Fake News Detection System

Mukesh Yadav, Rishab singh, Rahul Kumar,Priyvart Raghav

Department of Computer Science, IMS Engineering,
NH 24, Ghazibad,India

ABSTRACT

The proliferation of misleading information in everyday access media outlets such as social media feeds, news blogs, and online newspapers have made it challenging to identify trustworthy news sources, thus increasing the need for computational tools able to provide insights into the reliability of online content. In this paper, we focus on the automatic identification of fake content in online news. Our contribution is twofold. First, we introduce two novel datasets for the task of fake news detection, covering seven different news domains. We describe the collection, annotation, and validation process in detail and present several exploratory analyses on the identification of linguistic differences in fake and legitimate news content. Second, we conduct a set of learning experiments to build accurate fake news detectors. In addition, we provide comparative analyses of the automatic and manual identification of fake news.

Keywords: Naïve Bayes, Random Forest, NumPy, Fake News Detection, Neural Networks, Natural Language Processing.

1. INTRODUCTION

With people spending more time on the social media platforms, they are more prone to consume information from social media. Social media is free of cost, easy to access and help one to express opinions and hence it acts an excellent source for an individual to consume information from social media. But the quality of news on social media is generally lower than the traditional news organizations. It is because anyone can spread information they want in the social media and there is no regulating authority to control the information. Fake news, as a specific type of disinformation, means the false information that is spread deliberately to deceive people. Some individuals and organization use social media as a tool to spread disinformation for financial and political gains. It was approximated that over million tweets are related to fake news “Pizzagate” by the end of US presidential election. This consequence has adverse effects and the opinions of people are biased because of fake news. Thus, it is important to address this issue. Fake news detection is an important and technically challenging problem. In an attempt to tackle the growing misinformation, several fact-checking websites have been deployed to expose the fake news. These websites play a crucial role in clarifying fake news, but they require expert analysis which is time-consuming. Numerous articles and 2 blogs are written in order to distinguish fake news from the true news. However, they are not from authority sources and may be biased, which make the labels not fully reliable and convincing. Due to the volume and diversity of the social media, it is almost impossible to manually label the fake news and true news. Thus, to solve these challenges, we present a system FakeNewsTracker, to facilitate the community for studying fake news. Fig. 1

shows the various components in the FakeNewsTracker system. The major functionalities of FakeNewsTracker are as follows,

1. Fake News Collection: collecting news contents and social context automatically, which provides valuable datasets for the study of fake news;
2. Fake News Detection: extracting useful features and build various machine learning models to detect fake news;
3. Fake News Visualization: presenting the characteristics of fake news dissemination through effective visualization techniques

DESGIN AND OPERATION

Technologies Used: The system consists of the following technologies or field of work on :

• **Natural Language Processing** : Natural language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

• **Artificial Intelligence**:Artificial intelligence (AI) is an area of computer science that emphasizes the creation of intelligent machines that work and react like humans. Some of the activities computers with artificial intelligence are designed for include: Speech recognition.

Functional description of module

• **Natural Language Processing:** Then with the help of Natural Language Processing we will match the strings of the Sentence and from the trained data we will find the results .

Linguistic Features :To build the fake news detection models, we start by extracting several sets of linguistic features:

• **Ngrams:** We extract unigrams and bigrams derived from the bag of words representation of each news article. To account for occasional differences in content length, these features are encoded as tf-idf values.

•

• **Punctuation:** Previous work on fake news detection as well as on opinion spam suggests that the use of punctuation might be useful to differentiate deceptive from truthful texts. We construct a punctuation feature set consisting of twelve types of punctuation derived from the Linguistic Inquiry and Word Count software . This includes punctuation characters such as periods, commas, dashes, question marks and exclamation marks.

