A Machine Learning Approach to Fake News Detection Using Knowledge Verification and Natural Language Processing

Marina Danchovsky Ibrishimova and Kin Fun Li(✉)

Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8P5C2, Canada
{marinaibrishimova, kinli}@uvic.ca

Abstract. The term "fake news" gained international popularity as a result of the 2016 US presidential election campaign. It is related to the practice of spreading false and/or misleading information in order to influence popular opinion. This practice is known as disinformation. It is one of the main weapons used in information warfare, which is listed as an emerging cybersecurity threat. In this paper, we explore "fake news" as a disinformation tool. We survey previous efforts in defining and automating the detection process of "fake news". We establish a new fluid definition of "fake news" in terms of relative bias and factual accuracy. We devise a novel framework for fake news detection, based on our proposed definition and using a machine learning model.

1 Introduction

What is fake news? One of the aims of this paper is to define what constitutes fake news in the context of information warfare, and to propose an automated method for fake news detection based on this definition. Symantec 2019 Internet Threat Report lists information warfare as an emerging cybersecurity threat [1]. Information warfare includes the generation and the spread of fabricated claims for the purpose of manipulation.

Information warfare has existed for as long as warfare has existed. As Sun Tzu wrote in 5th Century BC, "All warfare is based on deception," in his book "The Art of War" [2]. Warfare and politics are deeply intertwined [3]. In 2016, the former President of Bulgaria Rosen Plevenliev warned that the Russian government is trying to influence the outcome of Bulgarian elections after losing the elections [8]. In 2017 Parkinson and Kantchev claimed in a Wall Street Journal article that a Bulgarian security agency allegedly obtained a document from a Russian spy outlining a Russian campaign to interfere in the Bulgarian elections. "The document offered advice on how to burnish the candidate's image by planting stories with Moscow-friendly news outlets. The stories were to be closely coordinated, publishing first in fringe blogs before entering mainstream media en masse to create maximum impact and ultimately become election talking points for the party." [9] Bulgarian security agency neither confirmed nor denied detecting Russian interference [10]. Russia denied all allegations, insinuating that they are fake news [11, 12].

L. Barolli et al. (Eds.): INCoS 2019, AISC 1035, pp. 223–234, 2020.
https://doi.org/10.1007/978-3-030-29035-1_22

If a piece of information is not supported by concrete evidence, then its factual accuracy cannot be established. What is “breaking news” to some becomes fake news to others. A universal definition is therefore required before an adequate computational solution can be created. In Sect. 2 of this paper we study various definitions of fake news and propose a new definition based on absolute factual accuracy and relative reliability of the source. In Sect. 3 we introduce previous work in automating the process of fake news detection. In Sect. 4 we propose a novel fake news detection framework, which utilizes both manual and automated knowledge verification and stylistic features. We discuss our results in Sect. 5. In Sect. 6 we discuss future work, and we draw conclusions in Sect. 7.

2 Defining Fake News

Defining fake news is problematic. People tend to regard any news as “fake” if it does not align with their views or agenda [13]. Edson et al. provided a typology of fake news definitions. They studied 34 different papers on fake news published between 2003 and 2017 and constructed a framework for the different types of fake news based on their definitions [14]. The different types, which include propaganda and advertising/public relations, can all be used in information warfare to influence public opinion on a particular topic [6, 7]. In this paper we focus primarily on “fake news” as a disinformation tool. Lazar et al. define fake news as “fabricated information that mimics news media content in form but not in organizational process or intent” [14]. However, the research of Horne and Adali suggests that there are notable differences in form especially when it comes to the titles of fake news [15]. By the definition provided by Lazar et al., the Wall Street Journal article that depicts Russian interference in Bulgarian elections is likely to be “fake news” because the main source that can verify the factual accuracy of the claims in the article, namely the Bulgarian secret service, is refusing to do so [10]. This makes the story appear fabricated.

So what is the likelihood that this story is in fact fabricated? Since its factual accuracy can be neither confirmed nor denied, then it is equally likely that it is fabricated and factually accurate. The Wall Street Journal has a good reputation, so this likelihood grows slightly larger but probably not by much. After all, even reputable news outlets have produced factually inaccurate news in the past [16, 17].

