Validating Fake News Detection Software using Metamorphic Testing

Dave Towey
School of Computer Science
University of Nottingham Ningbo China
Ningbo, China
Dave.Towey@nottingham.edu.cn

Yingrui Ma
School of Computer Science
Univerisity of Nottingham Ningbo China
Ningbo, China
scyym1@nottingham.edu.cn

***Abstract*—Since the popularization of social media, news has entered our lives in a digital way. While news is spreading broader and faster, fake news becomes an increasingly heated topic. Fake news detection is therefore more important in both social media and research area. With artificial intelligence technology, many fake news detection software has been developed. One of the biggest challenges is that there is no straight way to test the software. This is because in traditional software testing, one may need to know directly whether a specific output is correct. However, it sometimes takes very long time or there might be no way to figure out the correctness of the output of a fake news detection software. This type of problem is called oracle problem. Metamorphic testing, which was first introduced in 1998, has been applied successfully to solve oracle problems in many different areas involving artificial intelligence. In this paper, metamorphic testing is implemented for one fake news detection software called FakerFact.**

***Index Terms*—fake news, fake news detection, metamorphic testing**

I. Introduction

Fake news has been studied in many academic papers and defined in many different ways. There are three common types of definitions - satire, manipulation and propaganda. Satire is usually in the form of a jest [1], and its entertaining function is more important than its information [2]. An example source of satire is a news program called The Daily Show[1] [2], which uses humor to present news updates but in the style of TV news broadcast [3]. Manipulation (also called Photo Manipulation) refers to news using false narratives but with real images or videos [2]. Propaganda is one of the most prominent forms of fake news since the 2016 US election campaign [2], when fake news reportedly reached its peak of influence [4]. Propaganda is typically created by political institutions to impact on public opinions [2].

The detection of fake news has emerged in recent studies. However, quality assurance (QA) of artificial intelligence (AI) tools such as fake news detection remains a challenge because of the oracle problem. Metamorphic testing (MT) is used in this paper to test fake news detection tools.

II. Background

A. Fake News and its Impact

Recently, with technological developments and the popularisation of the Internet, the burgeoning concept of social media has changed the way news is formed and published, making news spread faster [5], [6]. Especially with the popularity of mobile devices, people can easily follow the trend [7] and it is more convenient to access social media than traditional news organisations [8]. While modern technologies are bringing more convenience to our lives, social media also has some disadvantages such as fake news. Fake news has entered public view and became a controversial topic since the 2016 U.S. election [6], [9]. It also became a major issue for social media companies during 2020 U.S. election (need reference), COVID-19 pandemic [6], and many huge events. Many people now feel entitled to use Internet to spread false or misleading information [6]. While they like, retweet, or send malicious news content on purpose, they can also encounter fake news unconsciously [6]. According to a recent estimate, 21% (out of 29 million) Twitter accounts are bots, which potentially spread fake news [6]. Fake news has been identified in the dissemination of information in many topics, including vaccination, nutrition, stock values [10]. Therefore, the detection of fake news is essential and necessary.

B. Fake News Detection

Due to the changing nature of online news publication, traditional fact-checking is usually not reliable [11]. Fact-checking organisations are often accused of involving political bias [6]. More tools are now using automatic AI techniques to detect fake news [12]. AI algorithms have been working quite well on many classification problems, such as image classification and voice detection, and this is partly because hardware is now cheaper, and bigger datasets are available [7]. Many fake news detection methods have been focusing on the 'fakeness' of fake news, checking how it is imitating real news and then gives an output. But most of them would not convince the users. While people are not tending to trust AI in a complete way (need reference), it is important and more reasonable that the AI would provide explanations or give

Identify applicable funding agency here. If none, delete this.

[1] http://www.cc.com/shows/the-daily-show-with-trevor-noah

suggestive outputs rather than directly tell the users whether the news is fake or not.

Scholars raised concerns about the transparency and accuracy of AI in tagging fake news on social media platforms [6]. That is why testing AI detection tools is necessary.

C. Related Software

There are lots of software in the market that works on fake news detection. In this project, we will focus on those using machine learning or AI techniques. Based on the output label styles, there are mainly two categories of software — the one that gives explicit outputs and gives suggestive outputs.

