Artificial Intelligence for Factchecking: Observations on the State and Practicality of the Art

Yudhanjaya Wijeratne, Dimuthu C. Attanayake
LIRNEasia, 12 Balcombe Place, Colombo, Sri Lanka (yudhanjaya@lirneasia.net)

LIRNEasia is a pro-poor, pro-market think tank whose mission is *catalyzing policy change through research to improve people's lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology.*

Abstract

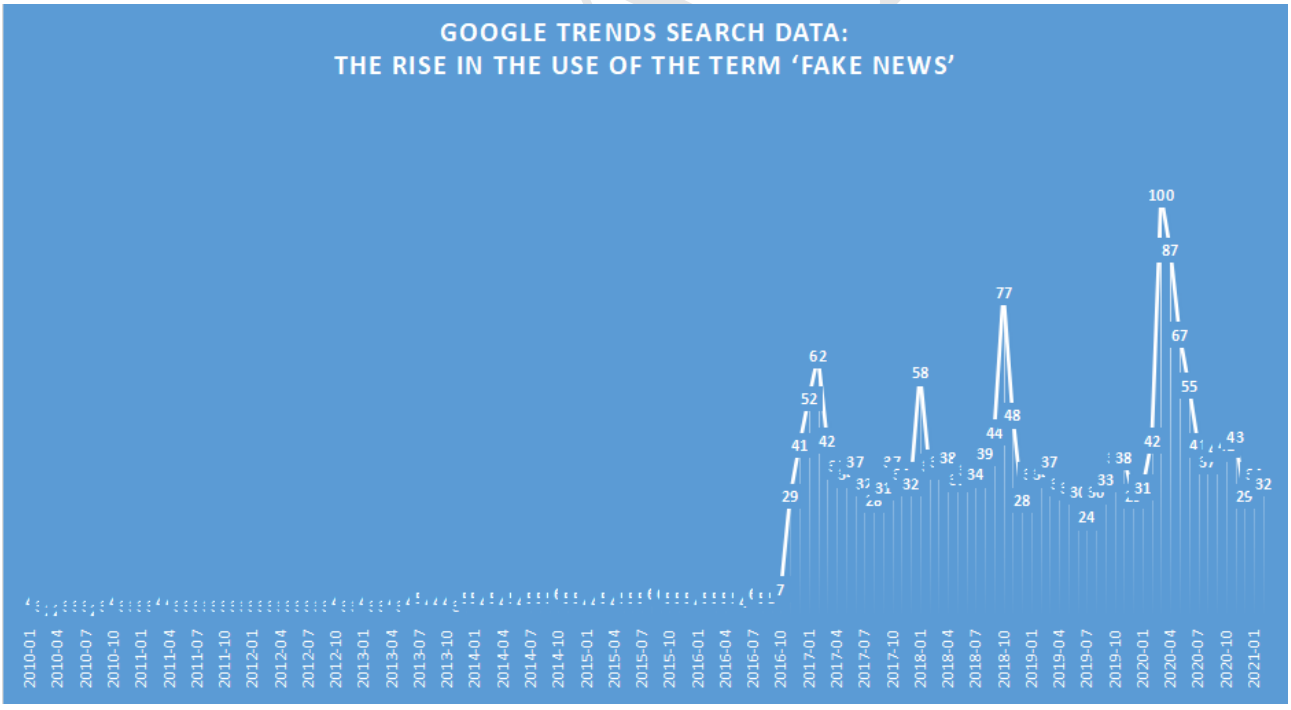
We summarize the state of progress in artificial intelligence as used for classifying misinformation, or 'fake news'. Making a case for AI in an assistive capacity for factchecking, we briefly examine the history of the field, divide current work into 'classical machine learning' and 'deep learning', and for both, examine the work that has led to certain algorithms becoming the de facto standards for this type of text classification task.

Keywords: machine learning, misinformation, text classification, natural language processing

Preface

While lying and fabrication may be as old as language itself, and possibly even the human species (Dor, 2017) [1], the 2016 US election brought with it a multi-disciplinary, mass awareness of misinformation and its effects. Various concerns have been raised about the flood of fabricated content and the erosion of notions of objectivity and balance in public discourse (Del Vicario et al., 2016 [2]; Lazer et al., 2018 [3]); these concerns spilled over from journalism to fields as far removed as economics (Allcott & Matthew, 2017 [4]).