- **Naïve Bayes:** Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set

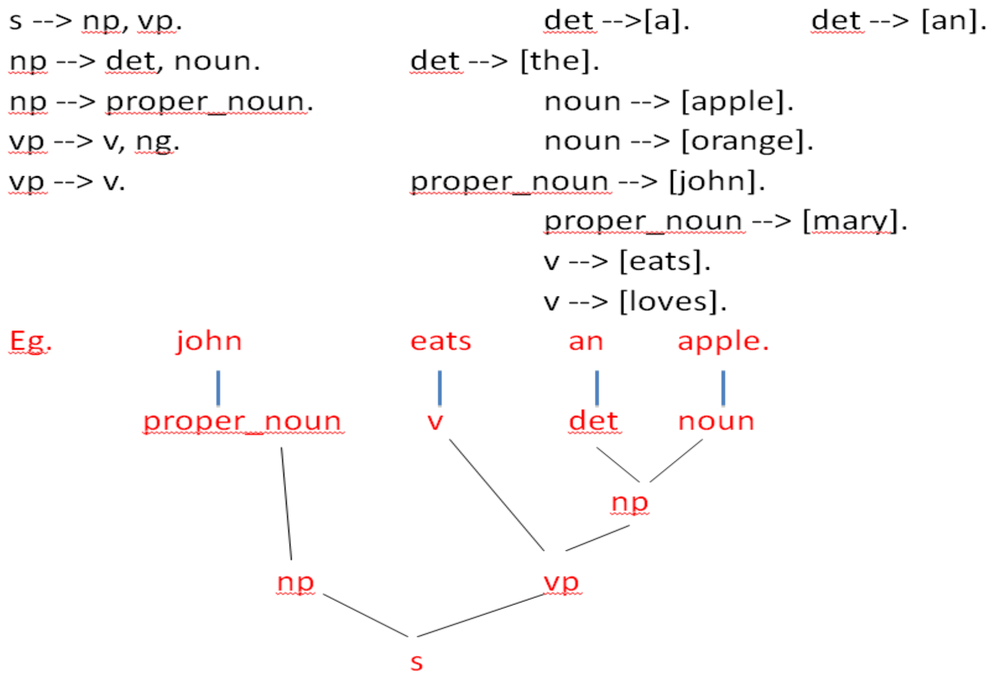
$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

.

- **Random Forest:** Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

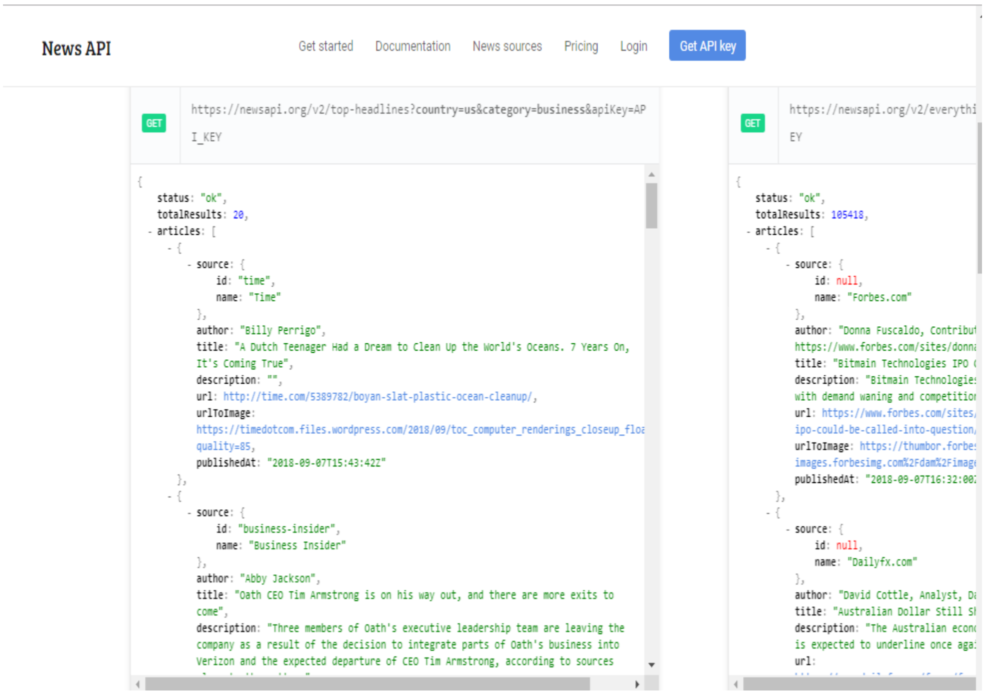
- **Readability:** We also extract features that indicate text understandability. These include content features such as the number of characters, complex words, long words, number of syllables, word types, and number of paragraphs, among others content features. We also calculate several readability metrics, including the Flesch-Kincaid, Flesch Reading Ease, Gunning Fog, and the Automatic Readability Index (ARI).

