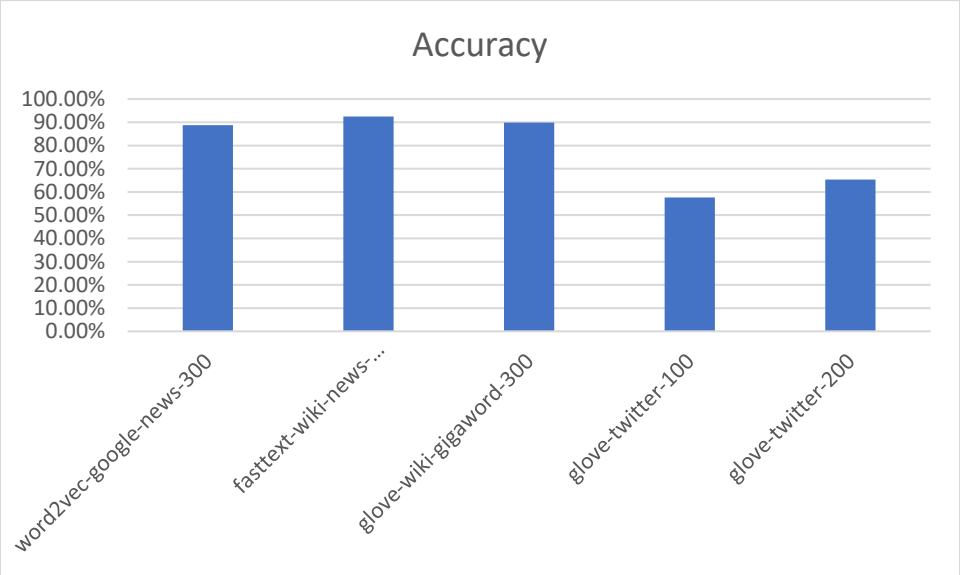
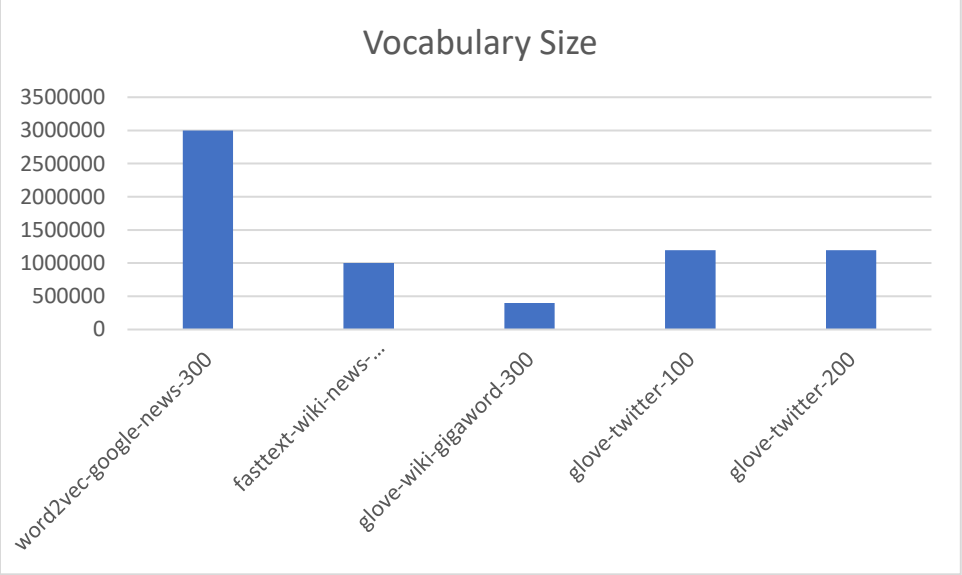
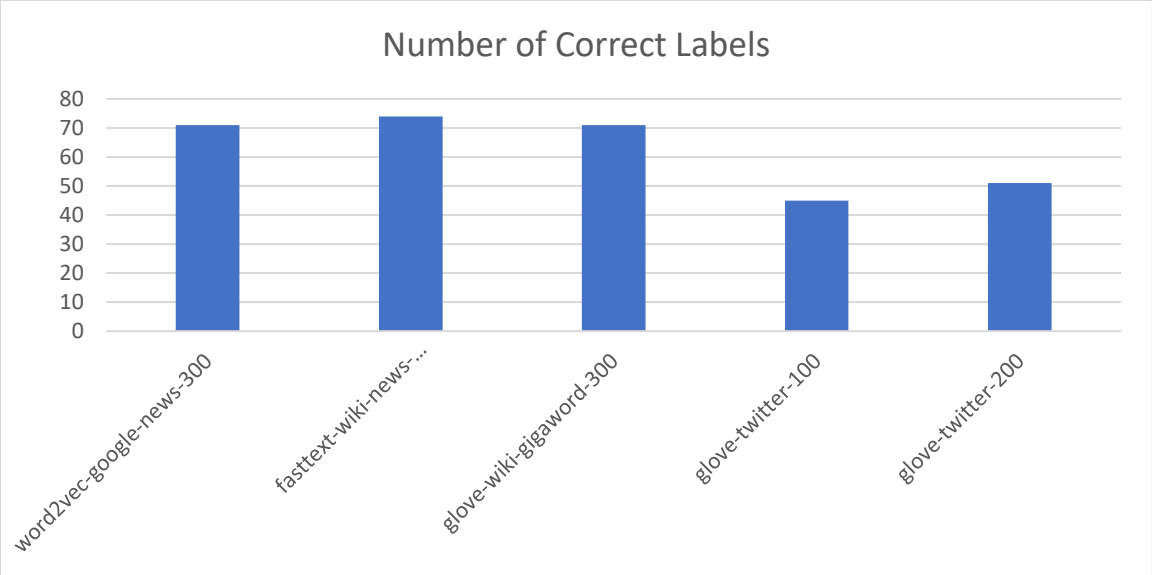
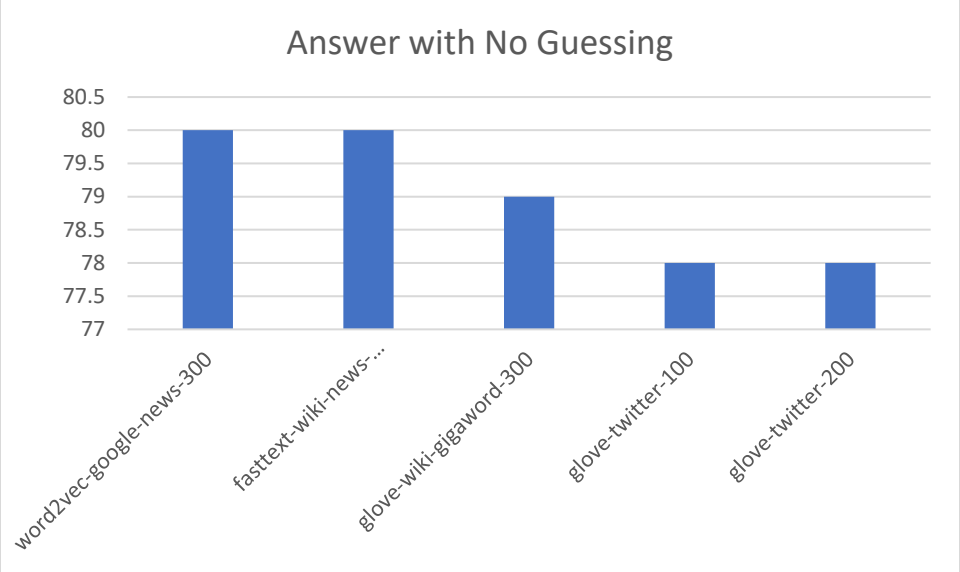
4 additional models were used on top of the Google model: fasttext-wiki-news-subwords-300, glove-wiki-gigaword-300, glove-twitter-100, and glove-twitter-200.