This general uptick is reflected in Google search traffic worldwide for 'fake news', the layman term for various types of misinformation:



However, interest does not necessarily equate to solutions. Two recent phenomena have brought the overwhelming power of today's misinformation back into the public eye. The first is misinformation around COVID-19, medical science and epidemics; many public health care services across the world have found themselves battling both a pandemic and rampant conspiracy theories and public mistrust (Limaye et al., 2020 [5]), to the point where people put both themselves and others at risk simply because they do not believe in the existence of a pandemic.

The second concerns the political sphere; former U.S. President Donald Trump, along with many elected representatives in the Republican Party of the United States of America, actively encouraged a repeatedly-debunked conspiracy alleging that the 2020 elections were fraudulent, culminating in a wave of armed protesters attempting an insurrection in Washington, D.C. Evidence indicates that these people had been stewing in misinformation for years, with conspiracy theories migrating from the fringes of the Internet to the media (Roose, 2021 [6]) and to the highest offices of the most powerful nation on the planet (Tollefson, 2021 [7]).

Drawing from experience with Watchdog Sri Lanka, a tech-heavy factchecking service[1], as well as un-recorded conversations with factchecking initiatives and researchers from Sri Lanka, India, Bangladesh, Ukraine, Singapore and South Korea), we posit that factcheckers around the world are woefully ill-equipped, with nowhere near enough resources to turn the tide.

In factchecking, tests invoking coherence and correspondence to test a truth consume both time and effort, while generating an untruth takes no such effort. In short, as Brandolini's law so elegantly puts it, "The amount of energy needed to refute bullshit is an order of magnitude larger than to produce it."[2]

While wisdom-of-the-crowds approaches have been proposed, such crowdsourcing is merely a form of *consensus gentium*. Rubin (2010) [8], comparing prior literature, pointed out that the human ability to detect misinformation and deception is quite pitiful in comparison, typically under 60% accuracy, with trained professionals performing only slightly better. Any wisdom of the crowds is therefore a poor determinant of truth. The same goes for authority (in many cases) and naive realism.

Therefore, at any given point in a crisis, the potential volume and velocity of misinformation far outstrips the ability of even large organizations to counter it. At a technical level this is a prime use case for artificial intelligence (AI). Much of the narrative around artificial intelligence is in its ability to automate and upscale work; what is usually considered a threat to jobs may, in this state of the world, be a relief to organizations that are drastically understaffed to face the challenge at hand.

What AI has got to do with all this? Definitions, limitations, and clarifications

We will first attempt some clarity by examining definitions. The term 'AI' - Artificial Intelligence - is the overall moniker for a wide variety of approaches towards machines that can learn and adapt as humans do; the stuff of both scientific aspirations and science fiction alike.

This term is quite broad, and such broadness is not useful in analysis. We shall therefore look to the term *machine learning*, which relies on computational pattern recognitionAnzai [9], and the algorithmic creation of pattern recognition systems ('models') that, having observed patterns in a training dataset, are not fit to observe said pattern elsewhere.

Machine learning stands in contrast to the rule-based expert system which once led the wave of AI. The former learns from data; the latter was a body of human knowledge, represented

[1] One of the co-authors of this paper is a co-founder of said service.

[2] https://statmodeling.stat.columbia.edu/2019/01/28/bullshit-asymmetry-principle/

mainly as if-then rules, with every step of every algorithm designed by hand, with clear, user-defined, and finite steps. (Liao, 2005) [10]). Early, impressive AI efforts, such as ELIZA, a pscyhologist program which was one of the first to attempt the Turing Test, was of the latter class.(Weizenbaum, 1966 [11]). However, this mode of programming rulesets by hand had serious limitations, and was supplanted by machine learning.

A strong subfield within machine learning, as a field, is natural language processing; and within this exists the dedicated subfield of content classification. Of late, it has risen in seeming reply to the rise in interest around misinformation.

Promising results have been obtained from building large collections, or corpora, of annotated text - words, sentences, paragraphs or articles. This text is generally labelled by humans to signify something concerning the truths presented therein: usually overtly labelling texts as 'true' or 'false', and sometimes labelling them with more nuanced classifications, like whether or not the headline agrees with the body text.