1) Oigetit Fake News Filter: This is an example of software that gives explicit outputs. Oigetit is a website[2], or an App (on App Store and on Google Play) that uses advanced AI system to distinguish whether the news from different social media platforms is real or fake by showing a label with a reliability score. The users can directly read news from its website. It also provides a search engine for users to enter a news topic they are interested in. There are three types of outputs: 'real news', 'mostly real' and fake news as shown in Table I.

TABLE I
THREE TYPES OF RESULTS

Predicted label	Corresponding reliability score
real news	≥ 65%
mostly real	≥34%, <65%
fake news	<34%

The outputs directly tell the users whether it is fake news or not simply based on its reliability. It does not provide any explanation.

2) FakerFact: FakerFact is an example of giving suggestive outputs. It is a website[3], or a browser extension (in Chrome and Firefox) where the users can enter the news URL (uniform resource locator) or they can enter the news content manually to enhance their understanding of it. Instead of giving explicit outputs, FakerFact gives more suggestive outputs: Journalism, Wiki, Opinion, Satire, Sensational, and AgendaDriven with a reason list explaining why the AI 'thinks' the article shows signs of that label.

Recently, some researchers think that instead of focusing on 'fake', we should focus more on 'news' of fake news [9]. Directly showing the categories of 'fakeness' is not strong enough to convince the users. Suggestive opinions from AI by examining the characteristics of news such as 'satire' or 'sensational' is more useful. After all, no matter how accurate it is, how beautiful the websites are designed, if the users are not convinced by the results, the software will be useless. A study has found that in fake news detection, explanations that provided by AI is very important to let people believe in AI [13]. For these reasons, FakerFact is more suitable for further investigation and testing. This paper will use FakerFact as the main experiment software.

D. Metamorphic Testing

According to Xie et al. [14], QA of machine learning (ML) algorithms remains a challenge. Traditional testing strategies usually focus on one execution of the software and the observation of its output [15]. It is often assumed by the tester that there is an oracle in software program, which means the correctness of the output produced by the software can be accurately decided in a reasonable amount of time [15], [16]. However, when testing fake news detection programs, the correctness of outputs may involve manual work, background reading, or even on-the-spot investigation, which may take a lot of time and may become very inefficient. This is a typical case of the oracle problem. According to Chen et al. [17] and Segura et al. [15], the oracle problem refers to the circumstances where verifying the test result of a given test case is extremely error-prone, difficult or impossible. Programs that suffer from this problem are often referred to as untestable [15].

An effective technique of testing, MT, can be used. Although initially used to generate new test cases based on successful ones [17], [18], MT can be used as an efficient software testing approach that alleviates the oracle problem [15]. One core element of MT is finding metamorphic relations (MRs), which are necessary properties in relation to multiple inputs (source test cases) and their outputs (follow-up test cases) [15], [17]. If a program violates a certain MR, then it must be faulty [15].

According to Zhou et al. [19], MRs can also help users or software testers better understand the software, then more easily achieve their goals. Recently, Segura et al. [20] completed a survey on MT covering 119 papers published from 1998 to 2015, presenting that MT has addressed approximately 295 real-life problems from various areas such as compilers and ML. According to Sagura et al. [15], GNU Compiler Collection (GCC) was detected using MT with more than 100 bugs. ML applications only took up 8% of total MT applications [15] and even less of them focused on fake news detection.

E. Metamorphic relation pattern

Inspired by Zhou et al. [19], metamorphic relation pattern is used to generate effective MRs. A metamorphic relation pattern is an abstraction that can extract a set of concrete MRs [19].

- Definition of Synonymy
  TODO — the example part needs to be updated.
  In natural language, synonymy could mean synonyms, synonymous sentences. In this paper, we define synonymy as news content with the exact same meaning. It could be either achieved by replacing words by their synonyms or use some mechanisms to get synonymous sentences such as back translation. In this project, we

[2] https://www.oigetit.com/breaking
[3] https://www.fakerfact.org/

use back translation, which means to translate a sentence into another language and then translate it back to get a synonymous sentence. For example, ‘Former snooker star Willie Thorne dies aged 66’ and ‘Former snooker star and Hong Kong favourite Willie Thorne dies’ are not an example of ‘similarity’ because the second news contains extra information ‘Hong Kong favourite’. Similarity can be achieved by replacing the keywords with synonyms. For example, “Former snooker star Willie Thorne dies aged 66” and “Old snooker player Willie Thorne dies aged 66” are an example of ‘similarity’.

