A Comparative Analysis of Deceptive Text Patterns Across Different Domains

Lucas Dufour | lud5@hi.is

March 24, 2023

Abstract:

Deceptive text is a pervasive issue in various forms of communication, ranging from phishing emails to fake news. This study conducts a comparative analysis of five different datasets containing deceptive and genuine text samples. The datasets include phishing emails, fake news, job scams, political statements, and product reviews. We employ various analytical techniques, such as text length distribution, word frequency, and t-SNE visualizations, to uncover patterns and trends within and across these datasets. Our findings provide valuable insights into the nature of deceptive text across different domains and can help inform future research and the development of more effective detection methods.

Introduction

Deception has become a significant problem in various areas of communication, including phishing emails, fake news, job scams, political statements, and product reviews. Identifying deceptive patterns across these domains can help in understanding their commonalities and differences, and subsequently, in developing more effective detection methods. In this study, we analyze five datasets containing deceptive and genuine text samples to uncover trends and patterns within and across these domains.

Materials and Methods

1. Datasets

We used five datasets containing samples from the following domains: - Phishing Emails - Fake News - Job Scams - Political Statements - Amazon Product Reviews

Each dataset contains text samples labeled as deceptive or genuine.

2. Data Analysis

We conducted the following analyses on each dataset: - Number of samples and distribution of deceptive labels - Text length distribution - Word frequency analysis using word clouds - t-SNE visualization for dimensionality reduction

3. Comparison Across Datasets

We compared the results of the analyses across the five datasets to identify commonalities and differences in deceptive text patterns.

Results

3.1 [Dataset 1: Phishing Emails]

- Number of samples and distribution of deceptive labels
- Text length distribution
- Word frequency analysis
- t-SNE visualization

3.2 [Dataset 2: Fake News]

- Number of samples and distribution of deceptive labels
- Text length distribution
- Word frequency analysis
- t-SNE visualization

3.3 [Dataset 3: Job Scams]

- Number of samples and distribution of deceptive labels
- Text length distribution
- Word frequency analysis
- t-SNE visualization

3.4 [Dataset 4: Political Statements]

- Number of samples and distribution of deceptive labels
- Text length distribution
- Word frequency analysis
- t-SNE visualization

3.5 [Dataset 5: Amazon Product Reviews]

- Number of samples and distribution of deceptive labels
- Text length distribution
- Word frequency analysis
- t-SNE visualization

3.6 Comparison Across Datasets

Discussion

Discuss the findings from the results section, elaborating on any notable patterns or trends observed across the datasets. Consider the implications of these findings on future research and the development of deception detection methods.

Conclusion

Summarize the key findings of the study and their significance in understanding deceptive text patterns across different domains. Highlight the potential implications for developing more effective detection methods and suggest directions for future research.

References

List the references used throughout the paper.