An Investigation into the Features of Disinformation that may lead to it being Spread Online

Billy Ly, School of Computer Science and Statistics, Trinity College Dublin
Supervisors: Prof. Owen Conlan, Mr. Dipto Barman, Mr. Ziyi Guo
Ackowledgements: This project was funded by the **Laidlaw Undergraduate Leadership and Research Programme,** Trinity ADAPT Research Centre

Abstract

The explosive growth of social media in recent years has given rise to very fertile grounds for disinformation; whose spread has been difficult to detect and contain. Using a dataset of pre-annotated tweets and news articles discussing the COVID-19 pandemic, we extracted various features of 'True'- and 'False'-labelled tweets, and then had four different machine learning models train on these features.
The results show that on average, models improved at detecting unreliable 'False' tweets when supplied with additional context and the overall sentiment of the text.

Introduction

The internet's rise has been undeniable, with an almost 80% jump in social media users in just five years [1]. In 2021, it was found that just under half of all U.S adults got their news from social media ‘often’ or ‘sometimes [2]. Worryingly, we have also seen that collective action on global issues, e.g. COVID-19, have been repeatedly impeded by the internet's spread of disinformation.
This period has thus been termed an ‘infodemic’, where excessive amounts of information has sown confusion and distrust among the population [3]. This project aims to identify the key features of said disinformation.

Dataset

Our project used the MM-CoVaR dataset; a multimodal dataset containing 2,593 COVID-19 related news articles and 24,184 related tweets, collected between February 2020-May 2021 [4]. Every tweet in the dataset referenced one of the news articles, and each was manually labelled as either 'True', 'False' or 'Inconclusive' based on the reliability of the referenced article and the stance of the tweet.

Tweet Stance

News Reliability	Support	Refute	Not Enough Information
Reliable	T	F	I
Unreliable	F	T	I

Key
T = True
F = False
I = Inconclusive

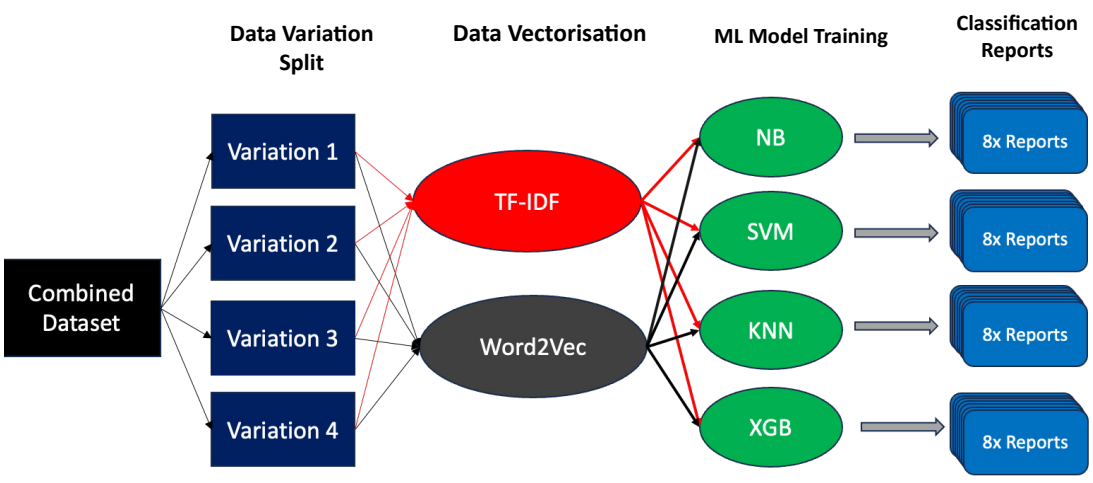
Fig 1. Tweet Labelling Process

Methodology

The machine learning algorithms: Naive Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbour (KNN) and Extreme Gradient Boosting (XGB) were chosen for this project due to their affinity for text classification tasks.

All conclusively labelled tweets and articles were first subject to sentiment and readability analysis using Python tools. A combined dataset then, with these new metrics added, was split up into four variations of data, as seen below. Some of the obtained data, such as readability metrics, ended up not being used to avoid overfitting the models.

Each variation of data then underwent both TF-IDF and Word2Vec vectoriation, for a comparative analysis later. This meant each ML algorithm was trained on 8 different variations of data, and produced 8 different classification reports.



Key
Variation 1: Tweet Tokens (Control)
Variation 2: Tweet Tokens + Tweet Sentiment
Variation 3: Tweet Tokens + Article Tokens + Article Reliability Label
Variation 4: Tweet Tokens + Tweet Sentiment + Article Tokens + Article Reliability Label

Fig 2. Methodology Visualization

Results

TF-IDF Vectorisation Results

Model / Variation	NB	SVM	KNN	XGB
1	Acc: 0.871	Acc: 0.922	Acc: 0.862	Acc: 0.879
2	Acc: 0.865	Acc: 0.921	Acc: 0.862	Acc: 0.879
3	Acc: 0.862	Acc: 0.942	Acc: 0.911	**Acc: 0.958**
4	Acc: 0.916	Acc: 0.944	Acc: 0.917	Acc: 0.957

Key
(SVM,2) = SVM model trained only on second data variation (tweet tokens and tweet sentiment)
Acc: 0.921 = Average accuracy of the model across folds is 0.921

Word2Vec Vectorisation Results

Model / Variation	NB	SVM	KNN	XGB
1	-	Acc: 0.615	Acc: 0.814	Acc: 0.865
2	-	Acc: 0.622	Acc: 0.809	Acc: 0.867
3	-	Acc: 0.757	Acc: 0.925	Acc: 0.951
4	-	Acc: 0.757	Acc: 0.903	**Acc: 0.955**

Conclusion

Across almost all models and variations, the addition of extra information such as article context and text sentiment contributed to the model's accuracy in identifying reliable and unreliable information. Such results suggest there are certain identifiable features that disinformative content possesses. This warrants further investigation; where future work could also use different datasets, ML models and features.

References

1. Dixon, S. (2022). Number of Social Media Users Worldwide 2010-2021. [online] Statista. Available at: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
2. Walker, M. and Matsa, K.E. (2021). News Consumption Across Social Media in 2021. [online] Pew Research Center's Journalism Project. Available at: https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/
3. UN,“Unmisinformationreport,”https://www.un.org/en/un-coronavirus-communications-team/un-tackling-%E2%80%98infodemic%E2%80%99-misinformation-and-cybercrime-covid-19,2020
4. Chen, M., Chu, X. and Subbalakshmi, K.P. (2021). MMCoVaR: Multimodal COVID-19 Vaccine Focused Data Repository for Fake News Detection and a Baseline Architecture for Classification. arXiv (Cornell University).