Machine Learning Project Report

On

Fake News Detection Using Knowledge Graph

Submitted by

Ayush Singh Chauhan (181500180)

Rohit Gupta (181500590)

Submitted To:

Mr.Hradesh Kumar

Department of Computer Engineering and Applications IET, GLA University-Mathura

Declaration

We hereby declare that the work which is being presented in the Machine learning project ”Fake News Detection Using Knowledge Graph”, in partial fulfillment of the requirements for Machine learning-Project viva voce, is an authentic record of our own work carried under the supervision of Mr. Hradesh Kumar

Signature of Candidates:

Name of Candidates: **Roll. No**

- Ayush Singh Chauhan 181500180
- Rohit Gupta 181500590

Course: COMPUTER SCIENCE AND APPLICATION

Year: THIRD YEAR

Semester: VITH

Acknowledgement

It gives us a great sense of pleasure to present the report of the B. Tech Machine Learning Project undertaken during B.Tech. Third Year.

This project in itself is an acknowledgement to the inspiration, drive and technical assistance contributed to it by many individuals. This project would never have seen the light of the day without the help and guidance that we have received. Our heartiest thanks to Dr. (Prof) Anand Singh Jalal, Head of Dept., Department of CEA for providing us with an encouraging platform to develop this project, which thus helped us in shaping our abilities towards a constructive goal. We owe special debt of gratitude to Mr. Pawan Kumar Varma, Hradesh Kumar, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. He has showered us with all his extensively experienced ideas and insightful comments at virtually all stages of the project & has also taught us about the latest industry-oriented technologies.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind guidance and cooperation during the development of our project. Last but not the least, we acknowledge our Teammate for their contribution in the completion of the project.

Abstract

The Internet has become an integral part of our lives. According to a report, the average time users all over the world spend on gregarious media is about 145 minutes a day. As we all ken that smartphones and the Internet have transfigured how we engender and consume news or information in general on the go. More than 40% of the world's internet users read their news online only. And in the time of social media where people consume news from social media, which acts like a double-barreled shotgun with one barrel bent rearwards. At one end, it has revolutionized the world by providing inexpensive, facile to acquire, expeditious circulation of information. On the other hand, it is the major cause of widespread 'fake news' or erroneous news. According to Lloyd's Register Foundation's World risk poll, Fake News ranked among major online jeopardies. The widespread distribution of counterfeit goods can have a wide-ranging negative impact on society and everything in the world. The cyber-world has the power to turn the world upside down in both senses, positively as well as negatively. Mendacious information is now becoming an authentic hazard because of the fact that it can be engendered, diffused, and consumed very facilely and it is a big conundrum to solve because of identification, tracking, and managing untrustworthy content. Fake News is additionally a conundrum to solve due to the arduousness of tracking, identifying, and managing untrustworthy content. Even if we got acquainted with any fake story being circulated online, expunging it or obviating people from sharing it could be perceived as an endeavor of intervention and censorship. Consequently, fake news detection on the internet has recently become an emerging research that is magnetizing tremendous attention. As a result, in order to address this issue, we proposed a model that makes use of Knowledge Graph, which is concerned with the veracity of information on the internet. Our algorithm validates the truthfulness of the text on the Web and tests the credibility of the top 15 Google search results by calculating the (Rp) and additionally by engendering a knowledge graph of the results and matching their knowledge graph; if they match over a threshold value, we can determine whether the news is genuine or fake.

CONTENT

INTRODUCTION

ABOUT

Fake news simple meaning is to incorporate information that leads people to the wrong path. Nowadays fake news spreads like water and people share this information without verifying it. This is often done to further or impose certain ideas and is often achieved with political agendas. Most of the information available on social media is in the form of images. This has given rise to fake news event distribution, which is misinforming the users. Hence, to tackle this problem, we propose a model which is concerned with the veracity analysis of information on various social media platforms available in the form of images. It involves an algorithm which validates the veracity of image text by exploring it on web and then checking the credibility of the top 10 Bing, Yahoo and Google search results by subsequently calculating the reality parameter (Rp), which if exceeds a threshold value, an event is classified as real else fake. In order to test the performance of our proposed approach, we compute the recognition accuracy, and the highest accuracy is compared with similar state-of-the-art models to demonstrate the superior performance of our approach.

WHAT WE HAVE IMPLEMENTED

- Exploration , analysis and cleaning of news dataset and then

- Create different entities of news data using a NLP model and then find relation between entities

- Using those entities and their relations create a knowledge graph of text.

- Use different search APIs to search the text on web and get top 10 results from all.

- Create knowledge graphs of all the results from searches and compare them with the knowledge graph we have created on the dataset.

- If 60% of the Knowledge graphs match with our dataset Knowledge Graph then we say the news is True else fake.