Machine learning algorithms that do well at classification are then let train on this text. This training process creates models that can then mimic the kind of classification performed by humans. Such a model would, in theory, able to process a significantly higher workload than a human, and could be a highly scalable tool in the factchecker arsenal.

Many sophisticated efforts have been made in pursuit of this goal. To examine this in detail, it is useful to divide machine learning again into two categories:

1) Classical machine learning

2) Deep learning

Classical machine learning is a set of approaches that uses well-defined algorithms that can parse data and perform some sort of analysis based on the algorithm and the data in question. This space that includes Bayesian approaches, decision trees, inductive logic programming, clustering, and model-free reinforcement learning.

Deep learning differs in that multiple layers of such smaller algorithms are clustered in layers. Each can perform some analysis on the data it receives, and, instead of passing the output directly to the user, can pass it amongst themselves, depending on the configuration. This is called a neural network. The total structure as a whole is therefore capable of more complex analysis than the single-algorithm model, but is more complex to build and more difficult to interpret.

This division allows us, as practitioners, to loosely cluster efforts by computational effort and complexity. Today, much of what we call AI falls into one of these two camps. For the subject of factchecking, we therefor have sophisticated pattern recognition for text; as yet we have not reached that nonhuman *intelligence* that the term 'AI' promises.

Classical machine learning

Three of the oldest and most established algorithms in the classical machine learning space are Logistic Regression, the naive Bayes algorithm (NB) (Rish, 2001 [12]) and Support Vector Machines (SVM) (Hearst et al., 1998 [13]). Early relevant literature shows us the roots of the

field: in classifying spam - especially in product reviews intended to mislead. (Jindal & Liu, 2008 [14]) concerns itself with study of opinion spam - specifically, of 5.8 million reviews and 2.14 million reviewers on Amazon.com, and showcases the effectiveness of Logistic Regression. Mihalcea & Strapparava(2009) [15] move more definitely towards misinformation: in their paper, which trials naive Bayes and Support Vector Machines on three data sets to detect falsehoods, they are able to state that "*Very little work, if any, has been carried out on the automatic detection of deceptive language in written text. Most of the previous work has focused on the psychological or social aspects of lying, and there are only a few previous studies that have considered the linguistic aspects of falsehood.*" While their classifiers are not particularly impressive by today's standards - in some cases, barely better than a coin flip - they definitively add to automated the detection of falsehoods.

Ott et al. (2011) [16] build upon this work, again sticking to the theme of opinion spam in reviews. While using a dataset that would be considered small by today's standards - 800 reviews, their contribution is in showing that Naive Bayes and SVMs significantly outperformed their human benchmarks, in some cases by almost 30%. Feng et al. (2012) [17], applying SVMs across multiple datasets, demonstrated both high classification accuracy (above 85%); a 2015 survey of the field by Conroy et al. (2015) [18] cite SVMs and Naive Bayes as being the state-of-the-art of the day.

These three algorithms consistently show up years later. For instance, Rubin et al. (2016) [19] used 360 Canadian and American satirical and legitimate news articles to trial an implementation using SVMs. Granik & Mesyura (2017) [20] utilized a simple Naive Bayes classifier to detect fake news based on specific words used in the text, based on 1,771 articles marked as true and false and no factual content categories. Ahmed et al. (2017) [21] trialled different algorithms, and concluded that a variant of SVMs, fed term frequencies - a representation of unigrams in a text and the number of times each appears - performed the best.

Meanwhile, the decision tree approach began to show up alongside SVMs and naive Bayes. Castillo et al.(2011) [22], in using a decision tree-based approach to assessing information credibility in tweets, compared their work with naive Bayes and SVMs and made a case for trees. Kwon et al.(2013) [23], studying rumor detection on Twitter, put random forests - a particularly high-performing type of decision tree - side-by-side with naive Bayes and SVMs; their 87% accuracy rate for random forests came within 2% of the high water mark set by SVMs in their test. By 2017, random forests had become part and parcel of the toolkit, as seen in Potthat(2017) [24] stylometric inquiries into hyperpartisan news, or Buntain & Golbeck (2017) [25]'s work on fake news on Twitter.

