Fake News :
Study & Application of Text Analytics

by

Pimpale Prakash Balaji

Dissertation work carried out at

Centre for Development of Advanced Computing, Mumbai

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)

April 2019

<Disclaimer for the public copy: I am like any other student. Documenting what I have done was my last priority that too in the very last days of the project submission ;) The copy was not properly re-read even by me. And I don't remember if it is the 'final' version or the 'final_final_before_submission' or the 'finally_last' version. I am not at all proud of this. Nor it should be criteria to judge organizations mentioned above. I am solely responsible for these things. And I have suffered penalty for this through the grades I got :(. Still I would like to make readers aware of this.

Also I have deleted some of the content which I may publish in the future! But I have kept all the content which will be of some use to others wanting to follow up on this.>

ABSTRACT :

News articles that are intentionally and verifiably false, and could mislead readers are called as fake news. These news articles are mainly used to spread hatred and manipulate public opinions among other serious purposes. These are mostly spread through digital media platforms like facebook, whatsapps and only online news sites. Avoiding fake news or making it easier to identify one is an important task for many digital platforms and even the whole society. Currently, the widely used methods to control the fake news are manual. Many platforms allow users to report such news so that the platforms can manually verify them. Some platforms pro-actively monitor the news and control the same. Given the volume and variety of the content and users, it’s a very challenging task.

There are attempts at making the process automatic using various natural language processing and machine learning techniques. Methods of stance detection and text classification have been applied. These methods have achieved the considerable ‘contest’ accuracy. But, fake news is a variable entity and plain text classification methods based on historical data are not sufficient in the real world.

In this work, an attempt to identify fake news using unsupervised and supervised methods is made. The devised system tries to verify the candidate fake news using other news sources over the web. It also makes use of text classification based on historical data to identify the fake news. The report details the explorations done to devise the system and reports some of the evaluations carried out.

Broad Academic Area of Work: Machine Learning

Key words : Information Extraction, Machine Learning, Natural Language Processing, Text Analytics

Acknowledgements

This is an opportunity to thank those who have been behind completion of my masters degree in technology from a reputed institution like BITS.

As my supervisor, Samadhan Manore has guided in many aspects and I thank him for the nice two year supervision. I also thank my additional examiner Mr. Prashant More for his valuable inputs. I would also like to thank my faculty mentor Dr. Aruna Malapati from BITS Hyderabad. Her question and concerns have helped me focus on what was more important for my topic – the data.

I also thank my organization and colleagues Sagun Baijal, Chandrashekhar Sharma, Archana Rane, Akshay Kumar, Manoj Kumar Singh and Balaji S K for various supports during the period of this course. I would like to specially thank Dr. M Sasikumar for allowing me to pursue this degree and encouraging not just me but everybody at C-DAC for higher education. There are not enough words to thank him for the many things that he has taught me and many of my colleagues, through his thoughts and actions. If I have any skills to look at the bigger picture of things, a big portion of those is learned from him.

This journey would not have been possible without motivation and support from my wife Pallavi. In fact she is the one who motivated me to enroll into this course and paid half of all my semester fees. I am thankful to her for all that and I am also thankful to have her. I thank my son Avishkar for supervising my studies. He has yelled at me many a times asking me to study seriously. I thank all my family members for supporting me through out my journey till date; specially, my father and grand father who have imbibed into me vision to look for learnings everywhere and not just in books!

I would use this opportunity to thank my friends Amol Suroshe, Parmeshwar Kamble and Nitin Bhande for all the things they have done for me.

At last, I thank two important people who were part of this journey, Chetan and Devashish (mentioned in alphabetical order). It really wouldn’t have been possible to pass any of the courses, without their encouragement and support during the last two years. I know, just thanking won’t be enough for these two and many others, I will need to throw a party to establish my thankfulness.

I thank BITS, its management and all the faculties for the opportunity to be part fascinating journey!

Table of Contents

1 Chapter 1 - Introduction

1.1 Overview

“To be news articles that are intentionally and verifiably false, and could mislead readers”

It is one of the a definitions of fake news. A recent study [1] used this to define Fake News. It further says that, the main motivations behind the production of fake news are financial and ideological. Both the aspects, financial and ideological are an important aspects of the human society. With fake news there is an effort to alter and manipulate those and that is fundamentally wrong and unethical.

Fake news is in existence for long time and many evidences can be referred from the history. What has changed is the speed with which the fake news travels. With the rapid developments in the communication and computing technologies, the spread of fake news is just click/swipe away. And with the same speed it affects the normal course of life of a society.