In 2002 the Associated Press fired one of their reporters after it was discovered that he had been fabricating facts and sources in his news reports for at least 2 years [17]. Just because a piece of information appears in one news outlet, it does not make it factually accurate even if the outlet is a respectable news agency with rigorous organizational processes in place. Nevertheless, a generally reliable source does slightly increase the likelihood that a piece of information is factually correct especially if the opposite cannot be proven. This increase is relative to the overall reliability of the source.

What other factors can increase the likelihood? Are there similar reports that offer more concrete evidence? In the case of the Wall Street Journal article on Russian interference in foreign elections there are. Although the Russians have repeatedly denied interfering in the political affairs of foreign countries [11, 12], a Czech secret service agency pointed at evidence to the contrary as early as 2008 [5]. The Czech Security Information Service unequivocally stated that "operations of intelligence services of the Russian Federation … are by far the most active ones in our territory". In addition, the Czech Security Information Service warned that Russia readopted and repurposed the tactics from Soviet times known as "active measures" [5]. Obviously, this information still does not prove the claim in the Wall Street Journal article about Russian election interference in Bulgaria. However, it does slightly increase the likelihood that it is factually accurate considering that Bulgaria and the Czech Republic, which was formerly a part of Czechoslovakia, are similar in that they were satellite states of the Soviet Union.

In cases where it is difficult to establish the factual accuracy of a piece of information, the line between a fabricated report and a factually accurate report becomes blurry. This is why the fact-checking website Politifact uses a Truth-o-Meter to score political facts [22]. In this paper we introduce "the fake news spectrum" [18]. It takes into account reports that can neither be factually verified nor disputed, even by professional fact-checkers, and even if they come from reliable sources.

In particular, if the accuracy of a report, or a claim can neither be confirmed nor denied, then there is a 50% chance that it is fabricated. If in addition the source is generally reliable or if a similar claim appears in a generally reliable source, then that percentage becomes slightly lower, relative to the overall reliability of the source(s). We can also take into account whether similar claims can be found in various other sources that follow rigorous organizational processes. In addition, we can also take into account whether the claim appears to be based on facts or opinions. In Sect. 4 we describe a machine learning model based on our previous work in incident classification and these observations. In the next section we discuss the existing methods for fake news detection.

3 Automating Fake News Detection

Establishing the factual accuracy of a claim is crucial in determining whether it is "fake news" by most definitions of "fake news". Several manually generated tools for identifying the factual accuracy of a given claim exist. However, such approaches rely on humans who may or may not be objective [18–20]. Additionally, evidence to support the factual accuracy of the claim might not be available as in the case of the Wall Street Journal article on Russian interference in Bulgarian elections. Various automated fact-checking methods have also been proposed. Thorne and Vlachos provide a comprehensive survey of existing automated fact-checking methods. One method uses Recognizing Textual Entailment (RTE) where "RTE-based models assume that the textual evidence to fact check a claim is given" as part of the claim [21]. Another method relies on checking a claim against a knowledge database of proven facts. Yet another method attempts to verify claims by profiling their source and implementing "credit history" of individual sources [21]. Thorne and Vlachos identify issues with all of these methods. Namely, RTE-based methods fail when there is no evidence to support the claim, the "database of proven facts" methods fail when presented with novel claims, and the "profiling the source" methods fail when the source is new. There is an even greater issue associated with fact-checking political news. As Coleman suggests, "Political truth is never neutral, objective or absolute" [4]. Even computational giants such as Google could not tackle this issue and had to shut down their fact-checking tool out of concerns over inaccuracy [22]. Although individually these methods all have weakness, it is worth studying different combinations of them.

In addition to the factual accuracy of a claim, researchers also studied extensively whether its stylistic form can reveal if it is fabricated [23–25]. Oshikawa et al. provide a comprehensive survey on methods using Natural Language Processing (NLP) [26]. Another survey by Groendahl and Asokan asserts that "while certain linguistic features have been indicative of deception in certain corpora, they fail to generalize across divergent semantic domains" [27]. However, they do admit that "some results have been replicated in multiple studies" [27].

