Fake News detection

Model building Perspective

Presented By-

- Sourav nandi
- Soumyadip Fadikar
- Rahul Singh
- Anish Chakrabarty

Under the guidance of

Karmaa Lab

The four observed flavors of “Possibly Misleading News”:-

- 1) *Clickbait* — Shocking headlines meant to generate clicks to increase ad revenue. Oftentimes these stories are highly exaggerated or totally false.
- 2) *Propaganda* — Intentionally misleading or deceptive articles meant to promote the author’s agenda. Oftentimes the rhetoric is hateful and incendiary.
- 3) *Commentary/Opinion* — Biased reactions to current events. These articles oftentimes tell the reader how to perceive recent events.
- 4) *Humor/Satire* — Articles written for entertainment. These stories are not meant to be taken seriously.

According to Prof **Kai Shu and Huan Liu, Arizona State University-**

- Social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information allow users to consume and share the news. On the other hand, it can make viral “fake news”, i.e., low-quality news with intentionally false information. The quick spread of fake news has the potential for calamitous impacts on individuals and society

For Example-

- the most popular fake news was more widely spread on Facebook than the most popular authentic mainstream news during the U.S. 2016 president election. Therefore, fake news detection on social media has attracted increasing attention from researchers to politicians.

History

- Fake news has existed for a very long time, nearly the same amount of time as news began to circulate widely after the printing press was invented in 1439

- Ref- http://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535

But the problem is-

- There is no clear cut definition of what can be called a fake news, and what cannot.
- Some papers regard satire news as fake news since the contents are false even though satire is often entertainment-oriented and reveals its own deceptiveness to the consumers .
- Other literature directly treats deceptive news as fake news ,which includes serious fabrications, hoaxes, and satires.

Definition

- Fake news is a news article that is intentionally and verifiably false
- There are two key features of this definition: authenticity and intent.
  - First, fake news includes false information that can be verified as such.
  - Second, fake news is created with dishonest intention to mislead consumers.
- This definition has been widely adopted in recent studies.
- Broader definitions of fake news focus on the either authenticity or intent of the news content

Which News is NOT fake

Following concepts are not fake news according to our definition:

- (1) satire news with proper context, which has no intent to mislead or deceive consumers and is unlikely to be mis-perceived as factual;
- (2) rumors that did not originate from news events;
- (3) conspiracy theories, which are difficult verify as true or false;
- (4) misinformation that is created unintentionally;
- (5) hoaxes that are only motivated by fun or to scam targeted individuals.

Motives for Finding Fake News

- Besides of the obvious motivation of not to influence individual's choices by distorted informaion, there are also other economic interests.
- The reality is that pretty much everyone mistakenly believes they can tell true from false, and fewer of us are willing to take several extra clicks to run an app that might flag content as true or false. The real stakeholders in this battle are all motivated by money.

The Reality

- **Aggregators like Facebook and Google** are already suffering from advertisers pulling their ads that were programmatically placed on sites that proved to provide fake news or other offensive content. Beyond that they have general reputational risk with their users who may visit less often if they are exposed to blatantly false or offensive material.

- **Legitimate news agencies** like Thomson Reuters or any of the major news broadcast or print organizations who take raw information feeds and convert it to news stories. They would suffer greatly if their material started to be compromised by falsehoods.

Source- Data Science Central

Fake News & Social Media- The famous Echo Chamber effect.

- Fake news on social media has its unique characteristics.
- For example, malicious accounts can be easily and quickly created to boost the spread of fake news, such as social bots, cyborg users, or trolls.
- In addition, users are selectively exposed to certain types of news because of the way news feed appear on the homepage in social media.
- Therefore, users on social media tend to form groups containing like-minded people where they are likely to polarize their opinions, resulting in an echo chamber effect.

Research Orientation

- Fake news detection on social media is a newly emerging research area. The survey [1] discusses related research areas, open problems, and future research directions from a data mining perspective. As shown in previous figure, research directions are outlined in four perspectives:
  - Data-oriented,
  - Feature-oriented,
  - Model-oriented,
  - and Application-oriented.

Problem Construction

Let a refer to a News Article. It consists of two major components: Publisher and Content. Publisher p(a) includes a set of profile features to describe the original author, such as name, domain, age, among other attributes. Content c(a) consists of a set of attributes that represent the news article and includes headline, text, image, etc.