The work of Chen et al.(2016) [26] created a technically superior version decision-tree approach that performed faster and better: the so-called eXtreme Gradient Boosting algorithm, or XGBoost. Derivatives of XGBoost exist - notably LightBGM and CatBoost(Daoud, 2019) [27] - operating on roughly the same principles. This algorithm started appearing alongside those named before - for example, as shown in Facebook-related work of Janze & Risius (2017) [28], which showed that the usual algorithms could be used to build highly performant classifiers on social media data, and made a case for platform operators to perform this kind of checking. By 2017, Shu et al. [29], in their widely cited survey of the field, were able to describe the general shape of this branch of research: "*most previously mentioned approaches focus on extracting various features, incorporating these features into supervised classification models, such as naive Bayes, Decision Trees, Logistic Regression, K-Nearest Neighbor (KNN), Support Vector Machines (SVM).*"

A year later, Helmstetter & Paulheim (2018) [30], operating on relatively large dataset of 401,414 tweets, applied almost the same series of machine learning algorithms - Support Vector Machines, Naive Bayes, Neural Networks, Random Forests and XGBoost. Shu et al. (2019) [31], in demonstrating a richer way of embedding social context into data for better classification, chose a similar palette for their testing - Logistic Regression, Naive Bayes, Decision Trees, Random Forests, XGBoost and AdaBoost. So did Gravanis et al. (2019) [32] when benchmarking algorithms for general-purpose misinformation classification work. Abonizio et al. (2020) [33] did a similar experiment with 9,930 news articles with American English, Brazilian Portuguese, and Spanish, and highlighted the effectiveness of SVM, Random Forest (RF), k-Nearest Neighbors (k-NN), and XGBoost. These same algorithms are used in misinformation detection related to COVID-19 (Felber, 2021) [34].

This gives us today's state of the art, where there seems to be broad consensus on common algorithms to deploy. The reason for their persistence may lie in their reliability and ease of use. If we compare with Rubin's human average (a mere 54%), the classical machine learning stack does far better: Felber's 2021 study reported that the LR, NB and SVM approaches were over 93% accurate in identifying fake news related to COVID-19. Kwon et al were not too far behind in 2013. These algorithms generally perform well in almost any case given to them.

Moreover, tooling commonly used for these algorithms - especially programming languages - have come so far in their ease of use that it is easy to deploy any of these algorithms in a '*theory-free*' manner - i.e. without necessarily knowing the minutae of how each algorithm performs. While deep learning approaches required some understanding of layers, neurons and concepts like backpropagation in order to fine-tune, the relative simplicity of the classical machine learning stack has made it easy to treat these algorithms as fairly reliable black boxes.

Deep learning

Beyond the complexity of the classical machine learning algorithms lies deep learning, in which neural networks of different architectures are brought to bear. Deep learning approaches typically require and perform best on large datasets, and thus an observer may draw parallels between the increasing availability of attention, public data and rise in methodological complexity.

Firstly, neural networks are not new to this field. Zhou et al's 2004 paper [35] shows neural networks outperforming other types of analysis, albeit with results in the 61.5% - 79.2% range depending on the dataset.

However, it was with the work of research teams led by Geoffrey Hinton, Yann LeCun and Yoshua Bengio (sometimes referred to as the Godfathers of AI [36]) that neural networks gained prominence as a potentially superior solution; this was the rise of deep learning. It combined with the burgeoning availability of increasingly more powerful processing via the utilization of GPUs, or graphical processing units; and the spread of software libraries such as Keras, Torch and Tensorflow that, in turn, enabled more researchers to engage in the kind of sophisticated models that deep learning required.

Neural networks are not a monolith, but divide themselves into architectures: specific arrangements of the algorithms (neurons) and layers that prove exceptional at one task of the other. A

number of these architectures have proved especially popular in the kind of classification task we are interested in. Recurrent Neural Networks (RNNs)(Mikolov et al., 2010 [37]) showed up early in text-related literature. Long Short-Term Memory architectures (LSTMs)(Greff et al., 2016 [38]), a variant of RNNs, became popular; LTSMs implementations appear in key work from Long (2017) and Rashkin et al. (2018), who posit that those implementations outperform prior work. As did Convolutional Neural Networks (CNNs); CNNs built on the work of seminal work of Kim (2014) [39] show up in the work of Wang (2017) [40], shown to be superior to implementations of LSTMs, SVMs and logistic regression (albeit by small margins).

