Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator

Terry Traylor
U.S. Marine Corps
Fargo, ND
terry.o.traylor4.mil@mail.mil

Jeremy Straub, Gurmeet, Nicholas Snell
Department of Computer Science
North Dakota State University
Fargo, ND
jeremy.straub@ndsu.edu

Abstract—Intentionally deceptive content presented under the guise of legitimate journalism is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most so-called 'fake news' is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as making excessive use of unsubstantiated hyperbole and non-attributed quoted content. In this paper, the results of a fake news identification study that documents the performance of a fake news classifier are presented. The Textblob, Natural Language, and SciPy Toolkits were used to develop a novel fake news detector that uses quoted attribution in a Bayesian machine learning system as a key feature to estimate the likelihood that a news article is fake. The resultant process precision is 63.333% effective at assessing the likelihood that an article with quotes is fake. This process is called influence mining and this novel technique is presented as a method that can be used to enable fake news and even propaganda detection. In this paper, the research process, technical analysis, technical linguistics work, and classifier performance and results are presented. The paper concludes with a discussion of how the current system will evolve into an influence mining system.

Keywords-component; Fake News, Machine Learning, Natural Language Processing, Attribution Classification, Influence Mining

I. INTRODUCTION

Intentionally deceptive content presented under the guise of legitimate journalism (or 'fake news,' as it is commonly known) is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most fake news is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as excessive use of unsubstantiated hyperbole and non-attributed quoted content. The results of a fake news identification study that documents the performance of a fake news classifier are presented and discussed in this paper.

II. BACKGROUND AND RELATED WORK

Fake news has been demonstrated to be problematic in multiple ways. It has been shown to have real influence on public perception [1]–[3] and the ability to shape regional and national dialogue [4]. It has harmed businesses [5] and individuals and even resulted in death, when an individual responded to a hoax [6]. It has caused some teenagers to reject the concept of media objectivity [7] and many students can't reliably tell the difference between real and faked articles [8]. It is even thought to have influenced the 2016 United States elections [9].

Fake news can be spread deliberately by humans or indiscriminately by bot armies [10], with the latter giving a nefarious article significant reach. Not just articles are faked, in many cases fake, mislabeled or deceptive images are also used to maximize impact [11]. Some contend that fake news is a "plague" on society's digital infrastructure [12]. Many are working to combat it. Farajtabar, et al. [13], for example, has proposed a system based on points, while Haigh, Haigh and Kozak [14] have suggested the use of "peer-to-peer counter propaganda."

The work presented herein builds on prior work in several areas. This section continues with a discussion of the characteristics of fake news. Then, prior fake news detection efforts are reviewed. Finally, fake news as a communication phenomena (including attribution considerations) is discussed.

A. Characteristics of Fake News

Fake news has been shown to be detectable in several ways. Obviously, fact checking is one way to identify and debunk fake news; however, this is slow and difficult to automate. Batchelor [15] has suggested tasking libraries to help with this task. Automated detection, however, can occur at or near the speed of transmission, limiting the level of human involvement in certain areas of operations. Fake news has also been shown to differ from legitimate journalism in structural and other ways. Horne and Adali [16] note that fake and legitimate news differ in title length and the simplicity and repetitiveness of body text. Rubin,

978-1-5386-6783-5/19/$31.00 ©2019 IEEE
DOI 10.1109/ICSC.2019.00086

et al. [17] propose the analysis of satirical cues, while Volkova, et al. [18] proposes the use of linguistic models.

B. Automated Fake News Detection

Given the problem posed by fake news, a variety of approaches have been proposed to detect it automatically [19]. Chen, et al. [20] make the case for the need for automated detection, for speed and convenience, among other prospective needs that would be met. Unlike crowdsourcing [21] and using human employees for review, automation can result in near-instant decisions and provides requisite scalability. Riedel, et al. [22], for example, proposed a headline stance based detection technique. Rashkin, et al. [23] use a language analysis-based approach, while Jin, et al. [24] propose a "hierarchical propagation" approach and Shu, et al. [25] use a datamining process.

Tools to automate this process are also being developed which will actualize the research performed into a tangible technology. Saez-Trumper [26], for example, has developed a tool to help identify users who promote fake news on Twitter. Jin, et al. [24] have developed a "Hierarchical Propagation" approach to content credibility evaluation.