Corporations and governments are trying hard to minimize the fake news and its repercussions. Technological interventions can be devised to tackle it to some extent. We can explore technological solutions and their applications to deal with the fake news and contribute to the efforts of minimizing it and its repercussions.

The repercussions of the fake news are so serious that it changes the course of progress of the society. Recently it has shown its impacts in three important events in the history. One, the American Presidential election, two the Brexit and three, the last Indian election and post election events.

Various studies have been conducted to asses impact of the fake news on various events. Authors of [2] from MIT studies fake news and its spread in comparison to the real news. They found out astonishing facts about fake news one of which is reproduced below for the ready reference.

> *Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories. The effects were most pronounced for false political news than for news about terrorism, natural disasters, science, urban legends, or financial information. Controlling for many factors, false news was 70% more likely to be retweeted than the truth.*

The finding is surprising as our, or least mine, assumption about the collective public opinion is that collectively humans make right decisions. This assumption is also supported by the fact that Wikipedia has been equally ‘good and erroneous’ source of information as Encyclopedia [3]. Similarly, it’s also evident from the success of the crowdsourcing. But when it comes to news it’s a different fact. The fake I.e false is upheld taller than the real. This contradiction can be subject of another study. There must be factors affecting these two different case differently.

The fake news is different from a real typical news in following special ways:

- Fake news is made up of content that gives rise to feelings like fear, disgust and surprise in the minds of readers, whereas real news gives rise to feelings like joy, sadness and trust.
- Fake news has shocking claim in the headline and sounds unbelievable
- It's mostly published on not-so reliable source or circulated on social media sites
- It's not professionally written, there may be spelling errors and bad formatting

This work aimed to contribute to the task of controlling fake news through application of text processing to the news articles. Following sections will help readers understand the technical definition of the fake news and use of various approaches to build the fake news identification system. The report will also talk about the build system, it’s advantages, limitations and future scope.

1.2 Problem Statement

Automatic identification of the fake news is a challenging task. As defined by many, it’s a news with intentionally wrong facts. The narrative around the wrong facts may be real or cooked up. But fake news can be verified through knowledge of the reality. FNC i.e. Fake News Challenge has split fake news into multiple tasks. The first in the series is Stance Detection. This task aims to identify the relation between the title of the article and the content of the article. They assume that it will help the manual factcheckers perform their task. The idea is to create stance identification system using data share created by [4].

Many researchers [5, 6] have defined the problem in its entirety, i.e. *given a news text news, identify if it’s fake or not.* Most of the approaches dealing with the fake news rely on the text classification at various stages. Be it stance detection or classifying a news as fake or real in its entirety.

Challenge with such an approach is that, the different fake news are not generated from a same collection/random machine. So the life of trained models won’t be very long. The trained models in that case have to be updated at very shorter duration.

Considering this challenge, this work tries to deal with fake news using a combination of the two approaches – exploring the text and supervised learning. So for the purpose of solving it, the problem is split into two parallel problems.

Exploration of the news:

- Analysis of the source of the news for reliability
- Analysis of the title of the news using comparison with the content of the news
- Analysis of the content of the news
    - To highlight important concepts
    - To extract the facts and compare them with other news articles from reliable sources

A sample of the facts in fake news

Sentence	The new Rs 2,000 notes are embedded with Nano-GPS Chips.
Entities	1. The new Rs 2,000 notes 2. Nano-GPS Chips
False Relation	3. Embedded with

Table : False facts in a fake news

This exploration will help user understand the reliability, topic and inclination of the news.

Prediction of the Fake news using text classification:

- Training a machine learning based binary text classifier
- Using the trained model to classify the candidate news fake or real

The prediction will directly help user know about the fakeness of the news. But as discussed, this is not viable solution when the facts in the fake news are relatively not so popular or never occurred in the past.

The another challenge of fake news is, there is not much reliable, historical data to train the classification systems on. Fake news identification also depends heavily on nature of the spread of the news. Fake news has different spread phenomenon compared to the real news. For that aspect to be used, we need access to platform information. The same is not available with most of the researchers, except the ones at companies like facebook and whatsapp.

2 Chapter 2 – Foundational Concepts and their Use

In this chapter we will talk about the various techniques of text and natural language processing that are foundational to this work. We will not just talk about the techniques, but also discuss the experiments carried out to create the parts of this work.

< -Deleted in public version- contains technical details of NLP Basics and Advanced concepts – for details please get in touch with me >

3 Chapter 3 – Text Classification

3.1 Text Classification

As objective of the task is to identify if a given news is fake or real, the same can be considered as a binary text classification task. The task of text classification for this purpose can be defined as: given input text features, identify the class of the news.

