Adversarial Attacks against NLP-based Fake News Detection Models

Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu

{kyriezoe, hkguan}@whu.edu.cn

{mbhat2, justhsu}@wisc.cs.edu

WUHAN UNIVERSITY 1893 武漢大學

THE UNIVERSITY of WISCONSIN MADISON

Background

- ***Fake news***

Deceptive, misleading or even harmful, especially when they are disconnected from their original sources and contexts

Background

Fake News Rattles Nigerian Election Campaign

Teresa Holland
@TezHolland
Follow

@SenJeffMerkley Jeff, Hope it wasn't a terrorist attack in Louisiana, was it? #ColumbianChemicals

3:07 PM - 11 Sep 2014

MENU the ONION FIFA Frantically Announces 2015 Summer World Cup In ... TOP HEADLINES

Fantasy Sports

PRESENTED BY FanDuel

FIFA Frantically Announces 2015 Summer World Cup In United States

Global Soccer Tournament To Kick Off In America Later This Afternoon

ARTICLE
May 27, 2015
VOL 51 ISSUE 21
Sports · World · Fifa · Soccer

Background

WhatsApp Tackling Fake News by Limiting Your Ability to Forward Messages

TECH | FACEBOOK | ARTIFICIAL INTELLIGENCE

Facebook is using machine learning to spot hoax articles shared by spammers

Humans still do the actual fact-checking, but AI is helping pick up the

By James Vincent | Jun 21, 2018, 10:21am EDT

This Google-Funded Company Uses Artificial Intelligence To Fight Against Fake News

ard Marr Contributor
rprise & Cloud

Lying

MIT Is Using Machine Learning to Sniff Out Fake News

November 10, 2018

Written by Reuben Westmaas

Existing approaches

- ***Linguistic* approaches**

Grammar feature, word pattern, term count and appearance of certain expressions...

- ***Network approaches***

Background information

- ***Hybrid approaches***

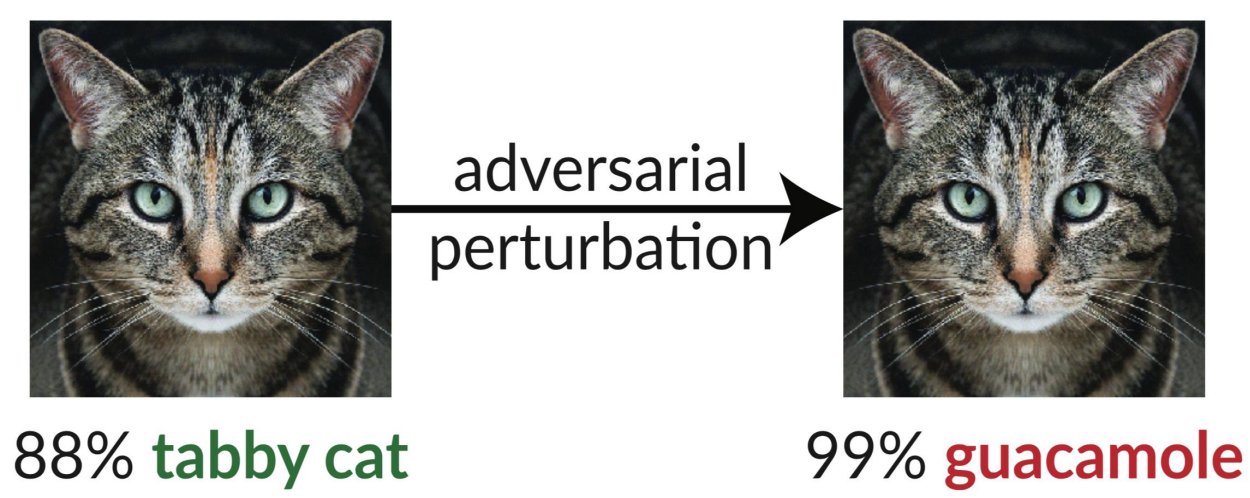
Current fake news detection models

The machine learning systems focus mainly on linguistic aspects of an article without fact checking (pure NLP systems)

Motivation

- *Machine learning (usually **computer vision** in previous research) systems are vulnerable to **adversarial attacks***

Adversarial example/image is a modified version of a clean image that is intentionally perturbed (e.g. by adding noise) to confuse/fool a machine learning technique, such as deep neural networks



Szegedy et al. 2013

Question: do adversarial attacks work on fake news detection models?

Our work

We propose three types of adversarial attacks and analyze their effectiveness on machine learning (NLP) systems

- ***Fact distortion***
- ***Subject-object exchange***
- ***Cause confounding***

Attack

- ***Fact distortion***

Definition

Exaggerating or modifying on some words. Character, time, location, relation, extent and any other element can be distorted

Example (before and after manipulation)

-12 people were injured in the shooting.