C. Attribution and Modern Communications

Multiple research teams have developed systems that use natural language routines to identify quotes and associated attributions. Pareti, et al. [27] and O'Keefe, et al. [28] developed machine learning classifiers that correctly identified direct and indirect quotes using machine learning methods. Muzny, et al. [29] also developed a method for identifying quote attributions using a multi-stage lexical sieve system. While there are multiple other studies that hand craft attribution detection systems, the Pareti and Munzy approaches to attribution will be integrated, herein, in order to develop a simple direct quote attribution system.

III. Methodology

The methodologies used to research the fake news phenomena, develop the research database, and evolve the qualitative model into a quantitative model are reviewed in this section.

A. Grounded Work and Theory Development

The research team implemented a mixed-methods approach to study fake news documents, develop a qualitative model for testing, and transform the qualitative construct into a quantitative system. Initial fake news observations and hand-crafted pattern analysis was performed using Glaser and Strauss's Grounded Theory [30] methods for theory building and coding. Grounded Theory is an inductive-based social-science research technique that is used to build theories and frameworks from existing data. When researchers use Grounded Theory to construct an understanding of a phenomena under study, the research team starts by observing the data and looking for patterns, trends, and differences. The trends and patterns that emerge from the analysis are grouped into codes and themes. Over time, the codes and themes become categories and form the basis for a new theory. As a hypothetical example, if one were to notice that all fake news documents started with the phase, 'trust me I am not lying to you,' the researchers observing this would eventually group enough data documenting this trend and form a hypothesis that all fake news documents start with that phrase. The emerged hypothesis would eventually become a rule to be tested. Grounded Theory was selected for use to facilitate building a theory inductively based on the data available.

The results of the initial qualitative work unearthed technical linguistic patterns unique to the fake news documents that were reviewed. The linguistic patterns were used to develop a machine learning grammar and hypothesis.

B. Fake News Identification Corpus

A new fake news identification corpus was developed in order to study fake news technical linguistic patterns and enable theory testing using a locally generated dataset. The research team constructed and validated the version of the corpus used for this work over a 7-month period.

At the time it was used for this work, the corpus contained 218 documents from over 40 different online sources. It contains validated fake and real news documents with assertion, belief, and fact quotations. The corpus was built by a research team with 10 different researchers. Document accuracy (whether or not a document was considered false content or fake news) was reviewed weekly by the research team and evaluated by other researchers on the team for corpus inclusion and acceptance. In short, each document that was added to the corpus was reviewed and accepted by multiple researchers before the document was added to the corpus for future use.

At the time this work was conducted, the corpus contained 421 quotations from documents that the research team classified as either real and fake media documents. While the corpus wasn't originally designed for quote attribution machine learning research, it includes all text inside a document and thus includes quotations. Each document in the corpus is subdivided into header and body parts. Work on a more robust corpus that can be publicly shared is the subject of a future publication.

C. Machine Learning Grammar Development

Machine learning grammars were built inductively and iteratively as technical linguistic patterns emerged from the Grounded Theory research approach. The emerged grammars became the basis for hypothesis development and experimentation.

IV. Emerged Tool Design

Based on the research team's initial Grounded Theory work and using the corpus, several key technical linguistic patterns were identified in the fake news documents that were used to develop a classifier model with supporting grammars. These grammars became the base feature extractors for the custom classifier.

A. Attribution and Key Fake News Features

Fake news documents exist across all forms of media and are particularly ubiquitous on social media platforms. At the

beginning of the research project, 30 false content and 30 real content documents were reviewed to develop an understanding of the phenomena and to begin building theory. Of the 60 documents reviewed, it was identified that the preponderance of the false content documents (28 of the 30 reviewed documents) that included quotes either lacked proper attribution or attributed quotes to non-named entities to assert a fact. While trends continue to be identified across the false content, the most dominant initial false content indicator was the lack of proper attribution. Attribution in the documents that were reviewed and classified as real news documents normally occurred within less than 50 character spaces from the beginning or ending of a direct quote.

With these trends and observations in mind, a custom classifier that used attribution as a sole fake news indicator was constructed. The system (described in section VI) measures the amount of attribution inside a document and based on a definable attribution tolerance, labels the document as either real or fake.

B. Custom Attribution Feature Extraction Classifier

The attribution classifier work and definitions originally put forth by multiple researchers were extended to build the attribution classifier and resultant one-feature fake news identification system. Specifically, the definitions and constructs originally proffered by Pareti, et al. [27] and by O'Keefe, et al. [28] were augmented. Both prior research teams defined attribution as a linguistic convention where a verb or attribution cue links a source to a quoted piece of text called content. Specifically, an attribution for a quote has a source span, cue span, and content span as defined in Table 1.

