Evaluating the Effectiveness of a Fact Verification Method on Social Media Posts

Oliver Barnes
1905121

Department of Computer Science
Swansea University

April 28, 2023

Declaration

This work has not been previously accepted in substance for any degree and is not being con- currently submitted in candidature for any degree.

Signed .... (candidate)

Date 28/04/2023

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .... (candidate)

Date 28/04/2023

Statement 2

I hereby give my consent for my thesis, if accepted, to be made available for photocopying and inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate)

Date 28/04/2023

Abstract

Misinformation remains a huge problem in social media spaces, with more created and spread every day. The COVID-19 epidemic exacerbated this, with false information about the virus, its causes, and potential cures permeating online spaces. This information could be harmful to those exposed to it, with those believing it or following its advice potentially sustaining injury.

This paper proposes a solution to flagging this information on Twitter, through verifying the content of the tweets, checking to see if the claims presented conflict with expert opinions. This is done using the BERT language model to check for semantic similarity. Four pre-labelled datasets containing misinformation relating to COVID are used to test this approach.

The semantic matching model was trained on the SNLI corpus, a natural language dataset for inference. A keyword-matching approach was used for evidence retrieval. The approach had an accuracy of around 50%, no better than random labelling, and an average F1 score of about 0.2 in testing. This paper analyses the approach used and attempts to explain why it performed as poorly as it did.

Table of Contents

Chapter 1

Introduction

Misinformation has always been an issue on the internet, especially on social media. Today it remains an effective tool for those who look to spread propaganda and influence others. In today's political climate it's especially prevalent, with misinformation about current events being easier than ever to find[1] – and often presented as fact to unsuspecting internet users[2].

Misinformation, by definition, is "false information, spread regardless of intent to mislead"[3]. While this definition may not present it as inherently harmful, there are cases where it can be. Regardless of intent, the consequences of its spread can be very real to those exposed to it[4].

COVID-19 is one of the most[5] popular topics for misinformation on social media. It has been reported the spread of misinformation about it has cost lives[6] and caused physical harm to numerous people. Whether directly about the virus itself, what causes it, or supposed cures, misinformation is a problem.

Social media is the most helpful tool to those looking to propagate false information, and it's effective to such a degree that 24% of social media users in the UK claim to have been exposed to COVID misinformation over the course of a given week[7]. Another survey states that Americans believe that 65% of the news they see on social media is misinformation[8]. The high volume of users, information, and the ease of spreading it are all factors in this.

This project explores a potential approach for flagging this kind of misinformation, and how it performs when tasked with identifying it in a social media dataset. The datasets this project will focus on are based on data from Twitter, specifically tweets containing dubious information about COVID-19.

[1] https://www.bbc.co.uk/news/blogs-trending-37846860

[2] https://hir.harvard.edu/futureofdisinformation/

[3] https://www.dictionary.com/browse/misinformation

[4] https://www.who.int/europe/news/item/01-09-2022-infodemics-and-misinformation-negatively-affect-people-s-health-behaviours--new-who-review-finds

[5] https://www.statista.com/statistics/1317019/false-information-topics-worldwide/

[6] https://www.bbc.co.uk/news/world-53755067

[7] https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/coronavirus-news-consumption-attitudes-behaviour

[8] https://www.poynter.org/ethics-trust/2018/americans-believe-two-thirds-of-news-on-social-media-is-misinformation/

1.1 Motivations

The primary motivation for this project is to attempt to help with the misinformation infodemic on social media. More specifically it is an attempt to flag information about given subjects, in this case ones that present a particular hazard to public health.

The future of misinformation, and especially disinformation, is a worrying one, with the advancement of technology that can be used to automatically generate false information [1] [2]. Disinformation being misinformation created with an intent to mislead those exposed to it. Developing methods to counter this problem is important to prevent the issues it could eventually cause.

Avoiding any discussions about censorship, this project merely is an attempt to separate neutral information from that which directly contradicts expert opinions and advice. Flagging this information can be useful to platforms like Twitter, especially in the context of ideas like "brand safety"[9], as press about them supporting the spread of misinformation could be harmful to their bottom line.