Groendahl and Asokan focus primarily on fake news detection methods at the document level as opposed to at the level of the news title. Their survey does not include the work of Horne and Adali who show that the title of a news article is often sufficient to detect if it is fake news [15]. An ordinary news title is typically written in a way to entice the reader to read the entire article. Political disinformation campaigns' main purpose is to spread their narratives to as many people as possible, including to people who do not like to read much. A fake news title as a tool for political disinformation is typically a summary of the entire article [15]. Horne and Adali explore a wide range of syntactic, psychologic, and stylistic features for machine learning models using several different datasets of political news and come to the conclusion that fake news titles are generally longer, have "significantly fewer stop-words and nouns, while using significantly more proper nouns and verb phrases". Of the many different features they test, Horne and Adali identify the top 4 features for classifying fake news' titles: "the percent of stopwords, number of nouns, average word length, and FKE readability." They achieve an accuracy of about 70% [15].

Researchers have also studied hybrid frameworks that employ natural language features and verification features. Conroy et al. survey the various different fake news detection technologies and outline a hybrid framework, which uses content cues (natural language processing tools to detect deceptive language) as well as information about the network, and source verification [28]. However, as Tschiatschek et al. point out, "[it is] difficult to design methods based on estimating source reliability and network structure as the number of users who act as sources is diverse and gigantic (e.g., over one billion users on Facebook); and the sources of fake news could be normal users who unintentionally share a news story without realizing that the news is fake," [29]. Tschiatschek et al. propose a system that relies on crowd-sourcing fake news detection using trusted users to flag potentially deceptive content, which is then forwarded to professional fact-checkers for further investigation. However, there are ethical implications to be considered when profiling users. Zhang and Ghourbani provide a comprehensive survey on fake news detection, "an exhaustive set of hand-crafted features, and the existing datasets for training supervised models" and also acknowledge the importance of having a clear definition of fake news [30].

4 Our Proposed Framework

We propose a hybrid framework for fake news detection, which repurposes the machine learning model for incident classification we previously described [31, 32]. Our incident classification model consists of 5 NLP features combined with 3 knowledge verification features in the form of questions related to the scope, the spread, and the reliability of the source. The ternary answers to these questions are obtained from the user submitting a textual incident report and can be verified independently by the system.