-24 people were killed in the shooting.

Attack

- ***Subject-object exchange***

Definition

Exchanging place of subject and object. With this attack readers will be confused as to who is the performer and who is the receiver of an action. It can be performed on sentence level

Example (before and after manipulation)

-A gangster was shot by the police.

-A policeman was shot by the gangster.

Attack

- ***Cause confounding***

Definition

Either building non-existent causal relationship between two independent events, or cutting off some parts of a story, leaving only the parts that an adversarial wants to present to his readers

Example (before and after manipulation)

-The condom policy originated in 1992 . . . The Boy Scouts have decided to accept people who identify as gay and lesbian. (unrelated events)

-The inclusion of gays, lesbians and girls in the Boy Scouts led to the condom policy.

Dataset for evaluation

McIntire's fake-real-news-dataset

An open-source dataset containing 6,335 articles. 3,171 are real and 3,164 are fake

Targeted model

Fakebox: a state-of-the-art NLP-based commercial fake news detector run by MachineBox

Checks several aspects of an article and gives a veracity score

- Title or headline: checked for clickbait;
- Content: analyzed to determine whether it's written like real news;
- Domain: some websites are known for hosting certain types of content, like hoaxes and satires.

Evaluation (applying McIntire's dataset to Fakebox)

Table 3: Normal-time accuracy of Fakebox with unsure cases excluded.

News type	Number of articles	Correctly classified	Classification accuracy
Real	2636	1159	43.97%
Fake	2721	2184	80.26%
Total	5357	3343	62.40%

Real and Fake are actual labels given by McIntire

It performs fairly well when handling "classical" fake news
What happens when it confronts our more subtle adversarial examples?

Evaluation (with our adversarial attacks)

- ***Fact distortion***

Practice example

For the article titled “Is the GOP losing Walmart?”, we substitute each “Walmart” in the content with “Apple”.

Outcome

The veracity score given by Fakebox drops down by only 0.0073, which is negligible for its judgement. **Still classified as real news**

Evaluation (with our adversarial attacks)

- ***Subject-object exchange***

Practice example

12 people were injured in the shooting.

-> 24 people were killed in the shooting.

Outcome

Veracity score doesn't change at all. **Still classified as real news**

Evaluation (with our adversarial attacks)

- ***Cause confounding***

Practice example

Simply mix two unrelated news together without subtle modification

Outcome

Veracity score doesn't change by much (and sometimes rises). **Still classified as real news**

It has no checking for correctness of causal relations

Implication

NLP-based fake news detectors are vulnerable to adversarial attacks

Possible reasons: word counts and term frequencies don't change; lack of fact checking mechanisms

Strawman proposal

One observation

Given that one of the essences of fake news is ***fact tampering, fact checking*** is of great help according to our analysis, but is largely ***missing*** in current NLP-based fake news detection models.

Another observation

Fake news usually floods on the ***early stage*** after an event happens, and local or well-informed people hear about the events faster and more accurately

Strawman proposal

- ***Extract*** key information from articles including causal relationships
- ***Compare*** it with a dynamically-updated news fact ***knowledge graph***
- A ***crowdsourcing*** way of building the knowledge graph may be feasible

Main ***drawback***: the difficulty of collecting high-quality information

"Accomplices"...

Future work

- Build a visualized interface for ***news knowledge graph crowdsourcing***

Users only need to fill in the “subject”, “action”, “object”, “time” and “location” entities, so as to make work as easy as possible for non-experts and stop fact-tampering fake news on early stage

Future work

- Look at the issue of fake news propagation from a different angle

Putting it in a social context and examining ***human factors*** in order to better understand the problem

Future work

- ***Natural language understanding (NLU)***

AI-hard

Future work

- ***Adversarial training***

Uses adversarial examples (automatic generation) to train machine learning models.

Some takeaways

- We explored whether NLP models, e.g., fake news classifiers, are vulnerable to adversarial examples
- We proposed three types of attacks: fact distortion, subject-object exchange and cause confounding
- We showed that these attacks are effective on a state-of-the-art model, Fakebox
- Our work suggests that fake news detectors based purely on NLP are not always effective
- Possible directions

Thank you!

Adversarial Attacks against NLP-based Fake News Detection Models

Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu

{kyriezoe, hkguan}@whu.edu.cn

{mbhat2, justhsu}@wisc.cs.edu

WUHAN UNIVERSITY 1893 武漢大學

THE UNIVERSITY of WISCONSIN MADISON