This project will investigate the viability of a BERT-based [3] fact verification method as a solution, and whether it is worth serious consideration for a larger-scale deployment. While the applications of a semantic BERT model to this problem may seem obvious, the actual performance of any potential implementations of it must be investigated to propose the best possible solution, hence this project.

1.2 Objective

The Objective of the project is as follows:

- To evaluate the performance of a fact verification approach for identifying misinformation. This involves:
    - Creating a robust pipeline for claim analysis and presenting a verdict based on a given claim (either *neutral* or *misinformation*).
    - Testing its performance across various datasets.
    - Testing a given set of (pre-classified) facts for claim comparison for each dataset.
    - Conducting three-fold cross-validation on each of the tests to accurately estimate performance.
    - Collecting metrics for accuracy, precision, recall, and F1 score.

[9] https://en.wikipedia.org/wiki/Brand_safety

1.3 Contributions

The main contributions of this work can be seen as follows:

- **Detailing the method and results of a failed approach**

  This paper outlines the approach implemented and shows the results of it on the given task of misinformation detection.

- **Analysis of the approach and results**

  Given that the implementation failed at the task this paper attempts to explain why and looks at how the method could be improved to give a potentially working method for the task.

Chapter 2

Background

Identifying misinformation is a complicated process. While it's something that humans can do (but aren't necessarily good at), developing some sort of automated solution is required for large quantities of data. As an example, around 500 million tweets are created every day[10], so having humans check even a small percentage of that data is not viable. One study [4] states that almost 25% of the COVID-related tweets they analysed contained misinformation, which is a sizeable proportion. Twitter does have its own counter-misinformation measures[11], a form of crowd-sourced fact-checking, but this is not automated and has a limited reach (it is only available in certain regions).

Research in this area has shown that there are many approaches that can be taken for detecting misinformation using language models. Different forms of analysis based on sentiment and stance can be applied [5].

In theory, either of these methods could be used to classify tweet data. However, several quirks related to tweet data make it more difficult to handle than more conventional text data like news articles, at least by conventional language models. As said by Barbieri *et al* [6], tweet data is mostly noisy, idiosyncratic text.

Sentiment analysis discerns an emotion from text. Extracting a sentiment value from given text information and understanding the response that it's supposed to elicit could potentially be very helpful in determining whether it is misinformation or not [7]. This approach has been used on COVID tweet data, with one method [8] achieving a 0.93 accuracy score on sentiment classification.

Stance detection (in an NLP context) assesses whether a statement supports or opposes a given claim. While like sentiment analysis in a sense, it is more semantically aligned, as a statement could have a negative sentiment while supporting a claim [9]. Stance detection on COVID tweet data has been done, with one approach [10] giving fairly accurate results.

Fact verification is effectively stance detection, where the stance taken by a claim is compared against a base of given facts. A judgement is then made on whether the claim is true or not based on the evidence given.

[10] https://www.dsayce.com/social-media/tweets-day/

[11] https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

The related work for this project has been separated into two sections, tweet data analysis and fact verification, as it's important to consider both the dataset and the approach in the context of this project.

2.1 Tweet Data Analysis

2.1.1 TWEETEVAL: Unified Benchmark and Comparative Evaluation for Tweet Classification

This study [6] covers a range of tasks and approaches for classifying tweet data, to establish a baseline for the tasks on the data type. This is necessary, as regular NLP task baselines can't be applied to tweet data, as already mentioned it is too different to conventional text.

Two of the tasks that this study handles are sentiment analysis and stance detection, classification tasks that can be used as a form of misinformation detection. Three different RoBERTa (BERT variant) models are used: a pre-trained RoBERTa base, one re-trained using tweet data, and one pre-trained from scratch on tweet data. The retrained model performed the best in their testing, with the base and scratch models achieving similar performance. The accuracy achieved by the approach on the stance detection task was roughly 70%, which is passable and shows promise, but also shows room for significant improvement.

The approaches in [6] have many similarities to this project, such as the use of tweet data and the application of a large language model. The insight provided by the analysis of the pre-trained model variants is helpful. While pre-training a BERT model is far outside the scope of this project, their results showed that the impact on performance was moderate but not enormous. Overall, the paper is not massively helpful to this project given the lack of detail regarding the implementation.