WHAT WE WILL LEARN

- What is a Knowledge Graph?
- How to Represent Knowledge in a Graph?
    - Sentence Segmentation- **The first step in building a knowledge graph is to split the text document or article into sentences**
    - Entities Extraction- Extract **the subject/object along with its modifiers, compound words and also extract the punctuation marks between them**.
    - Relations Extraction - To **build a knowledge graph, we need edges to connect the nodes (entities) to one another.**
- Build a Knowledge Graph from Text Data

We will learn How to Explore, analyze data and clean data

- Knowledge of Git
- APIS- We are Also using Some APIs for making Our more efficient and Reliable to use APIs like Google search , Bing search api

WHAT WE WILL NEED

Software: For this project we just need Jupiter Notebook, Machine learning libraries like spacy, sklearn etc., Git hub as our VCS (version control system).

NLP model used- en core web sm

Hardware Requirements:

We just need a laptop/pc/mac with i5 processor, 8GB of RAM

Development Hardware that We have used is listed Below:

- I5 7^{th} /9^{th} GEN HQ KABI-LAKE / Coffee Lake Refresh family
- RAM – 8/12GB
- 1/2TB HARDRIVE + 128GB/256GB SSD

Or

Use Google Colabaratory.

Project Application:

When someone (or something like a bot) impersonates someone or a reliable source to false spread information that can also be considered as fake news. In most cases, the people creating this false information have an agenda, that can be political, economic or to change the behavior or thought about a topic. There are countless sources of fake news nowadays, mostly coming from programmed bots, that can't get tired and continue to spread false information 24/7.Fake News are dangerous is various ways like Fake News Can Be Harmful to Your Health There are many fake and misleading news stories related to medical treatments and major diseases like cancer or diabetes. Trusting these false stories could lead you to make decisions that may be harmful to your health. The main Motive of News is make people aware of about right and wrong but fake news can turn things around by making wrong impression something right, Fake News Makes It Harder For People To See the Truth, There are also various Democratic impacts like Politician and their supporters and workers uses to spread propaganda and also fake news for defaming opposition, These are Possible because many even today think fake news are news article so we should believe it, A fake News detector will be various Useful for Detecting news he is fake or not if we will get to what is right and what is wrong then there will lesser conflict and world will be a better place.

Features:

Our consequential contributions to this area and the key features of the proposed technique can be summarized as follows:

- We propose a novel fake news authentication system for the detection of fake news on the internet. This model comprises five integrated units, namely, Sentence Segmentation from text, entity extractor, Extract Relations, Build a Knowledge Graph from Text Data scraping the web and processing unit, .

- We show that our system can reliably detect fake news on sundry internet platforms.
- To validate the proposed fake news detection framework, the dataset is used from kaggle.

DATASET:

Link:

https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

Dataset contains following labels

- id: unique id for a news article
- title: the title of a news article
- author: author of the news article
- text: the text of the article; could be incomplete
- label: a label that marks the article as potentially unreliable
    - 1: unreliable
    - 0: reliable

Exploratory Analysis

Dataset stats

Number of variables	4
Number of observations	23481
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	3
Duplicate rows (%)	< 0.1%
Total size in memory	67.9 MiB
Average record size in memory	3.0 KiB

Columns stats

```
1 profile.to_widgets()
```

```
/usr/local/lib/python3.7/dist-packages/pandas_profiling/profile_report.py:361: UserWarning: Ipywidgets is not yet fully supported on Google Colab (https://github.com/googlecolab/colabtools/issues/60).As an alternative, you can use the HTML report. See the documentation for more info
  "Ipywidgets is not yet fully supported on Google Colab (https://github.com/googlecolab/colabtools/issues/60)."
```

Overview | Variables | Interactions | Missing values | Sample | Duplicate rows

Overview | Reproduction | Warnings (5)

	Duplicates
Dataset has 3 (< 0.1%) duplicate rows	
title has a high cardinality: 17903 distinct values	High cardinality
text has a high cardinality: 17455 distinct values	High cardinality
date has a high cardinality: 1681 distinct values	High cardinality
title is uniformly distributed	Uniform

Report generated with pandas-profiling

date

Categorical

HIGH CARDINALITY

Distinct count	1681
Unique (%)	7.2%
Missing	0
Missing (%)	0.0%
Memory size	183.6 KiB

title

Categorical

HIGH CARDINALITY
UNIFORM

Distinct count	17903
Unique (%)	76.2%
Missing	0
Missing (%)	0.0%
Memory size	183.6 KiB

