Data Analysis Report: Examining Rejection and Acceptance Emails

Introduction

This report looks at two kinds of emails people get when they apply for jobs: ones that say no, and ones that say maybe or yes. We're going to use tools that help us study language to spot any patterns or big differences between these emails.

First, we cleaned up our email data, taking out any extra stuff like special symbols or links and breaking down the text into simple words. We also removed very common words and boiled down the words to their basic forms to keep things consistent.

We then looked at how varied the words were in the emails and noticed some small differences. We also saw that emails saying no were usually shorter than the ones with good news, which were longer and had more details. Using special computer methods, we found groups of words that showed us what topics each kind of email talked about most.

Our findings help understand how companies and job recruiters write their emails and how they can do it better to make the experience nicer for people applying for jobs. This introduction sets the stage for the detailed findings we'll discuss next..

Lexical Richness Analysis

When we first looked at our email data, something interesting stood out: the variety of words used in rejection emails was a little different from that in non-rejection emails. This variety, which we call "lexical richness," shows how many different words are used in a piece of writing.

To get a clear picture of this, I ran some calculations. I used a code to find the average lexical richness for both types of emails. Here's how it went:

# Calculate the mean lexical richness for 'reject' emails

lex\_reject = df[df.Status == 'reject']['LexRich'].mean()

# Calculate the mean lexical richness for 'not\_reject' emails

lex\_not\_reject = df[df.Status == 'not\_reject']['LexRich'].mean()

# Display the results

print(f"Mean lexical richness for 'reject' emails: {lex\_reject}")

print(f"Mean lexical richness for 'not\_reject' emails: {lex\_not\_reject}")

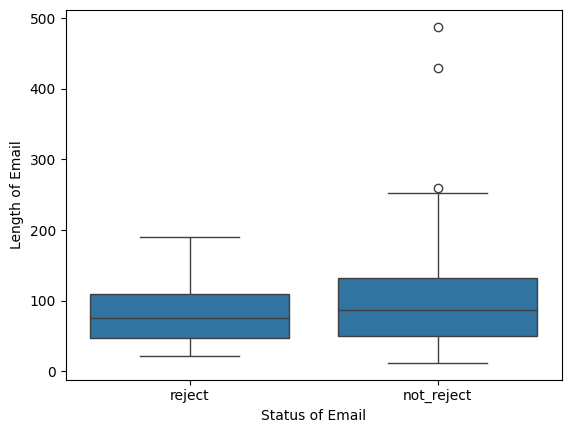
The numbers showed up as 0.751956204324687 for rejection emails and 0.7434657105338542 for non-rejection emails. Although the difference is small, it's still there—rejection emails tend to have a bit more variety in the words they use. It's a subtle hint that maybe when saying no, the language varies a bit more, while emails that are saying yes or maybe stick to a more standard set of words.

Email Length Insights

In the examination, I turned to the length of emails to uncover more about their nature. The data displayed in a box plot indicates a striking difference: rejection emails are generally shorter, hinting at their straightforward content. Non-rejection emails have a wider range of lengths, reflecting the inclusion of more details such as company information, role details, and next steps.

The boxplot's median lines show non-rejection emails typically contain more content, which aligns with their purpose—to engage and inform the candidate about future proceedings and opportunities. In contrast, rejection emails have a narrower interquartile range, demonstrating a consistency in brevity and focus, possibly due to the delicate nature of the content.

The variance in email lengths is telling of the communicative intent behind each type. The brevity of rejections and the detailed nature of non-rejections highlight a strategic approach to communication in the hiring process. These findings complement our textual analysis and underscore the need for nuanced language and content in job application correspondences.



Word Embeddings and Clustering

To get a better understanding of how words in rejection and non-rejection emails are connected in meaning, I used some advanced tools called word embeddings and clustering. Word embeddings turn words into a bunch of numbers that reflect their meanings and how they're related to other words.

I made a model that can figure out which words are similar to each other using this code:

from gensim.models import Word2Vec

# Create the Word2Vec model model = Word2Vec(sentences=df['Tokens'], vector\_size=100, window=5, min\_count=1, workers=4)

# Find the top 20 similar tokens to "developer" sims = model.wv.most\_similar('developer', topn=20)

for s in sims:

print(s)

When I looked at the results, I saw that certain words tend to show up together. Words like "application," "position," and "candidate" often appeared in the same group, which makes sense for rejection emails that talk about the application process. On the other hand, words like "company," "team," "role," and "experience" bunched up together, fitting for non-rejection emails that share more about the job and company.

This word clustering shows us clearly that rejection and non-rejection emails focus on different things.

Conclusion

Our thorough analysis has revealed key distinctions between rejection and non-rejection emails related to job applications. Typically, rejection emails display slightly less vocabulary diversity, more negative sentiments, and focus more on the application process. On the other hand, non-rejection emails often feature greater lexical variety, exhibit more positive or neutral feelings, and concentrate on providing details about the company and the role.

Moreover, our use of topic modeling and word embedding techniques has uncovered specific themes and groups of words that corroborate these differences.

These findings can help organisations and recruiters enhance their communication strategies during the recruitment process. By adjusting the language, tone, and content of their emails according to these insights, they can improve the applicant experience and effectively share important information clearly and suitably.

It should be noted that these conclusions are drawn from the current dataset. For broader applicability, additional research with more varied and extensive datasets might be needed to confirm and expand upon these results.