2.1.2 Stance Detection in COVID-19 Tweets

This study [11] showcases a stance detection and labelling approach for COVID-related tweets. It compares a variety of approaches for this, including BERT. The version of BERT used was pre-trained specifically for COVID Twitter data.

Four separate datasets featuring different subjects were used for testing the approaches. From their results the BERT-based approaches outperform everything else, but different BERT variants perform better than others on certain subjects.

There's a fair number of similarities between [11] and this project. Between the dataset used and the implementation of the BERT model, they are alike, however the actual approaches here as far as using BERT for sentiment analysis to inform the judgment is different from the approach this project takes.

2.1.3 Claim Identification and Verification on Twitter

This study [12] details two tasks for identifying misinformation in a tweet dataset. The first is analysis of tweets to evaluate whether the claims made need verification. The second is about identifying verified claims relevant to the classification of a new input claim. This second task is claim retrieval, which is integral to fact verification.

The method used for the claim retrieval task was comparison of similarity based on the TF-IDF weights of the claim against the verified facts. This comparison (and ranking) was handled by an SVM. The performance from their approach is stated as "reasonable" in the article.

The approaches to pre-processing of tweets and facts implemented here in [12] are relevant to this project and were used to simplify the process of claim retrieval. Given that this paper doesn't describe any approaches to the verification aspect, it has fairly limited relevance.

2.2 Fact Verification

2.2.1 BERT for Evidence Retrieval and Claim Verification

FEVER is a large dataset of labelled claims (in some cases with evidence) generated from Wikipedia articles, created for the purpose of testing approaches for fact extraction and verification [13]. This paper [14] describes a BERT model for each task, one for extraction and another for verification.

For the extraction stage, the relevant document is first retrieved and then processed by BERT for sentence extraction. This BERT model has been trained on the evidence set provided by FEVER. After this, another BERT model is used (trained on the claim, and the evidence set) and each claim is compared against the sentences extracted from the document. The final label is decided through aggregating the decisions made by the second model. This approach gave around 70% labelling accuracy.

The use of Wikipedia data as justification for claims has been incorporated into this project, even though it takes the form of conventional text instead of tweets. The approach to sentence extraction here is interesting but given the overheads an alternative was used. The actual classification stage (comparing claim against evidence) is very similar to the method used in this project.

2.2.2 An Empirical Comparison of BERT, RoBERTa, and Electra for Fact Verification

This study [15] compares two large language models (BERT and RoBERTa) against Electra, a transformer model used for training large language models, on the task of fact verification. Once again this uses the data provided by the FEVER set.

The study shows the Electra does not perform particularly well on the task despite how much faster it completes it. BERT and RoBERTa perform very well by comparison, with RoBERTa scoring the best on accuracy and F1.

This study shows that while BERT may not be the best model for the task of fact verification, it is certainly viable for it, as accuracy of 95% is not to be overlooked.

2.2.3 Combining Fact Extraction and Verification with Neural Semantic Matching Networks

This study [16] details an approach very similar to [14], in that it attempts roughly the same two tasks, fact extraction and verification. Again, this approach uses the FEVER dataset to benchmark. A variety of models were used to test the performance of the approach.

Semantic matching is the idea of evaluating the meaning of two sentences, and checking to see if they mean the same thing (entailment), mean the opposite (contradiction) or if there is no similarity in the meaning at all (neutral). This paper uses semantic matching to compare claims to sentences given as evidence. Documents were extracted using a keyword matching approach, with separate models handling sentence selection and claim verification. The approach for semantic matching detailed in [16] did not perform particularly well on fact verification with low accuracy scores (around 30-40%).

The key similarities here are the application of semantic matching to verify claims, the difference in this project being the application of the BERT model, as opposed to one that was purpose built. A keyword matching approach has been used in this project for sentence retrieval as opposed to document retrieval.

2.3 This Project

The intention of this project is to implement parts of these pre-existing approaches together to create a semantic equivalence fact verification method for tweet data.

This involves incorporating a semantic matching approach like in [14] and [16], which differs from the approaches used to identify misinformation in tweet data like in [6], [11], and [12].

This project is about evaluating the performance of this approach to see if it performs nearly as well as any of the methods put forward in these other papers, and in effect address the research gap between the two categories of research outlined here.