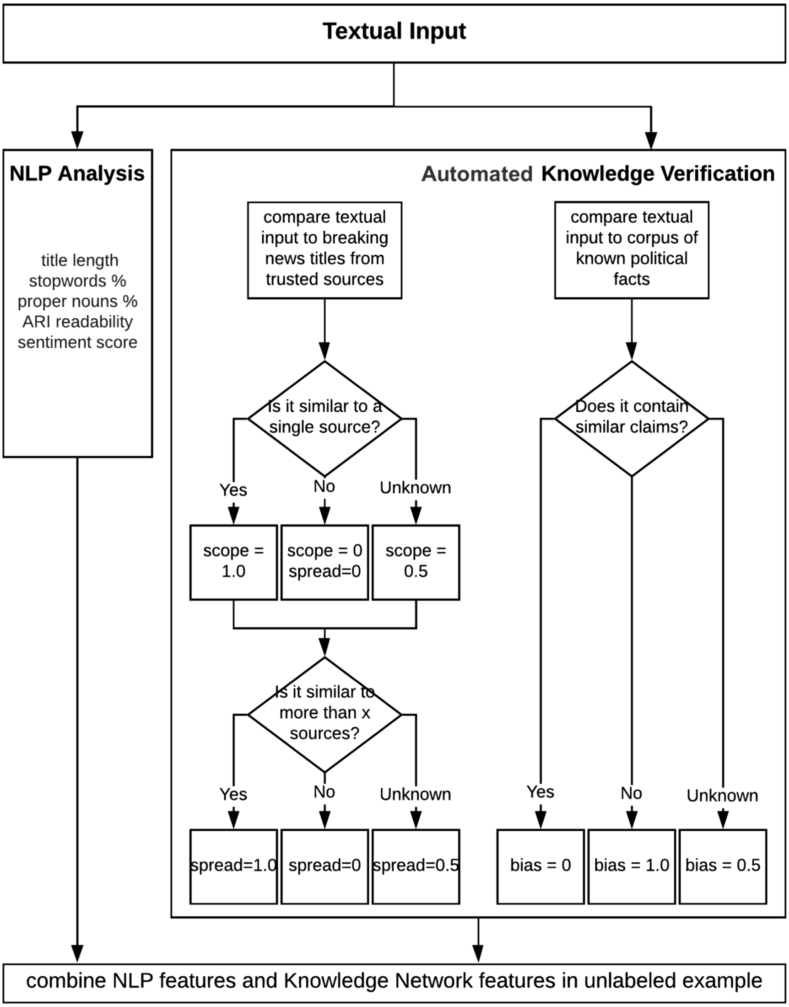

Fig. 1. Automated knowledge verification

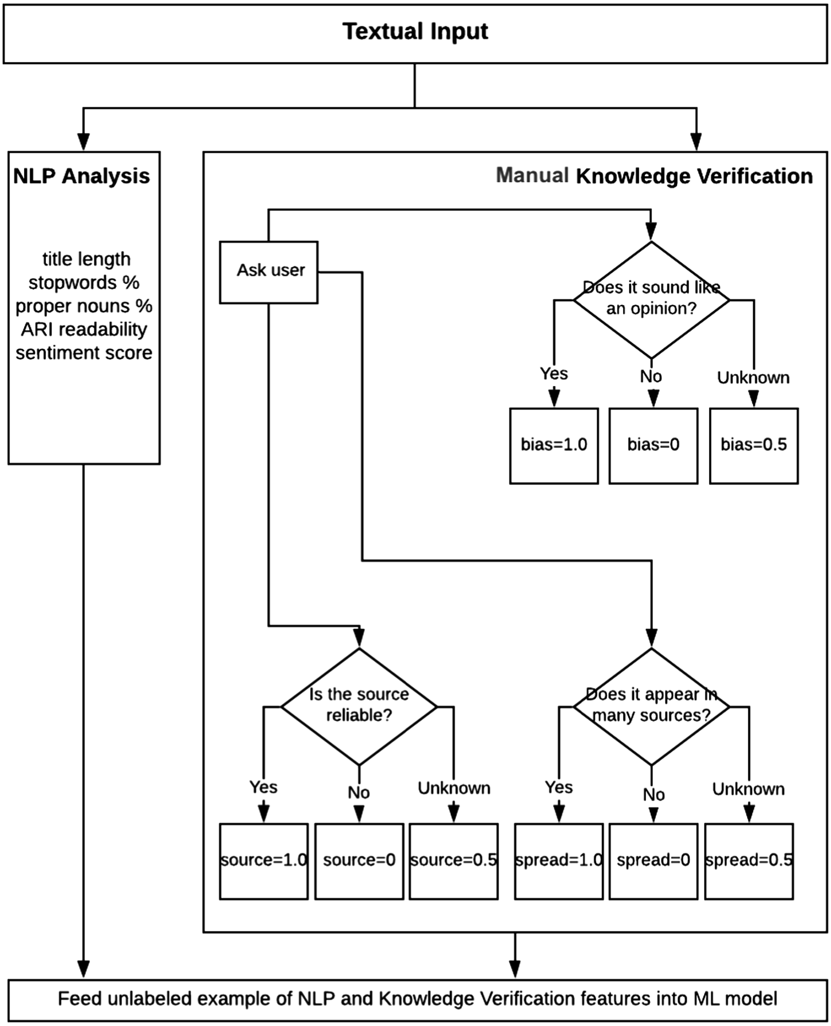

Fig. 2. Manual knowledge verification

We propose using the same general model for the detection of fake news titles where:

1. The 5 NLP features are: stopwords percentage; ratio of proper nouns to nouns; title length; ARI readability; overall sentiment of the text using Google NLP API for sentiment analysis.
2. The 3 knowledge verification features are ternary answers to the questions whether the title is similar to a recent title in a trusted source, whether similar titles appear in more than x sources, and whether the title appears to be based on facts or opinions.