Model Name	Vocabulary Size	Number of Correct Labels	Answer with No Guessing	Accuracy
word2vec-google-news-300	3000000	71	80	88.75%
fasttext-wiki-news-subwords-300	999999	74	80	92.50%
glove-wiki-gigaword-300	400000	71	79	89.87%
glove-twitter-100	1193514	45	78	57.69%
glove-twitter-200	1193514	51	78	65.38%









According to the Accuracy graph, there is no overall relation between the vocabulary size of a model and its accuracy, as our 2nd smallest model had the best accuracy, the smallest model had the second-best accuracy, and the biggest model had the 3rd best accuracy.

The embedding size, however, does matter, as all the models with an embedding size of 300 have an accuracy in the high 80s to low 90s, while the model with an embedding size of 200 has an accuracy close to 65.38%. The model with the smallest embedding size of 100 has the lowest accuracy of 57.69%. It thus follows that the smallest embedding size has the smallest number of correct guesses.

Between the 3 models with an embedding size of 300, there is no relationship between the size of the vocabulary and the accuracy.

Between the 2 models with the same corpus (glove-twitter), but different embedding sizes, the relationship between embedding and accuracy seems to be confirmed, as the model with an embedding

of size 100 has an accuracy that is roughly 8% lower than one for the model with an embedding size of 200.

Interestingly, the number of guesses changes more with the model/corpus than with the embedding size, as all models with an embedding size of 300 did not guess 79 to 80 times, while the two 100 and 200 models based on the same corpus did not guess 78 times. It must be noted that a guess here is a TRUE guess, meaning that neither the question-word nor the possible answer words were considered in the model.