Chapter 3

Approach

This project uses a BERT semantic matching approach for fact verification to identify misinformation. Four labelled text datasets were used in doing this, lists of tweets classified as misinformation or neutral information. All four were based around the COVID-19 epidemic, with two (the hydroxychloroquine and ivermectin) datasets being scraped from Twitter[12]. The other two (the 5G and fake cures sets) were provided from [17]. All four of these datasets were labelled automatically.

The automatic labelling may potentially mean that not all the sets are labelled accurately. This should in no way affect the testing process and the results presented, as the model's job is simply to evaluate the semantic similarity of two given claims. Whether either is factually accurate is irrelevant in this case, as it would be evaluated on the automatically generated labels. This would only cause a potential problem if a given claim occurred in the set more than once with different labels.

All code for this project was written in Python, using the TensorFlow library and the Transformers library for the BERT model, tokenizer, and optimizer.

3.1 Data Cleaning

All four datasets were originally saved in a .CSV format, with large amounts of unnecessary data packed in. The first step was cutting this down to just the text data and the labels, as these are all that is required for the verification process. Two of the datasets (hydroxychloroquine and ivermectin) also had data in multiple languages, so only the English data was preserved as the BERT model used was only pre-trained on English. Some records had no text, and these were removed also, and then labels were standardized across all the sets (1 for misinformation and 0 for neutral information).

The four datasets have varying sizes, displayed in table 3.1.

[12] The data was collected from 1/1/2020 to 31/12/2021.

Dataset	No. of entries (after cleaning)
Hydroxychloroquine (HCQ)	319,000
Ivermectin (IVM)	282,111
5G	51,107
Fake cures	16,984

Table 3.1: Datasets with sizes.

3.2 Dataset Building

Due to the way that the verification method works, each testing dataset had to be split into fact and claim sets. A 3-fold cross verification approach was employed (detailed in 3.4). Four tests were proposed for the approach, so with the four datasets and the cross-validation this meant 48 respective fact and testing sets were to be built, with one for each test.

Each of the fact sets and claim sets contained an equal proportion of misinformation and neutral information. This was done to establish a baseline of random labelling, and not skew the results.

These fact sets were a set containing the text content of two related Wikipedia articles split into sentences, a set with 5,000 labelled tweets, a set with 10,000 labelled tweets, and a set combining the 10,000 tweets with the Wikipedia data. In the Wiki sets, all data provided was labelled as neutral. This was done to see how the model would perform on varying quantities of facts, and if there was a difference in performance between using conventional text and tweets.

Each of the claim sets consisted of 5,000 claims, none of which should overlap with those in the fact set (as they were removed from the potential selection for fact sets when they were built).

To construct the testing and fact sets, three “buckets” of 5,000 statements were taken from each dataset. Upon taking each bucket it was removed from the full dataset to minimize overlap between them. Different combinations of these buckets were used for creating these sets, and in the case of 10,000 facts set, two of the buckets not used as the claim set would be appended together.

3.3 Verification Approach

As stated, a semantic matching BERT model is used for fact verification. The semantic similarity approach used was an adapted version of the example model provided by Keras[13]. The specific model used was BERT-base-uncased, with 64 trainable layers (427,267 params) added on top of the pretrained model. The Adam optimizer was used, and the model had a learning rate of 1e-3, with two training epochs.

The training data for the model was the same used in the Keras example, the SNLI (Stanford Natural Language Inference) corpus. The data provided is a set of sentence pairs separated by BERT [SEP] tokens. These are all labelled, either as entailment, neutral, or contradiction based on the content of the sentences. This model achieved 79% accuracy on its validation set (from the SNLI corpus). It is like the model used in [16], however that was trained on the FEVER set.

The model was not fine-tuned due to resource limitations, even with very small batches (as small as 4) it would not run with the sheer number of parameters. This is because the model was trained on a consumer GPU that couldn't support it (laptop RTX 3060). To check a given claim, it would need to pass through the semantic matcher with a relevant fact, then given the model's judgment the claim would be labelled accordingly. As stated in 3.2, multiple fact sets were constructed. Given how large these are, it would be infeasible to check any given claim against every fact in the fact set. For this reason, a keyword search was implemented.