A minor arms race of sorts has happened here; we see points for one and counterpoints for the other across these options. As is with neural networks, tuning of parameters and the fusion of different architectures make a difference in absolute accuracy gains - even if the increment is marginal. Arguably, Bhattarcharjee Balantrapu (2017) [41]'s approach of reducing class labels yield enormous dividends, taking CNN accuracy well above the 96% mark previously established on their choice of benchmark (Oshikawa et al., 2018) [42]. FNDNet, another CNN implementation, by Kaliyar et al. (2020) [43], demonstrated accuracy above the 98% mark. Recent work by Glenski et al. (2021) [44] benchmark these different approaches with LSTMs in a multi-domain, multi-language study, examining different ways that the input data can be processed before being passed into models and the performance impact thereof.

Meanwhile, the complexity of deep learning approaches leave space for variation in input and in architectures. Kaliyar et al, above, succeeded by increasing the 'depth' of their model; Bhattarcharjee Balantrapu used data reduction as well as a hybrid architecture of a shallow feature based classifier and a deep classifier working in tandem; on the other end of the spectrum, Singhania et at. (2017) [45]'s 3HAN was built on the three-layer heirarchical attention network architecture pioneered by Yang et al. (2016) [46]; 3HAN boasted above 96% accuracy in fake news detection, and made a powerful use case for a different architecture. Work by Vijjali (2020) [47] and (Gundapu & Mamid, 2021) [48] have even led to different approaches involving transformer architectures and large, pre-trained models.

Their versatility also lets them operate fairly accurate in domains beyond mere text. Ruchanksy et al. (2017) [49] examined RNNs for the capture of both article text and user interactions with it; TI-CNNs (Yang et al., 2018) [50] have demonstrated the ability to work with both text and images.

However, as with classical machine learning, we come to a sense of a 'state of the art' - albeit a lot more loosely than the previous category. CNNs, LSTMs, split between text and combinations of text and imagery, with heavy tweaking of input data features and layers for that last few percentage in optimized accuracy. A few approaches like HANs and transformers await on the sidelines. Most, if not all recent work performs above the classical machine learning stack.

There are epistemological boundaries: these techniques do not fact-check as a human agent does, but rely on linguistic features - such as the co-occurence of words and their relation to each other. Anecdotally, this is one of the strongest offhand reasons for dismissing automated, corpus-based AI methods, since the process of search, triangulation, and journalism that human factcheckers go through simply does not happen here. Any patterns not visible to the algorithms from the training corpus would be increasingly difficult to classify, and therefore, as public discourse and misinformation trends change, these tools become obsolete unless retrained or remade with new data. Furthermore, an astute observer may point out that satire may be impossible to interpret using these methods.

It it an intellectual fallacy to say that half of something is better than nothing. After all, half a kitten is not half the amount of fun and joy, but instead a messy butchered corpse that nobody wants to touch. And yet, given the inherent asymmetry of effort in the task at hand, and given Rubin's dismal human average (54% accuracy), these figures are extremely promising, even with their limitations. Which then behooves us to ask: what other limitations exist?

Given enough data . . .

An important caveat to the above is the term 'given enough data'. Much of the research we have cited is done with English data. Likewise, the experiment shows how easy it might be - in English, where tens of thousands of articles lie labelled and ready to download.

What of other languages? Glenski et al, who controlled for methods while testing across multiple languages, reported two significant findings. Firstly, a general model, attempting to work across multiple languages is outperformed by language-specific models: in this case English, Russian, German and Spanish. However, a review of the data underlying the research we have talked about makes the problem plain:

Dataset name or source	**Language**	**Access**
LIAR	English	downloadable via github repository
Kaggle Fake News dataset	English	downloadable via Kaggle
Harvard FakeNewsNet data	English	downloadable via github repository
Fake News Corpus	English	downloadable via github repository
KDNuggets Fake News dataset	English	obtainable by contacting the author
UNBiased dataset	English	obtainable by contacting the author
Kaggle-EXT	English	obtainable by contacting the author
Weakly Supervised Learning for Fake News Detection on Twitter	English	obtainable by contacting the author
Fake.Br Corpus	Brazilian Portuguese	downloadable via github repository
RumourEval	English	downloadable via github repository
WSDM 2019 challenge dataset	Mandarin Chinese	downloadable via kaggle
Ma-Weibo	English?	obtainable by contacting the author
Twitter15	English	obtainable by contacting the author
Twitter16	English	obtainable by contacting the author

