HACETTEPE ÜNİVERSİTESİ

HACETTEPE UNIVERSITY DEPARTMENT OF COMPUTER ENGINEERING

BBM409 LAB

ASSIGNMENT 2 REPORT

Name: Sanberk Satıcıoğlu - 21228654

Tfidf is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf-idf value increases proportionally to the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

For Multinomial Naive Bayes, TfIdf values were used they are extracted by "TfidfVectorizer"

PREDICTION ANALYSIS

Feature	Accuracy
Unigram w/ MultinomialNB	0,573
Bigram w/ MultinomialNB	0.452
Unigram and Bigram w/ MultinomialNB	0.564

As we can see, bigram and unigram and bigram performed worse than just unigram. There may be different causes for this. For example, adding extra features may lead to overfitting. Tf-idf is unlikely to alleviate this, as longer n-grams will be rarer, leading to higher idf values.