In each of the fact sets an additional value was added, being the keywords for each fact. This would be the plaintext fact stripped of stopwords and punctuation. When a claim was passed through the keyword search, it would also be stripped down to keywords, and the facts would be ranked based on the number of intersecting keywords. Only the top 20 highest scoring facts would be used to evaluate the claim. This keyword search approach was also used to work around quirks of the model, where in certain cases completely unrelated statements could result in non-neutral judgements.

Of the 20 evaluated facts, the one with the highest certainty (posterior probability) value for its result would be taken as the real evaluation, with the others being disregarded. Since the result returned from the model would be one of "entailment", "neutral", or "contradiction", this would have to be resolved to a "misinfo" or "neutral" verdict. This was done by checking the label on the fact used to generate the result. If the fact statement was labelled as neutral, then only those receiving a "contradiction" result were flagged as misinformation. If the fact statement was labelled as misinformation, then those receiving a result of "entailment" were flagged as misinformation.

[13] https://keras.io/examples/nlp/semantic_similarity_with_bert/

If a statement that was already in the fact set was passed to program, it would not be checked by the semantic matcher, instead it would receive the same label as the one in the fact set had. This saves processing power and time, as well as working around one notable quirk of the model, where in some cases the same statement could be passed through as both the claim and the fact and still receive a low certainty score.

3.4 Testing

3-fold cross validation was used in testing to give more accurate results. While ideally this project would've used 10-fold validation, this was not possible due to the small size of the datasets (namely the fake cures set) as 5,000 samples had to be used to test. It also would not have been viable for testing due to the sheer amount of data to be processed and the runtime of the program.

The tests were carried out in a Google Compute Engine instance running Debian 11. Command-line parameters were added to the Python script so it could be run automatically via a Bash script. Full testing of the approach took 10 days of continuous runtime, with an extra two days of supplemental testing taking place due to an issue with the script.

Chapter 4

Evaluation

Performance across all datasets (except for the fake cures set) was very poor. The fake cures set will be discussed separately and cannot be evaluated in the same way as the other three sets, as the approach used is flawed and so are the results.

4.1 HCQ, IVM, 5G

The hydroxychloroquine, ivermectin, and 5G sets all presented very similar results. Each variation of the fact bases performed roughly the same on accuracy, however the Wiki sets for each had far higher F1 scores. While it is obvious this approach failed to meet the objective, there are several potential reasons for this.

	Accuracy	**Precision**	**Recall**	**F1**
Wiki	0.424867	0.405627	0.322933	0.359587
5k	0.502533	0.508319	0.1548	0.237326
10k	0.5226	0.57431	0.174667	0.267866
Combi	0.516667	0.551781	0.1776	0.268711

Table 4.1: Results for hydroxychloroquine data.

One possible reason is that the fact sets were too small. As can be seen in table 4.2, the accuracy does improve somewhat from comparatively small Wiki set to the tweet-based ones. This trend is not continued in the other results however, and the difference is small to begin with (~4%) so this is doubtful.

	Accuracy	**Precision**	**Recall**	**F1**
Wiki	0.442133	0.438614	0.413467	0.425669
5k	0.479533	0.425158	0.116267	0.182599
10k	0.4862	0.452061	0.130133	0.202091
Combi	0.480667	0.434389	0.128	0.197734

Table 4.2: Results for ivermectin data.

One apparent feature of the results is the very low scores for recall. Most of the checked tweets were flagged as negative for misinformation, regardless of whether they were or not. Across all three of these sets, this only happens with the larger fact sets, with the Wiki set giving comparatively high scores for recall. This imbalance could be a result of the inclusion of neutral data in the fact set. It could be that the model failed to find any apparent similarities between most claims and the content of the fact bases. As a result, most of the claims would then be flagged as neutral, resulting in this large number of negative results.

	Accuracy	Precision	Recall	F1
Wiki	0.510133	0.508495	0.606533	0.553204
5k	0.4876	0.461916	0.1504	0.226916
10k	0.491133	0.473325	0.157333	0.236165
Combi	0.492133	0.476531	0.159733	0.239265

Table 4.3: Results for 5G data.

There are obvious flaws related to the model and the way in which it was implemented. The first being the fact that the model was trained on the SNLI corpus, a conventional text dataset, as opposed to tweet data. This mismatch between the training data and the test data resulted in the approach failing as it did. If the model were initially trained on semantically equivalent tweets, then there is some chance it would have performed better. Another potential approach that might have worked is a dataset of tweets, with their semantic equivalent as conventional text.