TABLE I. PARETI, ET AL. [27] ATTRIBUTION MODEL DEFINITIONS

	Definition
Source	The span of text that includes who put forth the quote or who the content is attributed to.
Cue	A verb or verb phrase that lexically links the source to the quote or content.
Content	The span of text that serves as the quote and is attributed.

To build the custom attribution machine learning classifier, attribution construct aspects from Muzny, et al's [29] work were also implemented. Muzny, et al.'s quote → mention, mention→ quote, and mention → entity linking attribution constructs helped extend Pareti, et al.'s definitions to build a simple technical attribution classifier, as depicted in Figure 1.

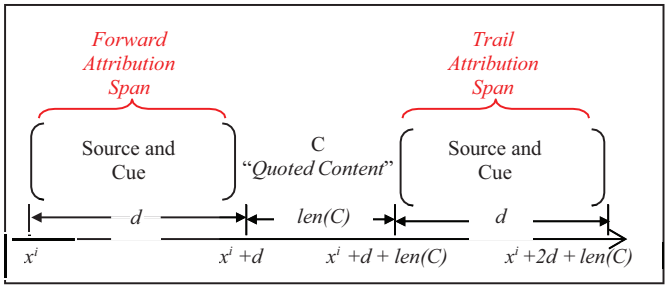

Fig. 1. Proposed extensions to the Pareti, et al. [27] definitions. Including the attribution span, attribution span absolute distance 'd', and length of the quote enables simple quote attribution searching for classification.

An attribution to a quote is defined using the following definition, to build the custom feature extractor: Let C be any content span of random span length len(C) for a quote that requires attribution. The attribution span is the absolute distance "d" in character spaces from the beginning or end of a content span marked by double quotes. So for any properly attributed quote:

$$(Source, Cue) \leq x^i + len(C) + 2d$$

$$\exists (Source, Cue) for\ C\ s.t. \{ \quad or \quad (Source, Cue) \geq x^i \tag{1}$$

The attribution span is divided into two searchable sub-spans called the forward and trail attribution spans. The classifier tool was built to search inside the forward and trail attribution space and to classify the quote as either attributed or not.

The resulting binary classification label is based on the presence of learned source and cue information inside the attribution spans. To identify a source, the custom classifier searched for named-entities or persons or organizations that could be attributed as having made a quote using named-entity recognition methods. Cue identification is based on learning associated cueing verbs or cue information contained inside the training set. Most informative cue words or phrases will be added to a living attribution "bag of words" model. Attribution feature extraction comes from applying machine learning algorithms to the forward and trailing attribution spans.

C. The Resultant Fake News Detection Pseudocode and Tool

The fake news detection tool uses the results of the outputs from the attribution classifications to assign a final label for the entire document. A simple scoring system, described in the next sub-section, was used to construct a final attribution score (called the attribution score or A-score) and assign a fake versus real classification label for every document containing quotes.

D. Fake News Detection Algorithm

The fake news detection algorithm is as follows. For each document in the document collection, the document's paragraphs are counted and tokenized. Each paragraph is also checked for quotes. If a paragraph has quotes, then these are processed using the custom attribution classifier (which uses the A-score algorithm). Positive attributions receive a +1 score and negative attribution classifications receive a -1 score. If the overall A-score (the sum of positives and negatives) is greater than or equal to 0, then the document is assigned a label of real. If the A-score is less than 0 then the document is assigned a label of fake. Note that the A-score threshold is, thus, a key area of potential configuration for this algorithm.

The A-score algorithm is used to label quotes as either real or fake based on the results of the machine learning classification. The pseudo code for both the algorithms is presented in Figure 2.

V. EXPERIMENTAL DESIGN

To test the fake news detection system, the corpus was devided into training (60% of the available data), development (10 % of the available data), and test (30% of the available data) sets. The training process tranis the algorithm to recognize

Algorithm: Fake News Detector with Attribution Key Extraction Feature

Algorithm: FakeRank Fake News Detector Main

```
1  For each doc in Documents
2  pCount ←Σ(number of paragraphs)
3  Tokenize doc by paragraph
4  FakeRank ← 0
5  For each paragraph in doc
6    qScore ← Σ (number of quotes)
7    if qScore > 0 then
8      A-score ← 0
9      For each quoteset
10       quote_cl classify.naivebayes(quoteset_attribution_space, d)
11       If quote_cl
12           A-score ← A-score +1
13       Else
14       A-score ← A-score -1
15       return A-score
16    FakeRank ← FakeRank + A-score
17 If FakeRank >=0 then
18    docLabel ← real
19 If FakeRank < 0 then
20 docLabel ← fake
```