- Definition of Antonymy
  TODO — the example part needs to be updated.
  Similar to synonymy, in natural language, antonymy could mean antonyms, antonymous sentences. In this paper, we define antonymy as news content with the exact opposite information. It could be either achieved by replacing words by their antonyms or use some mechanisms to get antonymous sentences. For example, “Former snooker star Willie Thorne dies aged 66” and “Former snooker player Willie Thorne does not die aged 66” are an example of ‘contrary’. ‘Former snooker star Willie Thorne dies aged 66’ and ‘Former snooker star Willie Thorne dies aged 166’ are also an example of ‘contrary’. Contrary can be achieved by replacing the keywords with antonyms.
- Definition of Permutation
  TODO

Metamorphic relations

- MR-1: Consistence with Similarity
  If we get the synonymous sentences of the news and then input both (original news and synonymous news) separately into the software, the output results should be the same because both sentences convey the same news content. Mathematically, let c_1 be the first news and c_2 be the second news with ‘similar’ content of the first one. Then the output results should be the same:
  $$l(c_1) = l(c_2)$$
- MR-2: Opposite Results after Contrary
  If we get the antonymous sentences of the news and then input both (original news and antonymous news) separately into the software, the output results should be different because they convey different news content. Mathematically, let c_1 be the first news and c_2 be the second news with ‘opposite’ content of the first one. Then,
  $$l(c_1) \neq l(c_2)$$
- MR-3: Consistence with Reason List
  A reason list is consist of sentences copied from original news. It shows why the software ‘thinks’ the result should be that predicted label. Theoretically, if the reason list sentences are entered as the second news, the predict label should be the same as or a subset of the original news’ result. Mathematically, let x be the input news, and let y be the reason list. It follows that,
  $$l(x) \supseteq l(y)$$
- MR-4: Permutation of Words
  If the words in the dataset sentences are re-ordered, the result should change because the sentence may have different main meanings. For example, “I ate apple” is changed to “apple ate I”. The meanings of these two sentences are different. Mathematically, let x be the news sentences, and let $perm(x)$ be the function of re-ordering:
  $$l(x) = l(perm(x))$$
- MR-5: Permutation of Sentences
  If the order of sentences is changed, the result should not change because the main idea of different sentences may not change. For example, “I ate apple. I hate milk.” is changed to “I hate milk. I ate apple.”. The meanings of these two content are the same. Mathematically, let x be the sentences, and let$perm(x)$ be the function of re-ordering:
  $$l(x) = l(perm(x))$$

Experiment Result

The dataset BS Detector[4] was published on Kaggle. It contains more than 20,000 URLs of news sources that can be inputted into FakerFact website.

TODO

Conclusion

TODO

Acknowledgment

TODO

References

[1] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, “Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News,” in *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*. San Diego, California: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-0802, Jun. 2016, pp. 7–17. [Online]. Available: https://www.aclweb.org/anthology/W16-0802

[2] E. C. T. Jr., Z. W. Lim, and R. Ling, “Defining “Fake News”,” *Digital Journalism*, vol. 6, no. 2, pp. 137–153, 2018. https://doi.org/10.1080/21670811.2017.1360143. [Online]. Available: https://doi.org/10.1080/21670811.2017.1360143

[3] G. Baym, “The Daily Show: Discursive Integration and the Reinvention of Political Journalism,” *Political Communication*, vol. 22, no. 3, pp. 259–276, 2005. https://doi.org/10.1080/10584600591006492. [Online]. Available: https://doi.org/10.1080/10584600591006492

[4] X. Zhou and R. Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities,” *Association for Computing Machinery*, vol. 53, no. 5, Sep. 2020. https://doi.org/10.1145/3395046. [Online]. Available: https://doi.org/10.1145/3395046

[4] https://www.kaggle.com/mrisdal/fake-news

[5] F. A. Ozbay and B. Alatas, “Fake news detection within online social media using supervised artificial intelligence algorithms,” *Physica A: Statistical Mechanics and its Applications*, vol. 540, p. 123174, 2020. https://doi.org/10.1016/j.physa.2019.123174. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378437119317546

[6] W. Ceron, M.-F. de Lima-Santos, and M. G. Quiles, “Fake news agenda in the era of covid-19: Identifying trends through fact-checking content,” *Online Social Networks and Media*, vol. 21, p. 100116, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468696420300562

[7] M. Granik and V. Mesyura, “Fake news detection using naive bayes classifier,” in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 2017. https://doi.org/10.1109/UKRCON.2017.8100379, pp. 900–903.