Another potential issue is the fact that the BERT model was not one pretrained on tweet data, like the one used in [11]. However, as demonstrated by [6], this pretraining element seems to just improve performance in approaches that already work, with no evidence suggesting that it fixes those that don't.

An argument could be made that the bad performance of the model is the fault of overfitting, and while this is possible, there are too many other possibilities to consider, namely that the model was not right for the input data to begin with. It's also unlikely to be an overfitting problem due to the performance of the validation set (being only 79% accurate). If anything, it is likely an underfitting problem, and it is possible that with more training the model's performance could've improved significantly.

4.2 Fake Cures

The reason for the seemingly good performance of the approach on this dataset (seen in table 4.4) is that the experimental approach was fundamentally flawed. This is down to an oversight on the dataset itself, namely the number of duplicates in it. Checking the whole set revealed that there were as many as 173 (near) duplicates of a single tweet present in it. More analysis of the test sets revealed that there were as many as 450 exact overlaps between each of the fact sets with their respective test sets, with potentially more being close matches.

	Accuracy	Precision	Recall	F1
Wiki	0.430733	0.437823	0.473363	0.454899
5k	0.622467	0.756645	0.383168	0.508719
10k	0.725333	0.856524	0.548225	0.668544
Combi	0.726333	0.861494	0.549488	0.670995

Table 4.4: Results for fake cures data.

These duplicates were then all correctly labelled, due to one of the aspects of the approach, being the automatic labelling of claims already in the fact set. The presence of these duplicates had a visible impact on the results, shown in table 4.4, with accuracy and F1 scores far higher than any of the other datasets received.

While it’s possible to conclude that the model could be helpful on datasets like this one, the accuracy shown is only 20% better than random labelling at best. Given that around 10% of the test data is identical to given entries in the fact set, with potentially more being nearly identical (only one or two words added or different), this result is not very impressive. Given the especially bad performance on the Wiki set (10% worse than random labelling), it could be hypothesised that there’s no real semantic matching happening.

Given the sheer number of duplicates, whether exact or just close, and the fact that claims identical to those in the fact set would receive an automatic label without being processed by the model, it is easy to see how these results have been inflated. In effect, the approach used on this set ended up being no better than a simple search for the text content of the tweets, while being far more time and resource intensive.

To obtain real results from this dataset, these duplicates would have to be removed. An alternative to this might be disabling the catch for identical statements, so that everything must pass through the model. Either of these measures would likely mean that the results more accurately reflect the performance of the model, as opposed to specific quirks of the dataset.

4.3 Overview

Due to the reasons given here, the approach used in this project was highly flawed, and as a fact verification method it was an outright failure. However, this does answer the research question, this method is not an effective one for fact verification for social media posts. All the objectives outlined in the introduction have also been met, as the pipeline was constructed, and multiple datasets were tested and the results cross-validated.

The software developed for this project does perform the task adequately enough to answer the research question, however it's not well designed and is inefficient, resulting in long runtimes and large resource consumption. It is run from a script that only makes use of a single thread, meaning that it does not benefit from GPU acceleration (as it is bottlenecked by the CPU performance). A more developed version of the method may use Python's multiprocessing library, or a different language altogether. Deployment of the software to the cloud instance was not difficult, likely due to its simplicity (i.e., not making use of a multiprocessing approach), which is one of its strengths.

Given the extensive (and obvious) limitations of the approach, it is easy to see why it did not work in hindsight. However, this project has delivered results that prove this outright.

Chapter 5

Conclusion

How effective is a fact verification method on social media posts? In the case of the method detailed in this project, not effective at all. Overall, while this project met all the objectives outlined and answers the research question, the approach itself was a failure. The contribution made here is minimal at best, simply offering proof that a classifier like this does not work when there is such a large mismatch between the training and testing data.

The quality of some of the actual research is questionable, for the reasons outlined in the fake cures section in the evaluation. The rest is serviceable for evaluating the performance of the method, however given that the cross-validation here is only three-fold (as opposed to the standard ten-fold), it is not perfect.