Dataset name or source	Language	Access
Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate	English	obtainable by contacting the author
Fake News Challenge (FNC-1) Data	English	downloadable via github repository
A multi-layer approach to disinformation detection on Twitter	English	obtainable by contacting the author
A multi-layer approach to disinformation detection on Twitter	Italian	obtainable by contacting the author
Horne and Adali(2017)	English	obtainable by contacting the author
FakeNewsCorpusSpanish corpus	Spanish	downloadable via github repository
Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking	English	downloadable via link provided in the paper
FEVER dataset	English	downloadable via github repository
Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter	English	downloadable via link provided in the paper
3HAN: A Deep Neural Network for Fake News Detection	English	obtainable by contacting the author
Buzzfeed election data set	English	downloadable via github repository
Burfoot and Baldwin(2009) data set	English	obtainable by contacting the author
PoliticalNews	English	obtainable by contacting the author
CREDBANK dataset	English	downloadable via github repository
NELA-GT-2018	English	downloadable via link provided in the paper
Some Like it Hoax:Automated Fake News Detection in Social Networks	English	obtainable by contacting the author
ISOT Fake News Dataset	English	Downloadable from the website link of University of Victoria

The overwhelming majority - especially those readily downloadable - are in English. Of other major languages there is barely a peep. Indeed, as pointed out in Wijeratne et al., (2019) [51], this problematic situation holds true for natural language processing in most of the world's languages. Because of the inherent structural differences between languages, especially those more distant in lineage to each other, algorithms that we take as par for the course perform differently; and because of the differences in data availability, most languages are far behind English when it comes to language processing.

What is to be done?

What is to be done? Firstly, as we have shown, the state of the art in English is quite advanced; English is a lingua franca of the modern world, and thus useful in many contexts. Furthermore, the kind of results we have surveyed in the literature and demonstrated in practice no longer require academic teams and complex hardware: commercial desktop computer hardware suffices. Given this ease, we suggest exploring other failure points in the adoption of these AI for dealing with misinformation: it could be anything from lack of user-friendly interfaces to information asymmetry not filtering these advances down to those who can use them. Once these issues are known, they can be rectified.

Secondly, the work of Glenski et al. shows that while we may not be able to make precise claims about the efficacy of machine learning in other languages, we may be reasonably confident that key algorithms from the domain will be within a few degrees of accuracy; an observation supported by Abonizio et al.(2020) [33], who, working with news articles in English, Portuguese and Spanish, highlighted the effectiveness of SVMs, RF (RF), and XGB.

However, for language-specific work, the fact remains that datasets must be built, these experiments must be trialed; only then can we make blanket claims about technological difficulty or ease. Until then, we exist in a fundamentally unequal state: for languages where the data is available, harnessing machine learning models to support factcheckers is - at least on a technical level - trivial. Datasets like FNC500k, Kaggle and KDNuggets are a few clicks away.

For those languages without the data, however, it is simply not possible. As once noted by William Gibson, the future is already here - it's just not evenly distributed.

Acknowledgements

This research has been made possible through a grant from the Asia Foundation.

Bibliography

[1] D. Dor, "The role of the lie in the evolution of human language," *Language Sciences*, vol. 63, pp. 44–59, 2017.

[2] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 554–559, 2016.

[3] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[4] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[5] R. J. Limaye, M. Sauer, J. Ali, J. Bernstein, B. Wahl, A. Barnhill, and A. Labrique, "Building trust while influencing online covid-19 content in the social media world," *The Lancet Digital Health*, vol. 2, no. 6, pp. e277–e278, 2020.

[6] K. Roose, "What is qanon, the viral pro-trump conspiracy theory?" *The New York Times*, 2020.

[7] J. Tollefson, "Tracking qanon: how trump turned conspiracy-theory research upside down." *Nature*, 2021.

[8] V. L. Rubin, "On deception and deception detection: Content analysis of computer-mediated stated beliefs," *Proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1–10, 2010.

[9] Y. Anzai, *Pattern recognition and machine learning*. Elsevier, 2012.

[10] S.-H. Liao, "Expert system methodologies and applications—a decade review from 1995 to 2004," *Expert systems with applications*, vol. 28, no. 1, pp. 93–103, 2005.

[11] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[12] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[13] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[14] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 international conference on web search and data mining*, 2008, pp. 219–230.