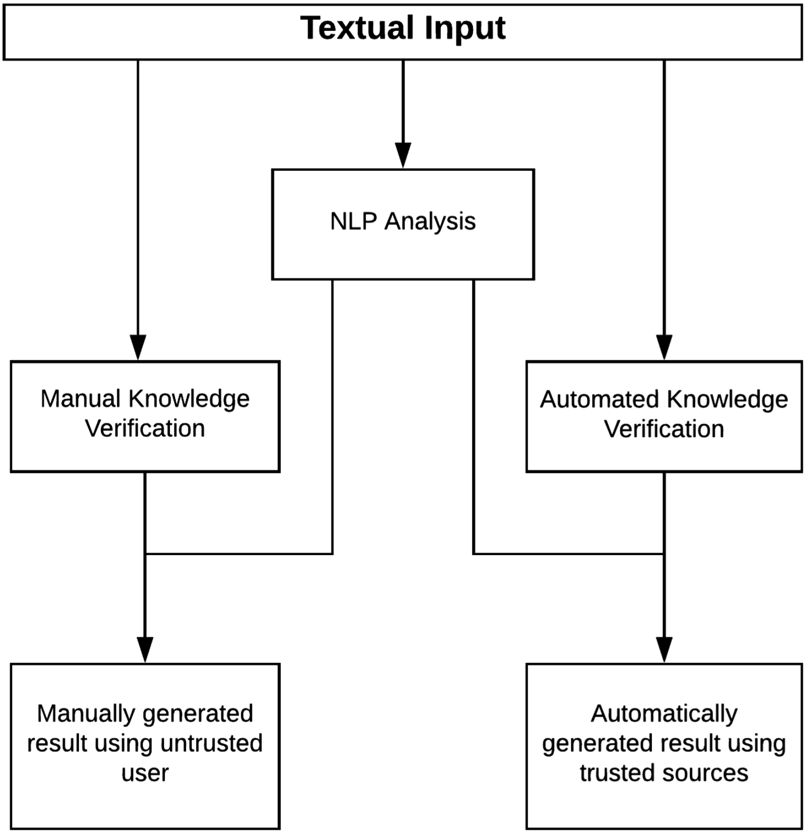

Fig. 3. Simultaneous manual and automated knowledge verification

Our model is different than Horne and Adali in that it uses knowledge verification features in addition to NLP features. Also, our model includes 1 more feature related to sentiment analysis, and we use the automated readability index (ARI), which is a readability test for English texts slightly different than the one used in [15]. Horne and Adali also observed that fake news use more proper nouns but less nouns overall [15], therefore we decided to use the proper nouns (entities) to nouns ratio instead of the number of nouns.

Our system is similar to the one outlined by Conroy et al. in that it is a hybrid framework. However, our proposed verification features are different in that the claim is checked for similarities with recent claims made by a trusted source. The spread of the claim over a variety of trusted sources is also measured. Finally the likelihood that the claim is fact-based or opinion based is established.

Our system can be used in various ways. Namely the verification process can be automated as described in Fig. 1 using trusted sources and datasets of known political facts such as FEVER: a large-scale dataset for Fact Extraction and VERification