- We also define Social News Engagements as a set of tuples $E = \{ei(t)\}$ to represent the process of how news spread over time among n users $U = \{u1, u2, ..., un\}$ and their corresponding posts $P = \{p1, p2, ..., pn\}$ on social media regarding news article a.
- Each engagement $ei(t) = \{ui, pi, t\}$ represents that a user ui spreads news article a using pi at time t. Note that we set t = Null if the article a does not have any engagement yet and thus ui represents the publisher.

Fake News detection

- Given the social news engagements E among n users for news article a, the task of fake news detection is to predict whether the news article a is a fake news piece or not.
- i.e., $F : E \rightarrow \{0, 1\}$ such that,
- F(a) =
  - 1, if a is a piece of fake news,
  - 0, otherwise.

  where F is the prediction function we want to learn.

Characterisation to Detection

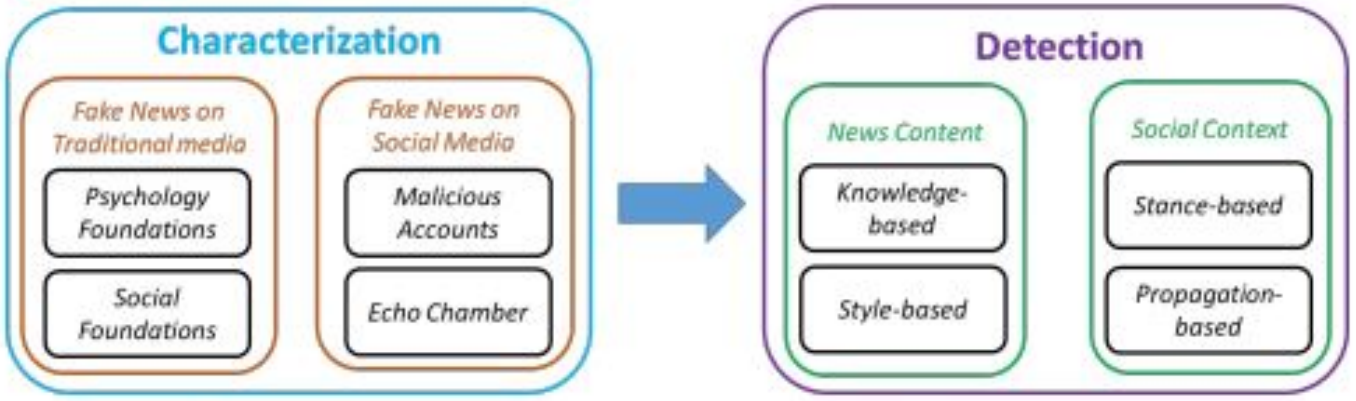

Figure 1: Fake news on social media: from characterization to detection.

Essentilly, this is a classification problem.

- Note that we define fake news detection as a binary classification problem for the following reason: fake news is essentially a distortion bias on information manipulated by the publisher.

- So, Standard Classification Algorithms will apply to here also.

Some Open Source Implementations Courtesy- Karmaa Lab

- https://nycdatascience.com/blog/student-works/identifying-fake-news-nlp/
- http://onlinelibrary.wiley.com/doi/10.1002/pra2.2015.145052010083/full
- https://web.stanford.edu/class/cs224n/reports/2710385.pdf

- https://arxiv.org/pdf/1708.01967.pdf
- https://towardsdatascience.com/i-trained-fake-news-detection-ai-with-95-accuracy-and-almost-went-crazy-d10589aa57c
- https://github.com/aldengolab/fake-news-detection

Main Packages & Technology Used

- Scikit-Learn
- Tensorflow
- Numpy
- TF-IDF vectorizer
- Binary cross-entropy loss
- K-fold cross-validation
- Naïve-Bayes Algorithm
- Adaptive Boosting
- Cross-validation, and Ensamble Techniques.

Evaluation Metrics

- Most existing approaches consider the fake news problem as a classification problem that predicts whether a news article is fake or not:
- • True Positive (TP): when predicted fake news pieces are actually annotated as fake news;
- • True Negative (TN): when predicted true news pieces are actually annotated as true news;
- • False Negative (FN): when predicted true news pieces are actually annotated as fake news;
- • False Positive (FP): when predicted fake news pieces are actually annotated as true news.

