|  |
| --- |
| **Instructions for \*ACL Proceedings** |
|  |
| **LAKHANI, Harsh Sunil**  **HKUST** |
|  |
|  |
|  |
|  |

Abstract

In this project, we address the problem of spam detection in comments, messages, and emails, a critical issue emphasized by Elon Musk (Add reference) and prevalent across various social media platforms. We curated a dataset of 6000 labeled instances by combining multiple sources of spam messages and emails. We employed supervised learning using state-of-the-art text classification models, including BERT, LSTM, and RoBERTa, to classify messages as spam or not. Subsequently, we applied unsupervised clustering techniques, such as Hierarchical clustering, K-means, and DBScan, to group similar spam messages. We leveraged a generative model to assign labels to each cluster, identifying the type of scam. Our aim and approach demonstrate the potential to detect spam and categorize it into different scam types, facilitating better content moderation on various online platforms.

Introduction

Spam messages, comments, and emails have become a pervasive issue in the digital era, polluting online communication channels and causing inconvenience for users. Beyond being a mere nuisance, spam often serves as a medium for scams, including identity theft, malware, financial fraud, and other potentially detrimental effects, thus posing significant risks to unsuspecting individuals. As these unwanted communications continue to evolve, the need for robust and efficient spam detection systems has become more pressing.

This report presents a novel approach to spam detection and classification by combining supervised learning techniques, unsupervised clustering, and generative models. We first use state-of-the-art text classification models, such as BERT, LSTM, and RoBERTa, to perform supervised learning on a curated dataset of 6000 instances. Afterward, we employ unsupervised clustering algorithms, including Hierarchical clustering, K-means, and DBScan, to group similar spam messages. Finally, we utilize a generative model to assign labels to each cluster, identifying the type of scam.

By implementing this approach, we aim to not only improve content moderation and provide a cleaner online environment for users but also to protect individuals from the harmful consequences of spam-related scams.

Related Work

Text Classification Models

* + 1. BERT

Bidirectional Encoder Representations from Transformers (BERT) is a powerful pre-trained language model introduced by Devlin et al. in 2018 (Add reference). BERT is based on the Transformer architecture proposed by Vaswani et al. in 2017 (Add reference). It leverages a bidirectional mechanism to understand the context from both left and right sides of a given input, providing rich contextualized word representations. BERT has achieved state-of-the-art performance in various NLP tasks, including text classification, due to its ability to capture complex language patterns.

* + 1. RoBERTa

RoBERTa (Robustly optimized BERT approach) is an optimized version of BERT, introduced by Liu et al. in 2019 (Add reference). RoBERTa modifies BERT's pre-training process, using larger mini batches, more data, and dynamic masking, which allows it to learn deeper contextual representations. This results in better performance in various NLP tasks, including text classification, as demonstrated by its improved benchmark scores.

* + 1. LSTM

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) introduced by Hochreiter and Schmidhuber in 1997 (Add reference). LSTM is designed to overcome the vanishing gradient problem encountered in traditional RNNs by utilizing a gating mechanism. This mechanism enables the LSTM to retain long-term dependencies, making it particularly suitable for sequence-to-sequence problems, including text classification.

Clustering Methods

* + 1. Hierarchical Clustering

Hierarchical clustering is an unsupervised learning method that constructs a tree-like structure (dendrogram) to represent data relationships. The method can be either agglomerative or divisive, depending on whether it follows a bottom-up or top-down approach, respectively. The key advantage of hierarchical clustering is its ability to provide insights into the data hierarchy, making it suitable for grouping similar instances in our spam detection problem.

* + 1. DBScan Clustering

Density-Based Spatial Clustering of Applications with Noise (DBScan) is an unsupervised clustering algorithm proposed by Ester et al. in 1996 (Add reference. DBScan identifies clusters based on the density of data points, grouping closely packed points together and treating sparse regions as noise. It is particularly effective in handling noisy datasets and discovering clusters with arbitrary shapes.

* + 1. K-means Clustering

K-means is a popular centroid-based clustering algorithm introduced by MacQueen in 1967 (Add reference). It aims to partition a dataset into K clusters by iteratively updating cluster centroids until convergence is achieved. K-means is simple, scalable, and efficient, making it suitable for various clustering tasks. However, it is sensitive to initial centroid positions and assumes spherical-shaped clusters, which can be limitations in some scenarios.

Generative Language Model

Generative language models, such as GPT-2 and GPT-3.5, introduced by OpenAI (Add references – 7 and 8) are pretrained models designed to generate coherent and contextually relevant text based on a given input. These models are built upon the Transformer architecture and trained on massive amounts of text data, allowing them to capture intricate language patterns. In our project, we leverage a generative language model to assign labels to spam clusters, identifying the type of scam and providing additional insights into the nature of the spam content.

Datasets

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Rows** | **Source** |
| Text messages | 3000 | Kaggle |
| Email | 3000 | Kaggle |
| NEED TO ADD LINKS PROPERLY |  |  |
|  |  |  |

* 1. Data Collection

## FILL IN

Data Preprocessing

## FILL IN

Methodology

Supervised Learning

* 1. Unsupervised Learning

Generative Models for Labeling

Experiments

Experiment Settings

Empirical Results

1. Future Work

(Train and fine-tune for social media comments)

Conclusion

Acknowledgments

An example acknowledgment.

References

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30, 5998-6008.

[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[4] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[5] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 226-231.

[6] MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1(14), 281-297.

[7] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog, 1(8).

[8] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Sutskever, I. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33.