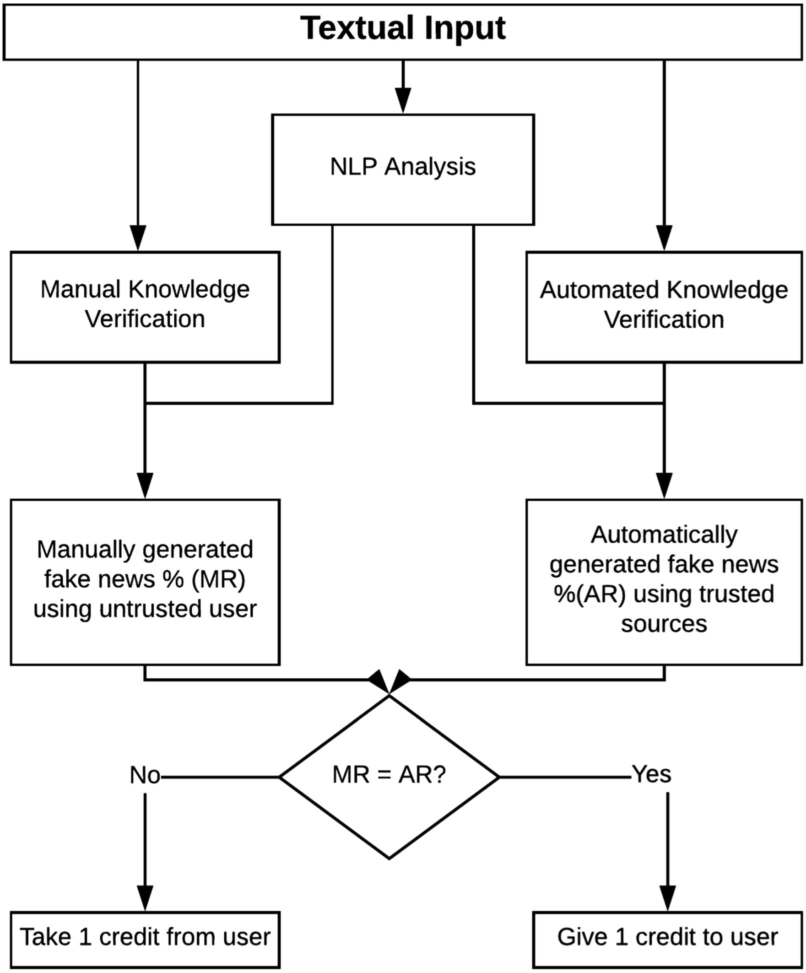

Fig. 4. Verifying user's ability to verify knowledge stated in news title

introduced by Thorne et al. [33] or ClaimBuster by Hassan et al. [34]. Figure 2 shows how the verification process can be manually generated by asking a user to answer questions related to the reliability of the source, the spread of the information, and its bias from the perspective of the user. We propose allowing users to take part in the investigative process by answering the verification questions described in Fig. 2. This way, a manually generated fake news percentage result (MR) is obtained. Behind the scenes an automatically generated fake news percentage result (AR) on the same trained model is also obtained as described in Fig. 1. The system then compares MR and AR as shown in Fig. 3. If the two results are very different or they both suggest the claim is fake, the claim is sent to a professional fact-checker for further investigation. Alternatively, the system can be used to test the fake news detection aptitude of users in a social network. Figure 4 describes one such application. This user-verification system is different from the DETECTIVE system described by Tschiatschek et al. [29] in that it does not rely on users to directly flag news they perceive as fake. Instead, it outsources the knowledge verification to the user and simultaneously performs an automated knowledge verification on its own. It creates two different labeled examples to be fed to a trained machine learning model, and finally it compares the 2 results returned by the model.

5 Result Analysis

We implemented the system described in Fig. 1 using Google NLP API [40] for the NLP analysis portion and News API [41] for the knowledge verification portion. We used the logistic regression model described in our paper on incident classification [31, 32] as our machine learning model for fake news detection. We trained it and validated it on Kaggle Fake News Dataset [42]. We generated a separate test dataset of actual news titles as reported by several verified news outlets with rigorous journalistic practices and we ran each example in our set through our implemented system. Table 1 shows the probability that these news titles are fake as determined by our system.

Table 1. The probabilities that real news titles are fake as determined by our system

Real news	Probability of being fake news
US expected to allow lawsuits against foreign companies doing business in Cuba	10%
Centre-right opposition wins Estonia election as far-right populists make inroads	30%
Vatican to open secret archives of wartime Pope Pius XII	12%
Netherlands summons ambassador to Iran amid diplomatic spat	39%
Residents evacuate as flash floods hit southern Afghanistan	39%
Rep. Dingell on what Cohen's testimony means for future investigations of Trump	45%
OxyContin drug maker mulls bankruptcy due to myriad lawsuits	37%
Fox News confirms exit of Eboni K. Williams	26%
AG William Barr not recusing himself from Russia probe, official says	5%

We generated a separate dataset of fake news by negating crucial words of news titles in the first dataset and we ran each example in our set through our implemented system. Table 2 shows the probability that these fake news titles are fake as determined by our system.