- **Syntax:** Finally, we extract a set of features derived from production rules based on context free grammars (CFG) trees using the Stanford Parser (Klein and Manning, 2003). The CFG derived features consist of all the lexicalized production rules (rules including child nodes) combined with their parent and grandparent node, e.g., *NNˆNP→commission (in this example NN –a noun– is the grandparent node, NP –noun phrase– the parent node, and “commissions” the child node). CFG-based features have been previously shown to be useful for linguistic deception detection (Feng et al., 2012). Features in this set are also encoded as tf-idf values.



```
s --> np, vp.                          det -->[a].        det --> [an].
np --> det, noun.           det --> [the].
np --> proper_noun.                    noun --> [apple].
vp --> v, ng.                          noun --> [orange].
vp --> v.                   proper_noun --> [john].
                                       proper_noun --> [mary].
                                       v --> [eats].
                                       v --> [loves].
```

Figure 1. *Processing of string*



```
https://newsapi.org/v2/top-headlines?country=us&category=business&apiKey=AP
I_KEY
```

```
{
   status: "ok",
   totalResults: 20,
 - articles: [
   - {
     - source: {
          id: "time",
          name: "Time"
       },
       author: "Billy Perrigo",
       title: "A Dutch Teenager Had a Dream to Clean Up the World's Oceans. 7 Years On,
       It's Coming True",
       description: "",
       url: http://time.com/5389782/boyan-slat-plastic-ocean-cleanup/,
       urlToImage:
       https://timedotcom.files.wordpress.com/2018/09/toc_computer_renderings_closeup_floa
       quality=85,
       publishedAt: "2018-09-07T15:43:42Z"
     },
   - {
     - source: {
          id: "business-insider",
          name: "Business Insider"
       },
       author: "Abby Jackson",
       title: "Oath CEO Tim Armstrong is on his way out, and there are more exits to
       come",
       description: "Three members of Oath's executive leadership team are leaving the
       company as a result of the decision to integrate parts of Oath's business into
       Verizon and the expected departure of CEO Tim Armstrong, according to sources
```

```
https://newsapi.org/v2/everythi
EY
```

```
{
   status: "ok",
   totalResults: 105418,
 - articles: [
   - {
     - source: {
          id: null,
          name: "Forbes.com"
       },
       author: "Donna Fuscaldo, Contribut
       https://www.forbes.com/sites/donna
       title: "Bitmain Technologies IPO (
       description: "Bitmain Technologies
       with demand waning and competition
       url: https://www.forbes.com/sites/
       ipo-could-be-called-into-question/
       urlToImage: https://thumbor.forbes
       images.forbesimg.com%2Fdam%2Fimage
       publishedAt: "2018-09-07T16:32:00
     },
   - {
     - source: {
          id: null,
          name: "Dailyfx.com"
       },
       author: "David Cottle, Analyst, Da
       title: "Australian Dollar Still Sh
       description: "The Australian econo
       is expected to underline once agai
       url:
```

Figure 1. *Code line of work*

4. FUTURE SCOPE OF THE PROJECT

- Modified to be the website who can effectively demolish the fake news.

- In Future can call more than 1000 websites to get the dataset from that to analyze the data.

- User friendly, Futuristic Website which can resolve the problem of Society

5. CONCLUSIONS

With an increasing focus of academic researchers and practitioners alike on the detection of online misinformation, the current investigation allows for two key conclusions. First, computational linguistics can aide in the process of identifying fake news in an automated manner well above the chance level. The proposed linguistics-driven approach suggests that to differentiate between fake and genuine content it is worthwhile to look at the lexical, syntactic and semantic level of a news item in question. The developed system's performance is comparable to that of humans in this task, with an accuracy up to 76%. Nevertheless, while linguistics features seem promising, we argue that future efforts on misinformation detection should not be limited to these and should also include meta features (e.g., number of links to and from an article, comments on the article), features from different modalities (e.g., the visual makeup of a website using computer vision approaches), and embrace the increasing potential of computational approaches to fact verification (Thorne et al., 2018). Thus, 3400 future work might want to explore how hybrid decision models consisting of both fact verification and data-driven machine learning judgments can be integrated. Second, we showed that it is possible to build resources for the fake news detection task by combining manual and crowsourced annotation approaches. Our paper presented the development of two datasets using these strategies and showed that they exhibit linguistic properties related to deceptive content. Furthermore, different from other available fake news datasets, our dataset consists of actual news excerpts, instead of short statements containing fake news information.

6. REFERENCES

1. A Gary D Bond and Adrienne Y Lee. 2005. Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. Applied Cognitive Psychology, 19(3):313–329.

2. Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. News in an online world: The need for an "automatic crap detector". Proceedings of the Association for Information Science and Technology, 52(1):1–4.

3. Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1):1–4.

4. Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, pages 171–175. Association for Computational Linguistics.

5. Wiki :https://en.wikipedia.org/wiki/Fake_news