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

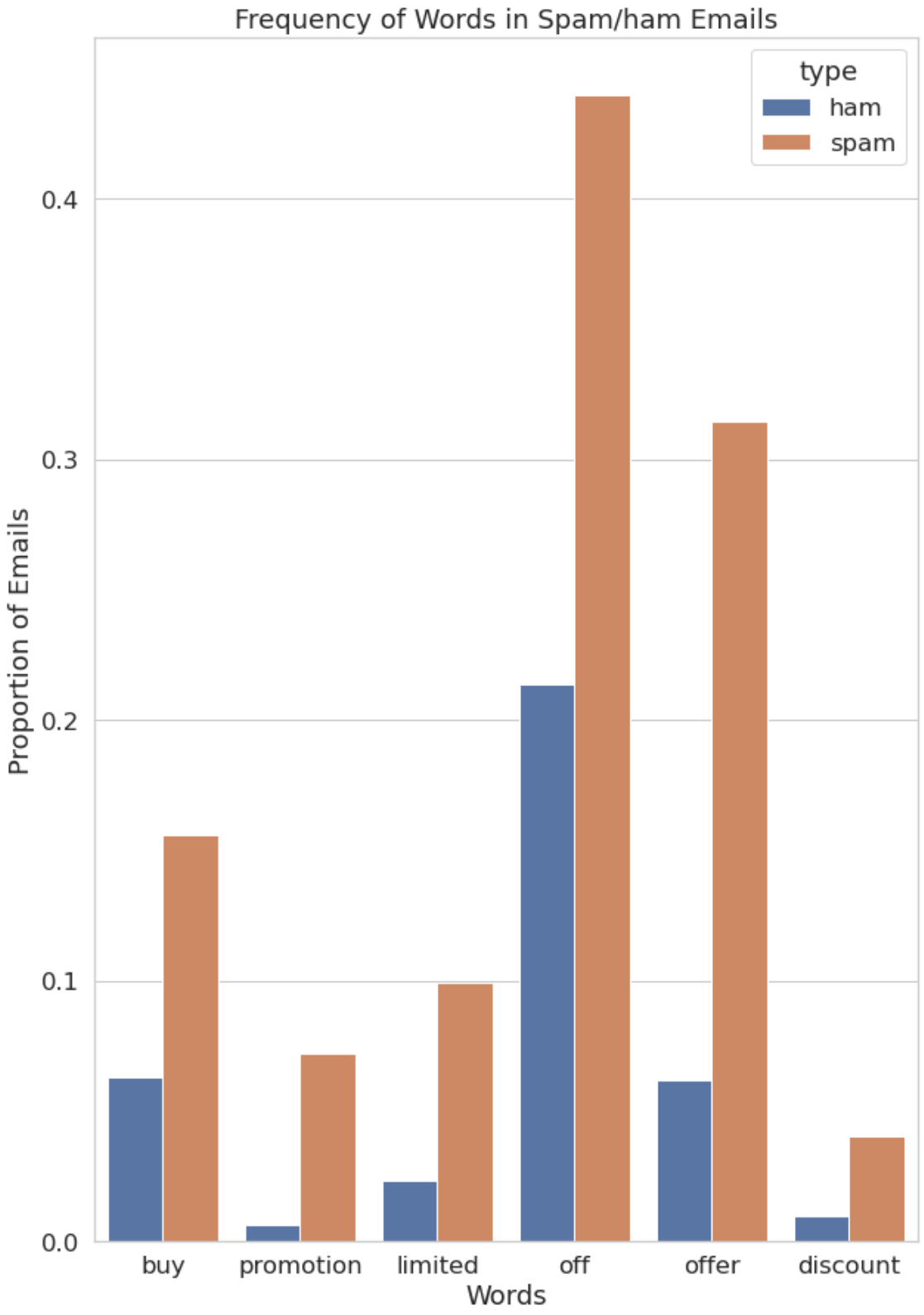
The spam email is in html format, while the ham is just text.

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 [41]:
```
train = train.reset_index(drop=True) # We must do this in order to preserve the ordering of em
words = ['buy', 'promotion', 'limited', 'off', 'offer', 'discount']
word_in_digit = words_in_texts(words, train['email'])
word_df = pd.DataFrame(word_in_digit, columns = ['buy', 'promotion', 'limited', 'off', 'offer'
word_df['type'] = train['spam'].replace({1: 'spam', 0:'ham'})
melt_df = word_df.melt('type')

plt.figure(figsize = (10,15))
sns.barplot(x = 'variable', y = 'value', data = melt_df, hue = 'type', ci = None)
plt.xlabel("Words")
plt.ylabel("Proportion of Emails")
plt.title("Frequency of Words in Spam/ham Emails");
```


Frequency of Words in Spam/ham Emails
type
ham
spam
Proportion of Emails
0.0
0.1
0.2
0.3
0.4
buy
promotion
limited
off
offer
discount
Words

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.

Since this predictor only gave us negative result:

the FP is always zero since there is no positive at all.

the FN is the number of positive samples since all of them are wrongly predicted as negative.

accuracy is the number of negative sample / total samples since we only correctly predicted negative samples.

recall is 0 since TP is 0.

0.0.3 Question 6e

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

There are more false positive and less false negative when using the logistic regression classifier from Question 5

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. Sightly better, but poor in generally speaking
2. These words are prevalent, but with no obvious bias indicating spam/ham
3. I prefer the logistic regression model. First it is slightly more accurate. Secondly,