Table 2. The probability that fake news titles are determined as fake

Fake news	Probability of being fake news
US expected to disallow lawsuits against foreign companies doing business in Cuba	80%
Centre-right opposition doesn't win Estonia election as far-right populists make	95%
Vatican won't open secret archives of wartime Pope Pius XII	71%
Netherlands doesn't summon ambassador to Iran amid diplomatic spat	68%
Residents don't evacuate as flash floods hit southern Afghanistan	80%
Rep. Dingell on what Cohen's testimony doesn't mean for future investigations of Trump	84%
OxyContin drug maker doesn't mull bankruptcy due to myriad lawsuits	71%
Fox News denies exit of Eboni K. Williams	74%
AG William Barr recusing himself from Russia probe, official says	40%

6 Future Work

The automated knowledge verification proposed in this paper relies heavily on the notion of similarity. In particular, it relies on establishing semantic similarity. Several state-of-the-art algorithms designed for this purpose have been introduced over the last 20 years, namely latent semantic analysis [35], latent relational analysis [36], explicit semantic analysis [37], temporal semantic analysis [38], distributed semantic analysis [39]. Future work in automating fake news detection using our proposed framework would involve evaluating these algorithms and deciding on one that best fits our purpose.

7 Conclusion

In this paper we define fake news in the context of information warfare. We briefly study the socio-political implications of fake news and we investigate previous efforts in automating fake news detection. We find that the most promising framework for fake news detection uses a combination of source and fact verification and NLP analysis, and we propose a hybrid framework based on our previous work in automating incident classification.

References

1. Symantec Internet Security Threat Report 2019. https://www.symantec.com/en/sg/security-center/threat-report. Accessed 07 Mar 2019
2. Tzu, S.: The Art of War. China. 5th Century BC
3. Kaiser, D.: Politics and War: European Conflict from Philip I1 to Hitler, pp. 149 and 172. Harvard University Press, Cambridge (1990)

4. Coleman, S.: The elusiveness of political truth: from the conceit of objectivity to intersubjective judgement. Eur. J. Commun. **33**(2), 157–171 (2018). https://doi.org/10.1177/0267323118760319
5. Abrams, S.: Beyond propaganda: soviet active measures in Putin's Russia. Connections **15**(1), 5–31 (2016). http://www.jstor.org/stable/26326426
6. Wall, T.: U.S. Psychological Warfare and Civilian Targeting. Peace Rev. **22**(3), 288–294 (2010). SocINDEX with Full Text. Web. 20 Feb. 2015
7. Fried, D., Polyakova, A.: Democratic defence against disinformation. https://www.atlanticcouncil.org/images/publications/Democratic_Defense_Against_Disinformation_FINAL.pdf. Accessed 04 Apr 2019
8. Correra, G.: Bulgaria warns of Russian attempts to divide Europe. BBC (2016). https://www.bbc.com/news/world-europe-37867591. Accessed 19 Feb 2019
9. Parkinson, J., Kantchev, G.: Wall Street J. (2017). https://www.wsj.com/articles/how-does-russia-meddle-in-elections-look-at-bulgaria-1490282352. Accessed 19 Feb 2019
10. Vassilev, I.: Russia's elections manual for BSP and the role played by Bulgarian sociological agencies. Bulgaria Analytica (2017). http://bulgariaanalytica.org/en/2017/03/31/
11. Sputnik News: While West Claims Moscow is Meddling in Bulgarian Elections, Sofia Blames Turkey (2017). https://sputniknews.com/europe/201703251051961027-bulgarian-election-meddling-accusations/, Retrieved 04.04.2019
12. Blyskov, P.: Dmitry Medvedev's interview with Bulgarian newspaper Trud. http://government.ru/en/news/35903/. Accessed 10 Apr 2019
13. Edson, C., Tandoc, Jr., Lim, Z.W., Ling, R.: Defining fake news. Digit. J. **6**(2), 137–153 (2018). https://doi.org/10.1080/21670811.2017.1360143
14. Lazer, D., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018)
15. Horne, B.D., Adali, S.: This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. arXiv preprint arXiv:1703.09398 (2017)
16. Wang, P., Angarita, R., Renna, I.: Is this the era of misinformation yet: combining social bots and fake news to deceive the masses. In: Companion Proceedings of the Web Conference 2018 (WWW 2018). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 1557–1561 (2018)
17. https://www.nytimes.com/2002/10/22/us/ap-says-it-couldn-t-find-45-of-fired-writer-s-sources.html. Accessed 10 Apr 2019
18. Factcheck. https://www.factcheck.org/about/our-mission/. Accessed 02 Apr 2019
19. Snopes. https://www.snopes.com/about-snopes/. Accessed 02 Apr 2019
20. PolitiFact. https://www.politifact.com/truth-o-meter/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/. Accessed 02 Apr 2019
21. Thorne, J., Vlachos, A.: Automated fact checking: task formulations, methods and future directions. In: Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018) (2018)
22. https://www.poynter.org/fact-checking/2018/google-suspends-fact-checking-feature-over-quality-concerns/. Accessed 10 Apr 2019
23. O'Brien, N., Latessa, S., Evangelopoulos, G., Boix, X.: The language of fake news: opening the black-box of deep learning based detectors. In: Workshop on AI for Social Good, NIPS 2018 (2018). http://hdl.handle.net/1721.1/120056
24. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: Truth of varying shades: analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2931–2937 (2017)
25. Long, Y., Lu, Q., Xiang, R., Li, M., Huang, C.R.: Fake news detection through multi-perspective speaker profiles. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (2017)