Algorithm: A-Score Fake News Detector Supporting

```
1 //Prepare quoteset_attribution_space
2 capture quotes (d)
3 FWD_A-Span ← quoteset_attribtuion_space-d
4 TRL_A-Space ← quotest_attribution_space_d
5 classificationspace ← [ FWD ∪ TRL] - stopwords
6   attribution_label ← classify.naivebayes(classificationspace,extractor
        = custom
7   return attribution_label
```

Fig 2. Pseudocode for proposed algorithms.

words inside the attribution space associated with real or fake news items. Pre-training data preparation focuses on removing extraneous common words to prevent these words from influencing association scores. Because common word data is not necessary for presentation to the classifier, no additional data preparation is done to the corpus for testing. As can be seen in the pseudo code and algorithm description, it was possible to tune the attribution span during testing, but it was decided to perform a simple run with the attribution space at d=45 for simplicity. Three types of experimental validations were conducted during experimentation. The accuracy of the quote attribution identifier, the custom quote attribution classifier, and finally the overall performance of the one feature fake news detection tool were tested.

VI. Initial Results and Analysis

This section summarizes the results of the three tests conducted to support the fake news detection tool.

A. Quote Attribution Identifier

The simple Python and Textblob system used to identify quotes in a paragraph worked well. Because the tool looks for the presence of double quotes inside strings, quote or content identification was simple. The tool identified 96% of the quotes in the training set. We assess that the tool was confused, in limited instances, by complex and malformed quotations.

B. Quote Attribution Classifier

The quote attribution classifier, which is the core of the system, functioned well, but did not perform at an acceptable level. Several runs were needed to properly calibrate the classifier and account for linguistic processing issues such as quotes inside quotes and single quotes inside double quotes. The classifier also had issues handling multiple quotes within short proximity to each other inside a text. For example, if one quote came right behind another quote and the source and cue data were within the joint attribution space (in front of or behind a quote) for both quotes, the system encountered challenges processing both quotes. The final overall Classifier Accuracy is 0.69 and the overall Classifier Error is 0.31. Additional classifier performance metrics are presented in Table II. While these numbers are subpar for attribution classification work, research is ongoing to improve the performance of the classifier. Tuning the attribution distance and potentially developing a fake news attribution dictionary are methods being used to improve the classifier performance.

TABLE II. Classifier Performance

	Fake	Real
True Positive	0.55	0.88
False Positive	0.13	0.45
True Negative	0.88	0.55
False Negative	0.45	0.13
Precision	0.85	0.61
F-score	0.66	0.72

The metrics in Table II are defined as follows. The true and false positive and negative rates are the number correctly (true) or incorrectly (false) identified over the total number identified with the relevant classification (positive or negative). For example, the true positive rate is the number of true positives divided by the number of true positives and false negatives. The precision is the number of correctly labeled items divided by the total number of elements belonging to the positive class. The total number of elements in the positive class includes both the true positives or correctly labeled items and the false positives or incorrectly labeled items. The F-score is a metric used in binary classification problems that measures the accuracy of a test. The F-score combines the precision and recall (or true positive rate) for a binary classification problem and is the harmonic mean of the precision and recall.

C. Overall Fake News Detection Tool Performance

The attribution-based fake news detection tool that uses the quote attribution classifier, performed suitably for a detection tool using only one feature extraction to classify a document; however, like the attribution classifier, it did not perform well enough for production use. After training and configuration, the tool correctly identified 69.4% of the fake and real news documents in the test set. Upon review, some of the missed labels were attributable to fake news documents with no quotations, fake news documents with attributed quotes of inaccurate statements, and fake news documents that quoted or cited other fake news documents. While, the overall performance results for this system are not as strong as desired, the initial performance is generally encouraging, because fake news is designed to deceive human targets, so an initial classification tool with only one extraction feature seems to perform well, given the complexity of the topic and the aims of the project.

VII. Conclusions and Future Work

This paper presented the results of a study that produced a limited fake news detection system. The work presented herein is novel in this topic domain in that it demonstrates the results of a full-spectrum research project that started with qualitative observations and resulted in a working quantitative model. The work presented in this paper is also promising, because it demonstrates a relatively effective level of machine learning classification for large fake news documents with only one extraction feature. Finally, additional research and work to identify and build additional fake news classification grammars is ongoing and should yield a more refined classification scheme for both fake news and direct quotes.