It's hard to say whether semantic matching for classifying misinformation is promising or not, given how flawed this approach was. More testing would have to be done using a greatly improved version of the model to decide that. It can be said that this project is not a good implementation of misinformation detection though, and this can absolutely be seen from the results.

The research gap that this project intended to fill may yet exist for a good reason. The "noisy, idiosyncratic" nature of tweet data [6] means that classification based on semantic values is likely an inherently flawed idea. Future language models and datasets may address this issue, but for now it seems that this approach is not effective.

Bibliography

[1] J. Bakdash, C. Sample, M. Rankin, M. Kantarcioglu, J. Holmes, S. Kase, E. Zaroukian and B. Szymanski, “The Future of Deception: Machine-Generated and Manipulated Images, Video, and Audio?,” *International Workshop on Social Sensing (SocialSens),* pp. 2-2, 2018.

[2] G. Spitale, N. Biller-Andorno and F. Germani, “AI model GPT-3 (dis)informs us better than humans,” *arXiv preprint arXiv:2301.11924,* 2023.

[3] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv preprint arXiv:1810.04805,* 2019.

[4] R. Kouzy, J. A. Jaoude, A. Kraitem, M. B. E. Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl and K. Baddour, “Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter,” *Cureus,* vol. 12, no. 3, 2020.

[5] Q. Su, M. Wan, X. Liu and C.-R. Huang, “Motivations, Methods and Metrics of Misinformation,” *Natural Language Processing Research,* vol. 1, no. 1-2, pp. 1-13, 2020.

[6] F. Barbieri, J. Camacho-Collados, L. Neves and L. Espinosa-Anke, “TWEETEVAL: Unified Benchmark and Comparative Evaluation for Tweet Classification,” *arXiv preprint arXiv:2010.12421,* 2020.

[7] C. Guo, J. Cao, X. Zhang, K. Shu and M. Yu, “Exploiting Emotions for Fake News Detection on Social Media,” *arXiv preprint arXiv:1903.01728 ,* 2019.

[8] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood and G. S. Choi, “A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis,” *PLoS ONE,* vol. 16, no. 2, 2021.

[9] A. ALDayel and W. Magdy, “Stance detection on social media: State of the art and trends,” *Information Processing & Management,* vol. 58, no. 4, 2021.

[10] Y. Hou, P. v. d. Putten and S. Verberne, “The COVMis-stance dataset: stance detection on twitter for COVID-19 misinformation,” *arXiv preprint arXiv:2204.02000 ,* 2022.

[11] K. Glandt, S. Khanal, Y. Li, D. Caragea and C. Caragea, “Stance Detection in COVID-19 Tweets,” *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing,* vol. 1, 2021.

[12] T. McDonald, Z. Dong, Y. Zhang, R. Hampson, Q. C. James Young, J. L. Leidner and M. Stevenson, “The University of Sheffield at CheckThat! 2020: Claim Identification and Verification on Twitter,” *CLEF (Working Notes),* 2020.

[13] J. Thorne, A. Vlachos, C. Christodoulopoulos and A. Mittal, “FEVER: a large-scale dataset for Fact Extraction and VERification,” *arXiv preprint arXiv:1803.05355,* 2018.

[14] A. Soleimani, C. Monz and M. Worring, “BERT for Evidence Retrieval and Claim Verification,” *Advances in Information Retrieval, Lecture Notes in Computer Science,* vol. 12036, no. 2, pp. 359-366, 2020.

[15] M. Naseer, M. Asvial and R. F. Sari, “An Empirical Comparison of BERT, RoBERTa, and Electra for Fact Verification,” *International Conference on Artificial Intelligence in Information and Communication,* 2021.

[16] Y. Nie, H. Chen and M. Bansal, “Combining Fact Extraction and Verification with Neural Semantic Matching Networks,” *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 33, no. 1, 2019.

[17] N. Micallef, B. He, S. Kumar, M. Ahamad and N. Memon, “The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic,” *2020 IEEE International Conference on Big Data,* pp. 748-757, 2020.

Appendix A

Results

This is the full set of raw results gathered from the project. As mentioned in the report, there are results for four separate datasets, with four tested evidence-bases, cross-verified three times.

