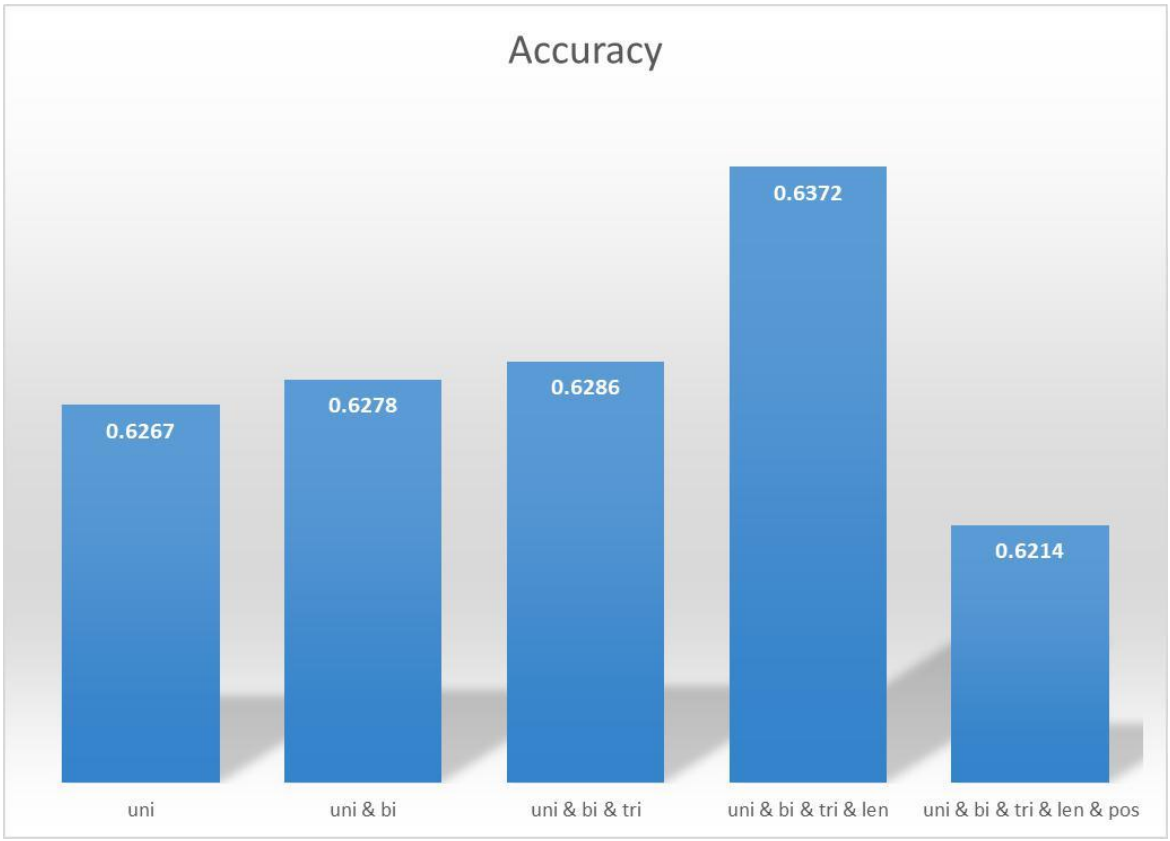
Naïve Bayes

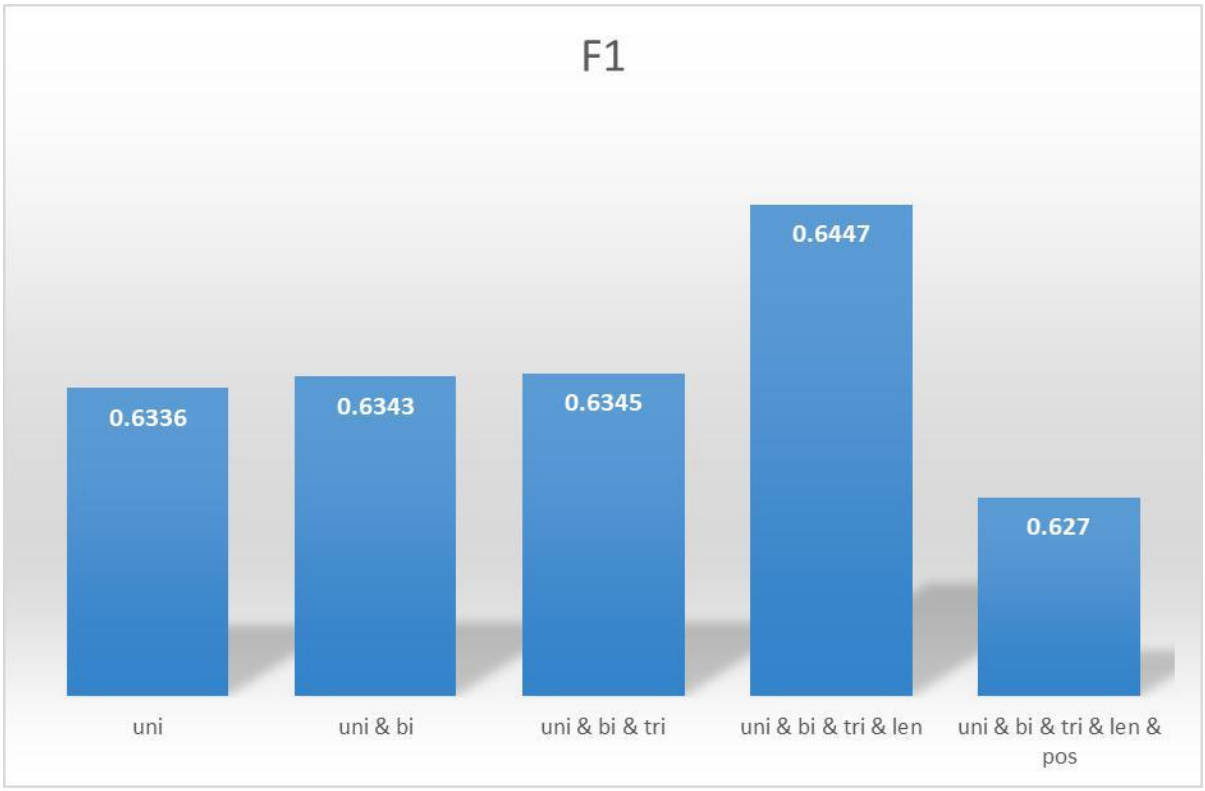
bag of words classifier leaves us with 62% accuracy, I think the main reason for accuracy loss is ignoring words sequence. in this classifier, we don't care about the sequence of words and that leads to the wrong answer in some cases. for example, a character may tend to make a mistake on a grammar rule or uses a word at the beginning of sentences frequently ... etc. this classifier ignores these properties of characters. these kinds of problems can be solved in sequence models which care about words order.

Maximum Entropy

in a series, characters usually have a specific set of words. for example, joey uses the word "Yeah" frequently and chandler uses the word "Monica" instead. So, using **unigrams** as a feature seems to be a good choice. similarly using **bigrams** and **trigrams** is reasonable. I also used **length of the sentence** as a feature because chandler sentences tend to be wordier. I tried using all POS tags as features but that turned out to be a wrong move.

Here are accuracy and F1 charts for using different features:




F1
0.6336
0.6343
0.6345
0.6447
0.627
uni
uni & bi
uni & bi & tri
uni & bi & tri & len
uni & bi & tri & len & pos