores adverbs like "not", "nearly", "almost" and so on. These words can shift the sentiment of a document toward either positive or negative.

The simplicity of this model makes it a good baseline with an F1 score of 0.71.

2 Notebook C

In this Notebook, the goal is to improve over the established baseline from the previous section.

2.1 Multinomial Naive Bayes

The first improvement over the baseline is done with the Multinomial Naive Bayes model. Two models are trained and evaluated against each other. The two models differ in one hyperparameter of the CountVectorizer.

2.1.1 CountVectorizer

The "binary" hyperparameter changes the resulting vectors representing the document.

$$\begin{bmatrix} 3 \\ \vdots \\ 0 \\ 2 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ \vdots \\ 0 \\ 1 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

If a particular word occurs once or more in the document the vectors will have a one in the corresponding location if "binary = True" else it will contain the number of occurrences. Only allowing one or zero, very common words do not skew the vectors, making the embeddings more distinguishable. This effect can also be observed in the classification result, by simply setting "binary = True" the classifier performs about 3% better and 10% better than the baseline (Test F1).

2.2 LogisticRegression

The Second model tested is the LogisticRegression model. Like in the above example a cross-validation and the Vectorizer is set to "binary = True". The best model is evaluated and it performs 3% better than the Multinomial Naive Bayes model (Test F1 0.85).

2.3 Final Assessment

The final experiment removes all tokens that are not contained in the two lists from Notebook B before feeding the filtered documents through the Vectorizer. This new classifier underperforms compared to the previous LogisticRegression classifier and performs about the same as the Multinomial Naive Bayes with the "binary = True". That this classifier performs worse than the LogisticRegression classifiers is expected since we lose information by removing all these tokens from the documents. That it still outperforms the baseline model is also expected, since it now can learn the positive and negative association of tokens from the training labels, thus "addicting" may be associated with a positive review, not a negative one.

All these approaches still ignore position and relationships between the tokens thus modifying adverbs can be lost in the BagOfWords.