The four datasets are: hydroxychloroquine, ivermectin, 5G and fake cures. The evidence bases are the contents of 2 related Wikipedia pages, 5000 labelled tweets, 10,000 labelled tweets, and 10,000 labelled tweets with the Wiki page content included.

The labels on the top of the table refer to the number of correctly labelled claims in the test set, true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

Hydroxychloroquine

Wiki Set

	Correct	**TP**	**FP**	**TN**	**FN**
1	2153	821	1168	1332	1679
2	2130	813	1183	1317	1687
3	2090	788	1198	1302	1712
total	**6373**	**2422**	**3549**	**3951**	**5078**

5K Set

	Correct	**TP**	**FP**	**TN**	**FN**
1	2555	399	344	2156	2101
2	2491	371	380	2120	2129
3	2492	391	399	2101	2109
total	**7538**	**1161**	**1123**	**6377**	**6339**

10K Set

	Correct	TP	FP	TN	FN
1	2655	459	304	2196	2041
2	2593	411	318	2182	2089
3	2591	440	349	2151	2060
total	**7839**	**1310**	**971**	**6529**	**6190**

Combi Set

	Correct	TP	FP	TN	FN
1	2601	457	356	2144	2043
2	2560	435	375	2125	2065
3	2589	440	351	2149	2060
total	**7750**	**1332**	**1082**	**6418**	**6168**

Ivermectin

Wiki Set

	Correct	TP	FP	TN	FN
1	2232	1043	1311	1189	1457
2	2165	1002	1337	1163	1498
3	2235	1056	1321	1179	1444
total	**6632**	**3101**	**3969**	**3531**	**4399**

5K Set

	Correct	TP	FP	TN	FN
1	2422	304	382	2118	2196
2	2386	287	401	2099	2213
3	2385	281	396	2104	2219
total	**7193**	**872**	**1179**	**6321**	**6628**

10K Set

	Correct	TP	FP	TN	FN
1	2454	341	387	2113	2159
2	2446	318	372	2128	2182
3	2393	317	424	2076	2183
total	**7293**	**976**	**1183**	**6317**	**6524**

Combi Set

	Correct	TP	FP	TN	FN
1	2383	341	458	2042	2159
2	2421	314	393	2107	2186
3	2406	305	399	2101	2195
total	**7210**	**960**	**1250**	**6250**	**6540**

Fake Cures

Wiki Set

	Correct	TP	FP	TN	FN
1	2187	1209	1520	978	1293
2	2096	1131	1547	965	1357
3	2178	1223	1508	955	1314
total	**6461**	**3563**	**4575**	**2898**	**3964**

5K Set

	Correct	TP	FP	TN	FN
1	3353	1162	277	2191	1370
2	3005	879	328	2126	1667
3	2979	891	338	2088	1683
total	**9337**	**2932**	**943**	**6405**	**4720**

10K Set

	Correct	TP	FP	TN	FN
1	3695	1450	216	2245	1089
2	3780	1452	201	2328	1019
3	3405	1253	279	2152	1316
total	**10880**	**4155**	**696**	**6725**	**3424**

Combi Set

	Correct	TP	FP	TN	FN
1	3752	1484	201	2268	1047
2	3688	1479	219	2209	1093
3	3455	1223	253	2232	1292
total	**10895**	**4186**	**673**	**6709**	**3432**

5G

Wiki Set

	Correct	TP	FP	TN	FN
1	2551	1513	1462	1038	987
2	2561	1511	1450	1050	989
3	2540	1525	1485	1015	975
total	**7652**	**4549**	**4397**	**3103**	**2951**

5K Set

	Correct	TP	FP	TN	FN
1	2421	354	433	2067	2146
2	2436	368	432	2068	2132
3	2457	406	449	2051	2094
total	**7314**	**1128**	**1314**	**6186**	**6372**

10K Set

	Correct	TP	FP	TN	FN
1	2483	422	439	2061	2078
2	2457	389	432	2068	2111
3	2427	369	442	2058	2131
total	**7367**	**1180**	**1313**	**6187**	**6320**

Combi Set

	Correct	TP	FP	TN	FN
1	2479	424	445	2055	2076
2	2440	376	436	2064	2124
3	2463	398	435	2065	2102
total	**7382**	**1198**	**1316**	**6184**	**6302**