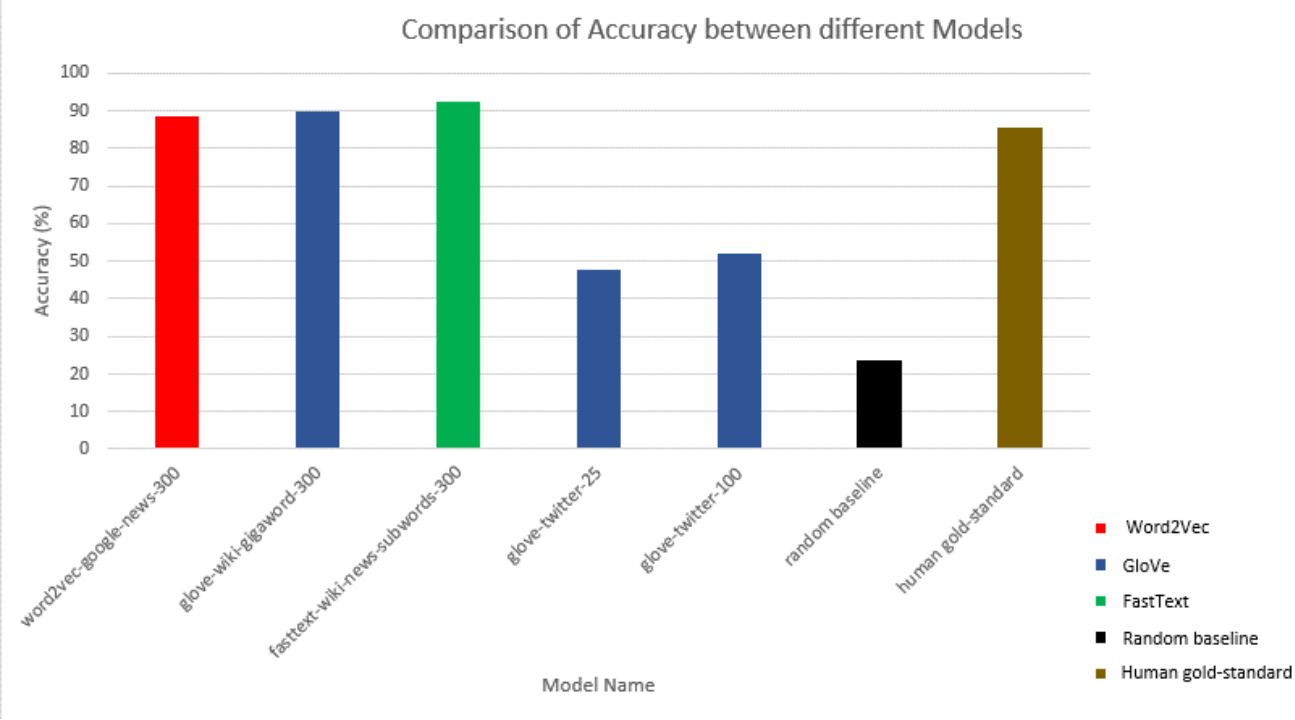

Analysis:

The graph above shows the relative accuracy of the different models that we used as well as that of a random baseline and the human gold-standard. The models with a higher embedding size have higher accuracies as a result. This is because a higher number for the embedding dimension size means that words can be grouped into more specific categories, making it easier to understand the context in which they are used. FastText had the highest overall accuracy, by a slight margin. This is because it is a slightly improved version of Word2Vec that decomposes each word into a set of n-grams.

The models that used twitter data show significantly lower values for accuracy, we believe that this is due to the lower embedding size but can also be due to the fact that certain words are used differently on social media platforms so the context in which the words are used can cause confusion. We then compared all the models to a random baseline that was generated by making random guesses using the options (1 in 4 chances of guessing correctly). As expected, all the models scored significantly better than the random baseline since they are using real data to make their choices. Other than the glove models, all the other models scored higher than the human gold-standard set by the students of COMP 472. This is impressive as it means that the models ae better than the average engineering student at identifying synonyms.