26. Oshikawa, R., Qian, J., Wang, W.Y.: A Survey on Natural Language Processing for Fake News Detection. Computation and Language (2018). arXiv:1811.00770v1 [cs.CL]
27. Gröndahl, T., Asokan, N.: Text Analysis in Adversarial Settings: Does Deception Leave a Stylistic Trace? Computation and Language (2019). arXiv:1902.08939v2 [cs.CL]
28. Conroy, N.J., Chen, Y., Rubin, V.L.: Automatic deception detection: methods for finding fake news. In: The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST 2015), 6–10 November, St. Louis (2015)
29. Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., Krause, A.: Fake news detection in social networks via crowd signals. In: Companion Proceedings of the The Web Conference 2018 (WWW 2018). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 517–524 (2018). https://doi.org/10.1145/3184558.3188722
30. Zhang, X., Ghorbani, A.: An overview of online fake news: characterization, detection, and discussion. Inf. Process. Manag. (2019). ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2019.03.004
31. Ibrishimova, M.D.: Cyber incident classification: issues and challenges. In: Xhafa, F., Leu, F.Y., Ficco, M., Yang, C.T. (eds.) Advances on P2P, Parallel, Grid, Cloud and Internet Computing. 3PGCIC 2018. Lecture Notes on Data Engineering and Communications Technologies, vol. 24 (2018)
32. Ibrishimova, M.D., Li, K.F.: Automating incident classification using sentiment analysis and machine learning. In: Traore, I., Woungang, I., Ahmed, S., Malik, Y. (eds.) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, ISDDC 2018. Lecture Notes in Computer Science, vol. 11317 (2018)
33. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and verification. In: NAACL-HLT (2018)
34. Hassan, N., Arslan, F., Li, C., Tremayne, M.: Toward automated fact-checking: detecting check-worthy factual claims by claimbuster. In: Proceedings of the 23rd ACM York, NY, USA, pp. 1803–1812 (2017). https://doi.org/10.1145/3097983.3098131
35. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Process. **25**(2–3), 259–284 (1998). https://doi.org/10.1080/01638539809545028
36. Turney, P.D.: Measuring semantic similarity by latent relational analysis. In: Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, IJCAI 2005, Edinburgh, Scotland, pp. 1136–1141 (2005)
37. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: Sangal, R., Mehta, H., Bagga, R.K. (eds.) Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI 2007), pp. 1606–1611. Morgan Kaufmann Publishers Inc., San Francisco (2007)
38. Radinsky, K., Agichtein, E., Gabrilovich, E., Markovitch, S.: A word at a time: computing word relatedness using temporal semantic analysis. In: Proceedings of the 20th International Conference on World Wide Web (WWW 2011), pp. 337–346. ACM, New York (2011). http://dx.doi.org/10.1145/1963405.1963455
39. Nguyen, T.H., Di Francesco, M., Ylä-Jääski, A.: Extracting knowledge from Wikipedia articles through distributed semantic analysis. In: Lindstaedt, S., Granitzer, M. (eds.) Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies (i-Know 2013), Article 6, 8 p. ACM, New York (2013). https://doi.org/10.1145/2494188.2494195
40. Google NLP API. https://cloud.google.com/natural-language/. Accessed 02 Apr 2019
41. News API. https://newsapi.org. Accessed 02 Apr 2019
42. Fake News Dataset. https://www.kaggle.com/c/fake-news/data. Accessed 02 Apr 2019