[8] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” *SIGKDD Explor. Newsl.*, vol. 19, no. 1, p. 22–36, Sep. 2017. https://doi.org/10.1145/3137597.3137600. [Online]. Available: https://doi.org/10.1145/3137597.3137600

[9] E. T. Jr., R. Thomas, and L. Bishop, “What is (fake) news? analyzing news values (and more) in fake stories,” *Media and Communication*, vol. 9, no. 1, pp. 110–119, 2021. [Online]. Available: https://www.cogitatiopress.com/mediaandcommunication/article/view/3331

[10] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018. https://doi.org/10.1126/science.aao2998. [Online]. Available: https://science.sciencemag.org/content/359/6380/1094

[11] N. K. Conroy, V. L. Rubin, and Y. M. Chen, “Automatic deception detection: Methods for finding fake news,” *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, 2015. https://doi.org/10.1002/pra2.2015.145052010082. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2015.145052010082

[12] C. C. Wang, “Fake News and Related Concepts: Definitions and Recent Research Development,” *Contemporary Management Research*, vol. 16, no. 3, pp. 145–174, Sep. 2020. https://doi.org/10.7903/cmr.20677. [Online]. Available: https://www.cmr-journal.org/article/view/20677

[13] J. H. Gallagher. (2021, Jan) When a Story is Breaking, AI Can Help Consumers Identify Fake News. [Online]. Available: https://news.rpi.edu/content/2021/01/21/when-news-breaking-ai-can-help-news-consumers-identify-fake-news

[14] X. Xie, J. W. Ho, C. Murphy, G. Kaiser, B. Xu, and T. Y. Chen, “Testing and validating machine learning classifiers by metamorphic testing,” *Journal of Systems and Software*, vol. 84, no. 4, pp. 544 – 558, 2011, the Ninth International Conference on Quality Software. https://doi.org/10.1016/j.jss.2010.11.920. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0164121210003213

[15] S. Segura, D. Towey, Z. Q. Zhou, and T. Y. Chen, “Metamorphic testing: Testing the untestable,” *IEEE Software*, vol. 37, no. 3, pp. 46–53, 2020. https://doi.org/10.1109/MS.2018.2875968.

[16] E. J. Weyuker, “On Testing Non-Testable Programs,” *The Computer Journal*, vol. 25, no. 4, pp. 465–470, 11 1982. https://doi.org/10.1093/comjnl/25.4.465. [Online]. Available: https://doi.org/10.1093/comjnl/25.4.465

[17] T. Y. Chen, F. C. Kuo, H. Liu, P. L. Poon, D. Towey, T. H. Tse, and Z. Q. Zhou, “Metamorphic Testing: A Review of Challenges and Opportunities,” *ACM Comput. Surv.*, vol. 51, no. 1, Jan. 2018. https://doi.org/10.1145/3143561. [Online]. Available: https://doi.org/10.1145/3143561

[18] T. Y. Chen, S. C. Cheung, and S. M. Yiu, “Metamorphic Testing: A New Approach for Generating Next Test Cases,” 2020.

[19] Z. Q. Zhou, L. Sun, T. Y. Chen, and D. Towey, “Metamorphic relations for enhancing system understanding and use,” *IEEE Transactions on Software Engineering*, vol. 46, no. 10, pp. 1120–1154, 2020. https://doi.org/10.1109/TSE.2018.2876433.

[20] S. Segura, G. Fraser, A. B. Sanchez, and A. Ruiz-Cortés, “A Survey on Metamorphic Testing,” *IEEE Transactions on Software Engineering*, vol. 42, no. 9, pp. 805–824, 2016. https://doi.org/10.1109/TSE.2016.2532875.