Feature: Terms – unigrams and n-grams
Class: Class of article **{fake, real}**

3.2 Data Collection

As we mentioned earlier reliable data sources for the fake news identification are very limited. Upon, research multiple open datasets were found as follows:

- FakeNewsNet: Created and published by [9]
  link to download : https://github.com/KaiDMML/FakeNewsNet
  Description: 420 news articles for two classes fake (209) and real(211). These are manually labeled by experts. This was found to be most reliable and usable data set. But this dataset has every article in JSON as independent file and had to be combined into one csv file.
- University of Maryland fake news Dataset: Created and published by [10]
  linke to download: https://github.com/jgolbeck/fakenews
  Description: This is also a labeled dataset with individual files for news articles. These articles are in plain text form with title, link and article on each line. The labels in this data are fake and satire.
- Kaggle fake news Dataset: Anonymous
  linke to download: https://www.kaggle.com/c/fake-news
  Description: This has articles classified into fake and real classes. Total train set is of 20.8K and test set is of 5.2K. But I was not able to verify the reliability of the data through multiple sources.
- LIAR fake news Data set: Created and published by [11]
  link to download: https://www.cs.ucsb.edu/~william/data/liar_dataset.zip
  Description: It has Training set size 10,269, Validation set size 1,284 and Testing set size 1,283. There is no fake news article as such in the data, every article is just a statement with author, context and justification information for assigned label. This was not relevant to the task that we are doing as we are expecting the complete article and not just single sentences.
- Snope fact checked Data set: Created and published by https://www.snopes.com/
  link to download:
  http://resources.mpi-inf.mpg.de/impact/web_credibility_analysis/Snopes.tar.gz
  Description: It has 4856 articles which are fact checked. There no fake or real labels as such but human explained fakeness of the article. This data set was not suitable for the classification task we targeted.

The above datasets are useful datasets for the fakenews identification, but more work needs to be done to clean them and combine into one good dataset for the task.

< -Deleted in public version- contains technical details about all machine learning experiments carried out and their results – for details please get in touch with me >

4 Chapter 4 – System Design and Development

As a part of this work, a system called **news.isFake()** has been developed using Java and Spring framework. The core functionalities are developed using Java and web interfaces is developed using Spring MVC framework.

Following sections details how different techniques have been collectively used create the system. For the first hand glance of the system a screen shot is given below.

