Evasive attacks on Fake News Detectors

I. Executive Summary

Machine learning has quietly been playing a pivotal role in our society in everyday decisions that computers are making on behalf of humans. There are multiple touch points with machine learning algorithms including the news and social content we consume each day. News media has traditionally been the window into the world for centuries. The print media ushered for a long time that news is mostly reliable and accurate. However, ever growing pace and sensationalism have been at the forefront pushing reliability and trustworthy news to the rear. Adding to this is the emergence of inadvertent and purposeful deviation of truth in some online news article. The impact is pervasive and has even claimed to have a major impact on national elections, posing a serious problem which merits our utmost attention. With speed, pace and ever growing nature of news, it is challenging for humans and social networks to vet all incoming news. Often times, this news or viral videos reach audiences across the world without any validation or fact checking on the source. To curb the menace of fake news, Social networks, news organizations, search engines try to validate the authenticity of the news articles and to stop the spread of false or fake news to consumers. However, the methods these organizations are employing to identify fake news might be riddled with adversarial evasive attacks. This paper discusses how evasive attacks on machine learning models can be used to disguise fake news as real news and cause even more confusion to misinformed people.

News notifies, influences and impacts our everyday life. A combination of content, source, and social network all play a significant role in humans trusting the news they read or watch. Today, the news is networked, abundant, and fast-flowing through social networks causing immediate repercussion. However, misleading or wrong information have a higher potential to become viral [1] and lead to negative discussions [2] and has a major social impact. Fake news is a hotly discussed topic after the US elections in 2016, which saw the proliferation of inaccurate and misleading news take the limelight. The fake news might originate as a rumor, while others are created for profit such as click-bait ads or few that are purposeful deceitful to sway public opinion on serious issues.

Fake news has been well studied in both, journalism and computer science. Studies have focused on either analyzing the spread of fake news or analyzing the content of fake news. This paper deals with the content of the fake news and how adversarial machine learning attacks can be employed to evade the detection by fake news detectors. The content of a fake and real news articles can be distinguished in a few ways. Fake news articles tend to be shorter in terms of content but use repetitive language and fewer punctuations. Fake news articles differ much more in their titles. Fake titles are longer, use few stop words, and fewer nouns but more proper nouns [3]. Bag of words, Fact-checking, Deep Syntax, Shallow Syntax, Rhetorical Structure and Discourse Analysis are some of the methods used in differentiating fake from real news. Often times, a combination of such methods are used to identify and stop the spread of inaccurate and misleading news.

This paper analyzes the weakness of a couple of text classification methods (Bag of words and Deep Syntax) and proposes exploratory attack on these methods to reduce the effectiveness of algorithms.The attack is indiscriminate as well as targeted in nature as the content of news article can be classified as Fake or Real News.The Goal of the attacker is to maximize the classification error and has limited knowledge about the machine learning model but the attacker can try a number of different news contents against the model to discover a way to get their fake news classified as innocuous.

II. Project Description

A. Approach Overview

The proposed plan is to identify strategies that can potentially evade common text classification methods employed to identify fake news.The dataset ”Getting Real about Fake News” from Kaggle contains text, metadata, and 6311 news articles which were collected over a period of 30 days in Jan 2017 from various media sources. The basic features of the datasets are mentioned in table 1. The label column in the datasets identifies each of the 6K news articles as fake or real.

Features of Fake News Dataset	
Features	Description
Title	Title of the news article
Text	Content of the news article
Label	There are 2 labels-Fake and Real

As a first step, data cleansing will be performed on this dataset to remove corrupt or inaccurate data, which will be followed by splitting the datasets into training and test data to be used by text classification methods. Bag of words model and Probability Context Free Grammars (PCFG) are two very popular techniques to identify fake from real news. Once an ML model is trained on the classification methods, an analysis would be performed on the false positives (Fake news classified as Real News) to identify the potential weakness of such approaches. Expanding on the idea, the paper would elaborate the weaknesses and vulnerabilities of such methods and how an evasive attack on techniques can guise a fake news as real news further decreasing the accuracy of the model.

The following metrics will be used in assessing the impact of evasive attack on the model.

% of fake news articles is calculated as

$$\frac{\text{No. of news articles labeled as fake}}{\text{total no. of news articles}}$$

% of real news articles is calculated as

$$\frac{\text{No. of news articles labeled as real}}{\text{total no. of news articles}}$$

or

(1 - % of fake news articles)

Efficiency of Attack is calculated as

$$\frac{\text{False Positive Rate after attack}}{\text{False Positive Rate before attack}}$$

If the number decreases then the impact of attack is sucessful.Confusion Matrix can also be used to demonstrate the efficiency of attack.

