CS5344 Project Proposal – Group 5

Members: Khoo Zheng Jie Anthony (A0248999W), Lee Seungryeol (A0243766W), Tan Hui Min (A0225203X), Wong Yu Hao (A0242691A), Zhang Yaowen (A0105745H)

Title of Proposal and Abstract

Demystifying the detection of spammers on finance related social media content

Spam started as an email advertisement for a mainframe product sent by a marketer of Digital Equipment Corporation in May 1978[1] to hundreds of users on ARPANET[2]. Presently, spam is any form of unsolicited communication (e.g. spam emails and messages on social media platforms), often malicious, causing detrimental effects to far-reaching entities and states. In the financial industry particularly, the occurrences of "pump and dump" scams[3] have also spiked as social networks have become an important part of life for more people. These platforms are favored mediums for spam or bots for the propagation of stock recommendations based on false, misleading or exaggerated comments, which are often followed or led by noticeable stock price developments, leading to price manipulation.

Introduction

Recent findings suggest: 1) spam campaigns are targeting COVID-19 pandemic, 2) about 85% of all emails are spam, 3) spam email costs companies about $20.5bn annually, 4) scams and frauds account for 2.5% of all spam email, yet phishing figures show identity theft comprises 73% of this amount, and 5) 26.5% of all emails linked to financial matters, e.g. financial news – the 3rd-largest spam email category[4]. Furthermore, 60% of social network users encounter spam comments on a daily and weekly basis[5]. Despite the prevalence of spam in social media, spam filtering techniques implemented by social networks are inferior in comparison to the email service providers (e.g. Gmail blocking over 99.9% of spam emails[6]), leading more users to be susceptible to spam.

As we focus on the severity of finance related spam, our project aims to leverage big data in the form of digital messages posted by social network users to build a comprehensive machine learning model. This model will detect spam on finance-related social media content such as stock market news and views within social networking platforms such as Twitter and Reddit. The model target audiences include investment firms, academia and general public communities.

Description of Data Sources

The raw data including search for stock tickers (where search term will be cashtag[7] e.g. $xxx or hashtag e.g. #$xxx or #xxx for at least 1 million data points) will be extracted from Twitter and selected Reddit communities such as r/WallStreetBets, where retail traders frequently share trading tips, using Python APIs. To create the training data, we will determine and label data as spam / non-spam, then transform the data into feature vectors.

Methodologies

The evolution of spam detection and filtering mechanisms initially developed as an identity-based model or a content-based model (where we detect keywords and patterns consistent across spam from the user itself or the messages they post), to the state-of-the-art Facebook's EdgeRank Algorithm which generates a score from a few features (e.g. content attributes including likes, comments, reposts, etc).

Prior to model decision, we will determine the input features using methodology proposed by Lin et al[8], which investigates three families of spammer behavior, 1) aggressive advertisement, 2) duplicate reposting, 3) aggressive following. Based on these root attributes, we can extract more features again after thorough data analysis. We plan to experiment with supervised learning (e.g. support vector machines (SVM), logistic regression (LR), Naive Bayes (NB), neural network) and unsupervised learning (e.g. DBSCAN). To assess the model, we will use F1 score (harmonic mean of precision and recall) and other performance metrics.

Potential Challenges / Risks

- Challenges:
  - Datasets for spam on financial content are not widely available – we may need to scrap and label user posts manually, a time-consuming process. Although we will explore the option of building a standalone machine learning model for labeling spams, if the results are not satisfactory, we will have to revert to manual labeling.
  - As social network constantly streams an aggregation of unfiltered data from various sources, the dataset is liable to data drift & concept drift.
- Risks:
  - If the model is trained on the dataset with incorrect labels for the ground truth, the model performance on spam detection may not be as satisfactory as being trained on more accurate dataset and may not identify spams accordingly.
  - To overcome the challenge of data drift & concept drift, we should ensure the dataset is extracted from the current and recent time period, else the model performance will be degraded.

References

1. Wikipedia definition of spam: https://en.wikipedia.org/wiki/Spamming
2. ARPANET refers to Advanced Research Projects Agency Network: https://en.wikipedia.org/wiki/ARPANET
3. Pump-and-dump is a scheme that attempts to boost the price of a stock through recommendations based on false, misleading, or greatly exaggerated statements: https://www.investopedia.com/ask/answers/05/061205.asp
4. DataProt spam report: https://dataprot.net/statistics/spam-statistics/
5. Foundation spam report: https://foundationinc.co/lab/social-media-spam
6. Google Gmail article: https://security.googleblog.com/2020/02/improving-malicious-document-detection.html
7. Twitter Glossary: https://help.twitter.com/en/resources/glossary
8. Lin, C., He, J., Zhou, Y., Yang, X., Chen, K., & Song, L. (2013). Analysis and identification of spamming behaviors in Sina Weibo microblog. SNAKDD '13: https://www.semanticscholar.org/paper/Analysis-and-identification-of-spamming-behaviors-Lin-He/2ed188f97a0b2929c8384616535f6bdac132a6d6