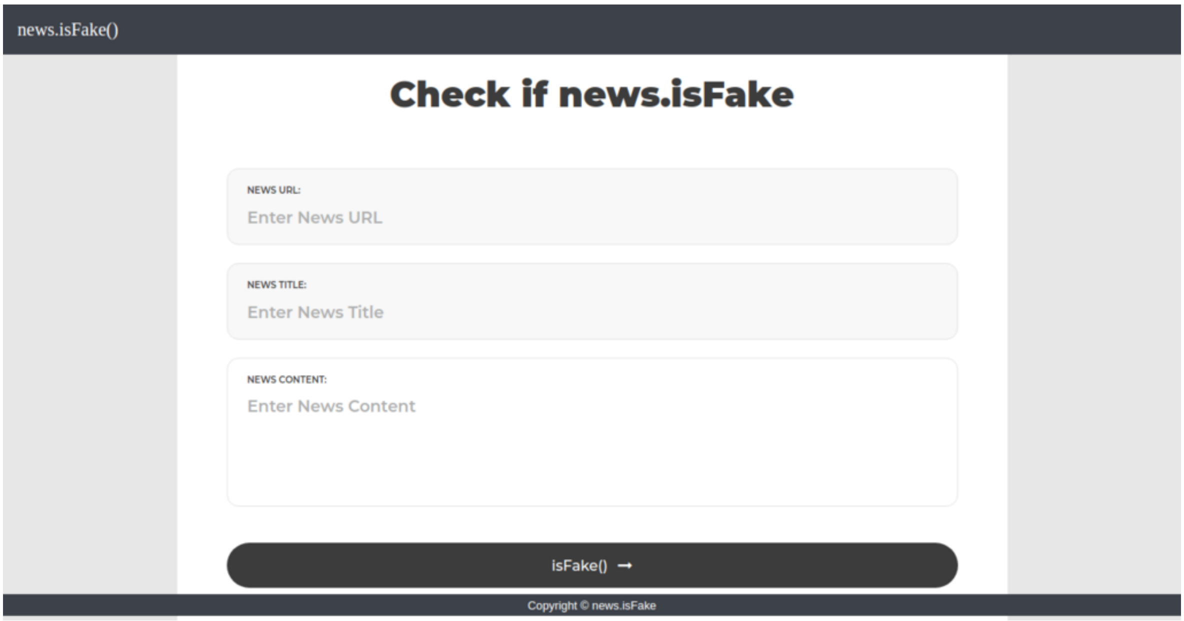
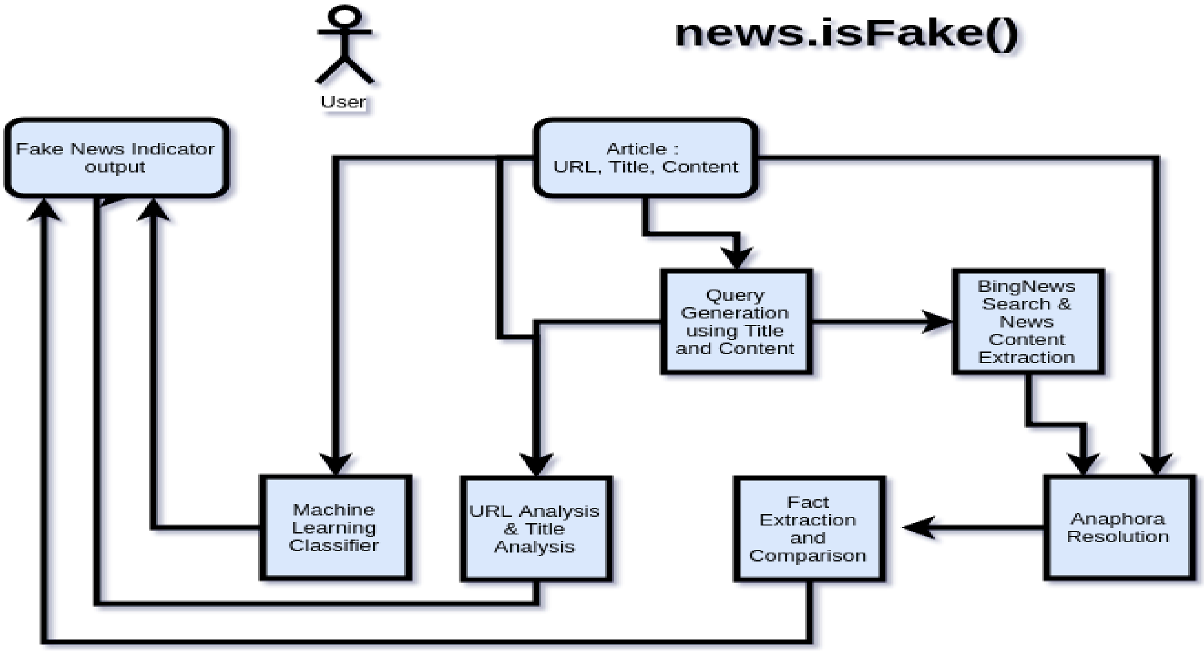


Figure: Home Page of the System

4.1 Block Diagram



news.isFake()
User
Fake News Indicator output
Article : URL, Title, Content
Query Generation using Title and Content
BingNews Search & News Content Extraction
Machine Learning Classifier
URL Analysis & Title Analysis
Fact Extraction and Comparison
Anaphora Resolution

Figure: Block Diagram of the System

4.1.1 Article Input

A news Article is define as a combination of the URL, Title and Content of the article. This taken as input from the user. The appropriate cleaning methods are applied to clean title and content.

News class in the system has URL, Title, Content and List of facts in the News. Whereas, facts are nothing but entities and relation between them.

We can represent that here as:

NEWS:

URL

TITLE

Content

List of FACTS

FACT:

Entity one

Entity two

Relation

4.1.2 Query Generation

Using the title and content a query is generated for firing on the search engine. Custom query generation method is devised and also there is provision to use the title as it is for the query.

The query generation, takes place in three steps: stop word removal, lemmatization, Ngram generation and intersection of the title and content ngrams. The cleaned query is formed from this process which is then fired to the Bing News Search API.

Example:

Title: Pope Francis has endorsed Hillary Clinton for President.