[15] R. Mihalcea and C. Strapparava, "The lie detector: Explorations in the automatic recognition of deceptive language," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009, pp. 309–312.

[16] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," *arXiv preprint arXiv:1107.4557*, 2011.

[17] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 171–175.

[18] N. K. Conroy, V. L. Rubin, and Y. 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.

[19] V. L. 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*, 2016, pp. 7–17.

[20] M. Granik and V. Mesyura, "Fake news detection using naive bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. IEEE, 2017, pp. 900–903.

[21] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*. Springer, 2017, pp. 127–138.

[22] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.

[23] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *2013 IEEE 13th international conference on data mining*. IEEE, 2013, pp. 1103–1108.

[24] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," *arXiv preprint arXiv:1702.05638*, 2017.

[25] C. Buntain and J. Golbeck, "Automatically identifying fake news in popular twitter threads," in *2017 IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 2017, pp. 208–215.

[26] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[27] E. Al Daoud, "Comparison between xgboost, lightgbm and catboost using a home credit dataset," *International Journal of Computer and Information Engineering*, vol. 13, no. 1, pp. 6–10, 2019.

[28] C. Janze and M. Risius, "Automatic detection of fake news on social media platforms." *Pacis*, vol. 261, 2017.

[29] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[30] S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news detection on twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 274–277.

[31] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 312–320.

[32] G. Gravanis, A. Vakali, K. Diamantaras, and P. Karadais, "Behind the cues: A benchmarking study for fake news detection," *Expert Systems with Applications*, vol. 128, pp. 201–213, 2019.

[33] H. Q. Abonizio, J. I. de Morais, G. M. Tavares, and S. Barbon Junior, "Language-independent fake news detection: English, portuguese, and spanish mutual features," *Future Internet*, vol. 12, no. 5, p. 87, 2020.

[34] T. Felber, "Constraint 2021: Machine learning models for covid-19 fake news detection shared task," *arXiv preprint arXiv:2101.03717*, 2021.

[35] L. Zhou, J. K. Burgoon, D. P. Twitchell, T. Qin, and J. F. Nunamaker Jr, "A comparison of classification methods for predicting deception in computer-mediated communication," *Journal of Management Information Systems*, vol. 20, no. 4, pp. 139–166, 2004.

[36] J. Vincent, "'godfathers of ai' honored with turing award, the nobel prize of computing," Mar 2019. [Online]. Available: http://www.theverge.com/2019/3/27/18280665/ai-godfathers-turing-award-2018-yoshua-bengio-geoffrey-hinton-yann-lecun

[37] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

[38] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[39] Y. Kim, "Convolutional neural networks for sentence classification," 2014.

[40] W. Y. Wang, "" liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648*, 2017.

[41] S. D. Bhattacharjee, A. Talukder, and B. V. Balantrapu, "Active learning based news veracity detection with feature weighting and deep-shallow fusion," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 556–565.

[42] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," 2020.

[43] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "Fndnet–a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, pp. 32–44, 2020.

[44] M. Glenski, E. Ayton, R. Cosbey, D. Arendt, and S. Volkova, "Towards trustworthy deception detection: Benchmarking model robustness across domains, modalities, and languages," *arXiv preprint arXiv:2104.11761*, 2021.

[45] S. Singhania, N. Fernandez, and S. Rao, "3han: A deep neural network for fake news detection," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 572–581.

[46] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1480–1489. [Online]. Available: https://www.aclweb.org/anthology/N16-1174

[47] R. Vijjali, P. Potluri, S. Kumar, and S. Teki, "Two stage transformer model for COVID-19 fake news detection and fact checking," in *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL), Dec. 2020, pp. 1–10. [Online]. Available: https://www.aclweb.org/anthology/2020.nlp4if-1.1

[48] S. Gundapu and R. Mamidi, "Transformer based automatic covid-19 fake news detection system," 2021.

[49] N. Ruchansky, S. Seo, and Y. Liu, "Csi," *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Nov 2017. [Online]. Available: http://dx.doi.org/10.1145/3132847.3132877

[50] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, "Ti-cnn: Convolutional neural networks for fake news detection," *arXiv preprint arXiv:1806.00749*, 2018.

[51] Y. Wijeratne, N. de Silva, and Y. Shanmugarajah, "Natural language processing for government: Problems and potential," *International Development Research Centre (Canada)*, 2019.