Future planned research efforts involve combing attribution feature extraction with other factors that emerge from the research to produce tools that not only identify potential false content, but influence based content designed to compel a reader or target audience to make inaccurate or altered decisions.

Acknowledgments

Thanks are given to Alex Thielen, Zak Merrigan, Brian Kalvoda, Riley Abrahamson, Dibyanshu Tibrewell, William Fleck, Ben Bernard, Brandon Stoick and Bonnie Jan who collected and classified the news articles in the database that was used for this work. Thanks are also given to the numerous participants in research efforts related to fake news detection, cybersecurity and information warfare topics at NDSU. The work done in these projects has undoubtably benefited the development of this paper.

References

[1] M. Balmas, "When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism," Communic. Res., vol. 41, no. 3, pp. 430–454, 2014.

[2] C. Silverman and J. Singer-Vine, "Most Americans Who See Fake News Believe It, New Survey Says," BuzzFeed News, 06-Dec-2016.

[3] P. R. Brewer, D. G. Young, and M. Morreale, "The Impact of Real News about '"Fake News"': Intertextual Processes and Political Satire," Int. J. Public Opin. Res., vol. 25, no. 3, 2013.

[4] D. Berkowitz and D. A. Schwartz, "Miley, CNN and The Onion," Journal. Pract., vol. 10, no. 1, pp. 1–17, Jan. 2016.

[5] C. Kang, "Fake News Onslaught Targets Pizzeria as Nest of Child-Trafficking," New York Times, 21-Nov-2016.

[6] C. Kang and A. Goldman, "In Washington Pizzeria Attack, Fake News Brought Real Guns," New York Times, 05-Dec-2016.

[7] R. Marchi, "With Facebook, Blogs, and Fake News, Teens Reject Journalistic "Objectivity"," J. Commun. Inq., vol. 36, no. 3, pp. 246–262, 2012.

[8] C. Domonoske, "Students Have 'Dismaying' Inability o Tell Fake News From Real, Study Finds," Natl. Public Radio Two-w., 2016.

[9] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," J. Econ. Perspect., vol. 31, no. 2, 2017.

[10] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer, "The spread of fake news by social bots."

[11] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy," in WWW 2013 Companion, 2013.

[12] E. Mustafaraj and P. T. Metaxas, "The Fake News Spreading Plague: Was it Preventable?"

[13] M. Farajtabar et al., "Fake News Mitigation via Point Process Based Intervention."

[14] M. Haigh, T. Haigh, and N. I. Kozak, "Stopping Fake News," Journal. Stud., vol. 19, no. 14, pp. 2062–2087, Oct. 2018.

[15] O. Batchelor, "Getting out the truth: the role of libraries in the fight against fake news," Ref. Serv. Rev., vol. 45, no. 2, pp. 143–148, Jun. 2017.

[16] B. D. Horne and S. Adalı, "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News," in NECO Workshop, 2017.

[17] V. L. Rubin, N. J. Conroy, Y. Chen, and S. Cornwell, "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News," in Proceedings of NAACL-HLT 2016, 2016, pp. 7–17.

[18] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, "Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 647–653.

[19] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News," in Proceedings of ASIST, 2015.

[20] Y. Chen, N. J. Conroy, and V. L. Rubin, "News in an Online World: The Need for an "Automatic Crap Detector"," in Proceedings of ASIST 2015, 2015.

[21] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez, "Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation."

[22] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, "A simple but tough-to-beat baseline for the Fake News Challenge stance detection task."

[23] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, and P. G. Allen, "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, 2017, pp. 2931–2937.

[24] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang, "News Credibility Evaluation on Microblog with a Hierarchical Propagation Model," in Proceedings of the IEEE International Conference on Data Mining, 2014.

[25] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective."

[26] D. Saez-Trumper, "Fake Tweet Buster: A Webtool to Identify Users Promoting Fake News on Twitter," in Proceedings of HT'14, 2014.

[27] S. Pareti, T. O'keefe, I. Konstas, J. R. Curran, and I. Koprinska, "Automatically Detecting and Attributing Indirect Quotations," in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 18–21.

[28] T. O'keefe, S. Pareti, J. R. Curran, I. Koprinska, and M. Honnibal, "A Sequence Labelling Approach to Quote Attribution," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 12–14.

[29] G. Muzny, M. Fang, A. X. Chang, and D. Jurafsky, "A Two-stage Sieve Approach for Quote Attribution," in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, vol. 1, pp. 460–470.

[30] B. G. Glaser and A. L. Strauss, The discovery of grounded theory: strategies for qualitative theory. New Brunswick: Aldine, 1967.