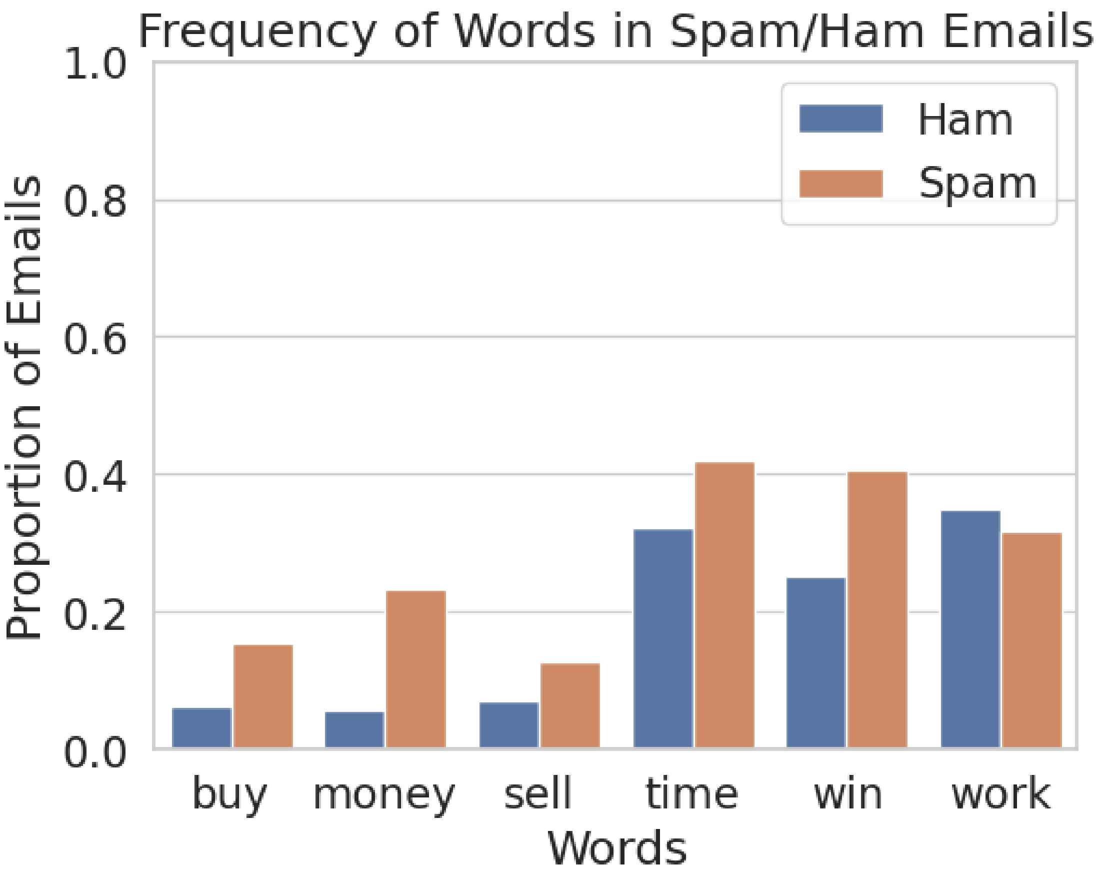
Discuss one thing you notice that is different between the two emails that might relate to the identification of spam.

The first difference I noticed is the url. For instance, the url for ham emails most often ends with “.net”, “.org”, or “.edu”. However, the spam emails often contain only “.com”. The second difference is the name and signature of the author. For instance, in the ham email the author of the email include their names, but most often in spam emails this is not practiced.

0.0.1 Question 3

Create a bar chart like the one above comparing the proportion of spam and ham emails containing certain words. Choose a set of words that are different from the ones above, but also have different proportions for the two classes. Make sure to only consider emails from `train`.

```
In [12]: train = train.reset_index(drop=True) # We must do this in order to preserve the ordering of em
         spam_ham = train[['spam']].replace({'spam': {0: 'Ham', 1: 'Spam'}})['spam']
         words = ['buy', 'sell', 'work', 'money', 'win', 'time']
         words_in_email = words_in_texts(words, train['email'])
         df = pd.DataFrame(data = words_in_email, columns = words)
         df['spam/ham'] = spam_ham
         df_melted = df.melt('spam/ham').groupby(['spam/ham', 'variable']).agg(np.mean).reset_index()
         sns.barplot(data = df_melted, x = 'variable', y = 'value', hue = 'spam/ham')
         plt.legend(title = '')
         plt.ylim([0,1])
         plt.xlabel('Words')
         plt.ylabel('Proportion of Emails')
         plt.title('Frequency of Words in Spam/Ham Emails')
         plt.show()
```



0.0.2 Question 6c

Comment on the results from 6a and 6b. For **each** of FP, FN, accuracy, and recall, briefly explain why we see the result that we do.

6a: FP is zero because our model never predicts positive, so we have zero FP. Taking into account FP, FN is the number of spam emails because our model falsely predicts spam emails as ham. 6b: recall is zero because TP is zero as our model predicts all emails as ham. Accuracy is the correct ratio of the ham emails that the model label correctly over all emails (spam + ham).

0.0.3 Question 6e

Are there more false positives or false negatives when using the logistic regression classifier from Question 5?

FN = 1699 and FP = 122. FN is way more than FP.

0.0.4 Question 6f

1. Our logistic regression classifier got 75.76% prediction accuracy (number of correct predictions / total). How does this compare with predicting 0 for every email?
2. Given the word features we gave you above, name one reason this classifier is performing poorly. Hint: Think about how prevalent these words are in the email set.
3. Which of these two classifiers would you prefer for a spam filter and why? Describe your reasoning and relate it to at least one of the evaluation metrics you have computed so far.

1. The accuracy rate of the logisitic regression classifier is greater than the accuracy rate of the zero predictor (75.76% > 74.47%).
2. The words we use to predict spam and ham emails do not help us distinguash between these two class. This is why our model performs poorly meaning that the accuracy rate is low.
3. Based on accuracy, the logistic regression model perfroms better than zero predictor, so I choose the logistic regression model.