Feature Engineering

Anas Salamah

My username in Kaggle is Anas My approach with this problem of feature collection was to get the basic information first and test and record my results. After that, add more features that are not so obvious but can collect like the length of the example. The interesting thing was that adding bigram information did not help the accuracy. Bellow is a graph showing my attempts:

Accuracy	Features
0.434184	words in one big string
0.722087	W: word_x
0.729004	L: Length of example, W:
0.737506	morphy_stem(), L: Length of example, W:
0.789764	stopwords, morphy_stem(), L: Length of example, W:
0.7828	*POS, stopwords, morphy_stem(), L: Length of example, W:
0.789933	stopwords, morphy_stem(), L: Length of example, W:
0.783423	BIGRAM: bigram_x, stopwords, morphy_stem(), L: Length of example, W:
0.792022	CNTword_x, stopwords, morphy_stem(), L: Length of example, W:
0.789284	CNTword_x, stopwords, morphy_stem(), L: Length of example, W

W: word_x: all the words starting with a “W:
L: Length of example
Morphy_stem(): a function that stems the word
Stopwords: removed stopwords
BIGRAM: added bigram information with a the mentioned tag
CNTword_x: all words counts information