Evaluation Metrics

By formulating this as a classification problem, we can define following metrics,

$$Precision = \frac{|TP|}{|TP| + |FP|}$$

$$Recall = \frac{|TP|}{|TP| + |FN|}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

Performance Measurement

- Receiver Operating Characteristics (ROC) curve provides a way of comparing the performance of classifiers by looking at the trade-off in the False Positive Rate (FPR) and the True Positive Rate (TPR). To draw the ROC curve, we plot the FPR on the x axis and and TPR along the y axis. The ROC curve compares the performance of different classifiers by changing class distributions via a threshold

$$TPR = \frac{|TP|}{|TP| + |FN|}$$

$$FPR = \frac{|FP|}{|FP| + |TN|}$$

Use of NLP

- The news articles are not in numeric format.
- So, we have to use Natural Language Processing (NLP) techniques to deal with it.
- The suggestion of using Language Technologies (NLP, NLU etc.) to design solutions for modern online media phenomena such as “fake news”, “hate speech”, “abusive language”, etc. is receiving rapidly growing interest in the form of shared tasks, workshops and conferences

The Continued Influence Effect

- Yet, at the same time, it is being acknowledged that the problem is much more complex than anything that can be solved by exploiting current state of the art techniques alone.
- The effect known as “belief perseverance” or “continued influence effect” (Wilkes and Leatherbarrow, 1988) and its influence on modern media and politics is described by (Nyhan and Reifler, 2015), who state that-
  - “ reasoning based on facts that have shown to be false, remains in place until an alternative line of reasoning has been offered.”

An Example-
Courtesy- William Vorhies(DS Central)

Which of the following headlines is fake?

- **a) Police Investigating Clinton-backed Pizza Shop Pedophilia Ring**
- **b) The Pope Endorses Donald Trump for President**
- **c) Eastern European Fake News Sites Impact Pro-Trump Anti-Clinton Voter Sentiment**

The first two are in fact fake

- The third headline however is true, but not with the implications you may think.
- These were well circulated during last US Election coverage.
- Depending on which conspiracy theory you subscribe to, either the Russian government or someone else produced huge volumes of pro-Trump fake news with the intent of influencing the election.

Who is doing this

- *BuzzFeed* discovered last November is that at least 140 pro-Trump web sites were being run out of one small Macedonian town of Veles (population 45,000 and no not part of Russia) by a relatively small group of teens and young adults cashing in on the Google AdSense pay per click revenue stream.

What is the motive

- Their motive - strictly the cash. Apparently you can have a pretty good life style in Veles from just this source. Their political knowledge – none beyond the fact that these extremist web sites once on Facebook drew tons of cash generating clicks.
- So what's driving fake news is the change in the fundamental business model of news reporting, from a few well fact-checked news agencies to the new internet enabled anyone-can-write-a-headline-and-get-paid free for all. Essentially version 2.0 of the Nigerian prince email racket.

So, what can be done? A Possible Approach

- A) Data Collection and Preprocessing
- B) Sampling, including Bootstraping techniques.
- C) Classifier Building
- D) Improvement of Classifier

A) Data Collection

- Collect news articles from a set of credible and non-credible websites. Get training labels from OpenSources, a professionally curated database.
- They provide a continuously updated database of information sources for developers to leverage in the fight against fake, false, conspiratorial, and misleading news. The database is maintained by professionals who have analyzed each source, looking for overall inaccuracy, extreme biases, lack of transparency, and other kinds of misinformation.

B)Sampling, including Bootstraping techniques

Sample from the corpus in such a way that the training set contains an even number of unique articles from both credible and non-credible sources for each day of data collection.

C) Classifier Building

- This has to be done in accordance with the mathematical modelling.

To Build an ensemble classifier that considers the predictions of two separate models:

1. "Content-only" model (Multinomial Naive Bayes)
2. "Context-only" model (Adaptive Boosting)

D) Improvement of Classifier

Each classifier is retrained daily and subjected to cross validation testing to obtain updated accuracy scores. These scores are used to update weights in the final ensemble classifier.

Thus, using the future information about performance, our model could be made even better.

Finally, Caution About Credible Sources and the need for Cross-Checking from multiple sources-

- Sources that circulate news and information in a manner consistent with traditional and ethical practices in journalism Are Credible sources.
- But, even credible sources sometimes rely on clickbait-style headlines or occasionally make mistakes.
- No news organization is perfect, which is why a healthy news diet consists of multiple sources of information.

References

- Wikipedia
- Github
- Stack-Exchange
- [1] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In WWW'11.
- [2] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. In ASONAM'16.
- [3] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: Recognizing clickbait as false news. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, pages 15–19. ACM, 2015.
- [4] Justin Cheng, Michael Bernstein, Cristian DanescuNiculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In CSCW '17.
- [5] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Transactions on Dependable and Secure Computing, 9(6):811–824, 2012.

THANK YOU