Content: News outlets around the world are reporting on the news that Pope Francis has made the unprecedented decision to endorse a US presidential candidate. His statement in support of Hillary Clinton was released from the Vatican this evening: “I have been hesitant to offer any kind of support for either candidate in the US presidential election but I now feel that to not voice my concern would be a dereliction of my duty as the Holy See. A strong and free America is vitally important in maintaining a strong and free world and in that sense what happens in American elections affects us all. With that at the forefront of my mind I must express my strong reservations about Mr. Donald Trump. His demeanor and temperament should preclude him from becoming President. I fear he may be disastrous to the security, stability, and prosperity of the United States and to the world. I believe that Secretary Clinton would be a better, more stable choice. Though I don’t agree with Secretary Clinton on some issues I’m asking, not as the Holy Father, but as a concerned citizen of the world that Americans vote for Hillary Clinton for President of the United States.

Generated Query: pope francis have hillary for president

4.1.3 Bing News Search and Content extraction

The created query is fired to the Bing news search API which returns the title, url and date of publication of the news. It doesn’t return the content of the news. The news content is fetched and cleaned for HTML and other tags using JSOUP utility.

4.1.4 Anaphora Resolution

Both uncleaned articles, i.e. not stop word removed and not punctuation removed are processed to replace the pronouns with their references.

4.1.5 Fact Extraction and Comparison

The anaphora resolved articles are processed to extract the facts from them. The extracted facts from the input news articles are compared will all the facts from these searched news sources. The record of matched facts is retained and further presented to user for his conclusion.

4.1.6 URL Analysis and Title Analysis

In the URL analysis the news URL is looked up in the master list of reliable sources. The result of same is represented to the user for his conclusion.

Along with this, the title is analyzed to see which all concepts are supported by the article. This is done using the ngram generation and matching.

4.1.7 Machine Learning Classifier

As detailed in the last chapter, a machine learning model is called on the content of the news article and results are presented to the user. This gives straight forward conclusion about the fakeness of the news.

4.1.8 Fake News Indicator Output

The user is presented with analysis of the URL, Title and Content. The analysis of the content includes most frequent words from the article excluding the stop words. This will help user get a glance of the over content in one shot.

The user is presented with the facts from the article, that the system was able to verify from other sources. The list of sources with which the facts were compared is also presented to the user.

Also the result from machine learning classifier is presented which is conclusive decision by the system for the article as Fake or Real.

< - Deleted in public version – for details please get in touch with me >

5 Directions for future work

For the fake news identification, this is just a start. Good dataset for text classification approach are need and can be developed with compilation of the available data sources.

Apart from that, for fake news identification it's also important to know the spread of news and network information regarding the same. That information is available mostly with the platforms where the news spreads. These platforms can share this information with users like trail of email, where we should be able to know the source and path of the news. This meta information can also be used to identify the fake news.

Problem, with fake news identification by comparing it to available sources need the coverage of the news to be done by so called reliable sources. If they miss some genuine news coverage for the events, even the real news will be termed as fake as it's not covered by the reputed or reliable news sources.

There are many fact databases that are being built, those fact databases along with the these approaches to compare the facts will be helpful and can be looked upon for further development.

References:

1. Tandoc, Edson & Wei Lim, Zheng & Ling, Rich. (2017). Defining “Fake News”: A typology of scholarly definitions. Digital Journalism. 1-17. 10.1080/21670811.2017.1360143.
2. Soroush Vosoughi, Deb Roy, and Sinan Ara, The Spread Of True And False News Online, Mit initiative on the digital economy research brief, MIT, 2018.
3. Jim Giles, “Internet encyclopaedias go head to head,” Nature, December 15, 2005
4. Ferreira, W. and Vlachos, A., Emergent: a novel data-set for stance classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies (pp. 1163-1168).
5. Popat K. Assessing the Credibility of Claims on the Web. InProceedings of the 26th International Conference on World Wide Web Companion 2017 Apr 3 (pp. 735-739).
6. Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R. Automatic detection of fake news. arXiv preprint arXiv:1708.07104. 2017 Aug 23.
7. Angeli G, Premkumar MJ, Manning CD. Leveraging linguistic structure for open domain information extraction. InProceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) 2015 (Vol. 1, pp. 344-354).
8. Pawar S, Palshikar GK, Bhattacharyya P. Relation Extraction: A Survey. arXiv preprint arXiv:1712.05191. 2017 Dec 14.
9. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286. 2018 Sep 5.
10. Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Paul Cheakalos, Jeannine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly Hoffman, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn Iv, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Alexandra Steinheimer, Aditya Subramanian and Gina Visnansky. 2018. Fake News vs Satire: a Data Set and Analysis. Proceedings of the 10th ACM Conference on Web Science. Amsterdam, the Netherlands.
11. Wang WY. " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648. 2017 May 1.
12. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.