Bag of words approach aggregates and analyzes individual words or n-grams frequencies to reveal the fake news. This approach includes most frequent words as numeric training feature and trains this on a Machine Learning classifier (Random Forest). Deep Syntax is implemented through Probability Context Free Grammars (PCFG). Sentences are transformed to a set of rewrite rules (a parse tree) to describe syntax structure, for example, noun and verb phrases, which are in turn rewritten by their syntactic constituent parts [4]. The final set of rewrites produces a parse tree with a certain probability assigned. This method is used to distinguish rule categories (lexicalized, unlexicalized, parent nodes, etc.) for deception detection. Some of the weakness of PCFG parsing model are lack of sensitivity to lexical information and lack of sensitivity to structural preferences. This can be analyzed further to find a pattern and initiate evasive attacks.

B. Related Work

Fake News Detection can be approached from three different aspects: content, propagation and information source. This paper deals with introducing adversarial attacks by analyzing the content of the news articles as there is a clear distinction in the content of satire, fake, and real news [3].Undoubtedly, without proper domain knowledge, people can hardly distinguish between fake news and real news. [5] provides a typology of several varieties of veracity assessment methods emerging from two major categories linguistic cue approaches (with machine learning), and network analysis approaches. Recently, many researches have been devoted to automatic fake news detection on social media. However, Rhetorical Structure Theory (RST) on content data from NPRs Bluff the Listener achieved a 63% prediction accuracy over a 56% baseline [6]. Recently, the natural language techniques are being used in the detection of fake news.

Another challenging problem for news media is identifying and preventing click-baits in online news. Some of the research related to click-baits also used content-based approaches. A recent work [7] attempted to detect clickbait Tweets in Twitter by using common words occurring in click-baits, and by extracting some other tweet specific features. The browser extension Downworthy [8] detects clickbait headlines using a fixed set of common clickbait phrases and then converts them to meaningless garbage text.

To the best of my knowledge,there are no works related to adversarial attacks on fake news detection.This paper might use some of the ideas from good word attack on spam filters. Good Word Attack [9] which is one of the most popular techniques frequently employed by spammers. This technique involves appending spam messages to a set of good words that are common to legitimate e-mails (ham) but rare in spam. Spam messages inflated with good words are more likely to bypass spam filters.

C. Broader Impact

The proliferation of fake news is claimed to have an impact on even the 2016 US Presidential Elections. Fake news is a serious problem with multiple technology companies trying to curtail the spread of fake news by implementing natural language processing and machine learning algorithms. Hence, it is prudent to understand the limitations and weakness of fake news detectors.

D. *Intellectual Merit*

To the best of my knowledge,there are no works related to adversarial attacks on fake news detection.So some of the good word attack strategies that have been used in spam detection are being used for the attack on fake news detectors.One of the challenge as part of this project is the use of active attack.There are two different ways to carry out word attacks-active and passive. Active attacks use feedback from the methods whereas passive attack use hit and trial strategy. Also, Active attacks are more effective but are difficult to perform and technically challenging.

III. Project Plan

Week 1: April 29
Data Preprocessing and create bag of words model with random forest classifier

Week 2: May 6
Implementation of PCFG.

Week 3: May 13
Introduce Evasive attacks in bag of words model.

Week 4: May 20
Introduce Evasive attacks in Deep Syntax(PCFG) Implementation.

Week 5: May 27
Analyze the results and complete the report.

References

[1] Bessi A, Coletto M, Davidescu, G. A, Scala A, Caldarelli, G; and Quattrociocchi, W., Science vs Conspiracy: Collective Narratives in the Age of Misinformation, PLoS ONE 10(2):e011809317, 2015.

[2] Zollo F, Novak P. K, Del Vicario, M, Bessi A, Mozetic I, Scala A, Caldarelli G, Quattrociocchi W, Emotional Dynamics in the Age of Misinformation. PLoS ONE 10(9):e013874022, 2015b.

[3] D. Horne and Sibel Adal?, Benjamin Rensselaer,This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News,Polytechnic Institute 110 8th Street, Troy, New York, USA, 2016.

[4] Feng, S., Banerjee, R.Choi, Y, Syntactic Stylometry for Deception Detection in 50th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 171175, 2012.

[5] Niall J. Conroy, Victoria L. Rubin, and Yimin Chen, Automatic Deception Detection: Methods for Finding Fake News, Language and Information Technology Research Lab (LIT.RL), Faculty of Information and Media Studies, University of Western Ontario, London, Ontario, Canada, 2015.

[6] Rubin, V. L, Conroy, N. J, Chen Y, Towards news verification: Deception detection methods for news discourse in Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, HICSS48, 2005.

[7] M. Potthast, S. Kpsel, B. Stein and M. Hagen, Clickbait detection, Advances in Information Retrieval, 2016, Springer.

[8] A. Gianotto, Downworthy: A browser plugin to turn hyperbolic viral headlines into what they really mean. [online] Available:downworthy.snipe.net/

[9] D. Lowd and C. Meek,Good word attacks on statistical spam filters in Proceedings of the 2nd Conference on Email and Anti-Spam, 2005.

[10] Dhruv Ghulati, ”Introducing FactmataArtificial intelligence for automated fact-checking”,Factmata.