text

Categorical

HIGH CARDINALITY

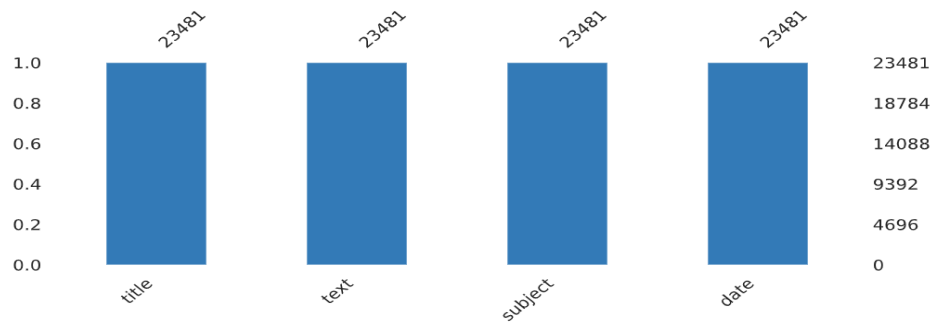
Distinct count	17455
Unique (%)	74.3%
Missing	0
Missing (%)	0.0%
Memory size	183.6 KiB

subject

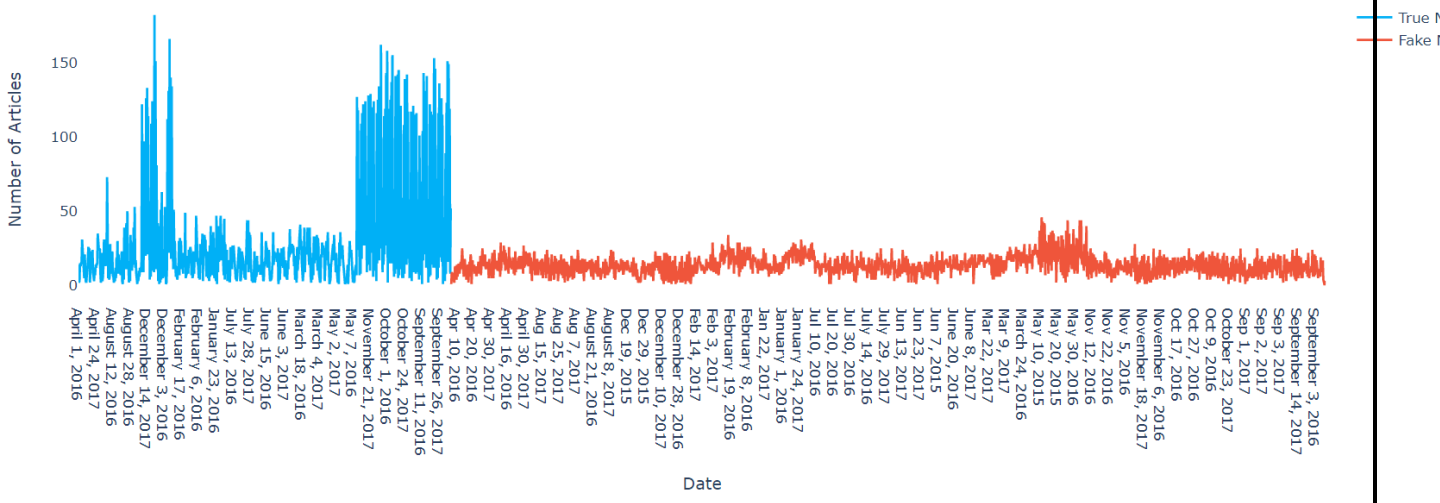
Categorical

Distinct count	6
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	183.6 KiB

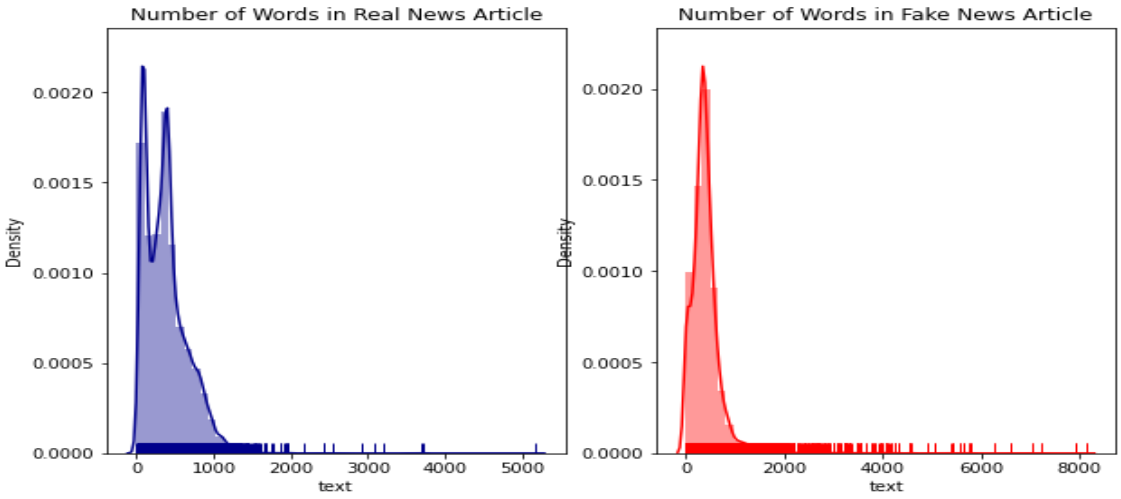
Missing Value Count



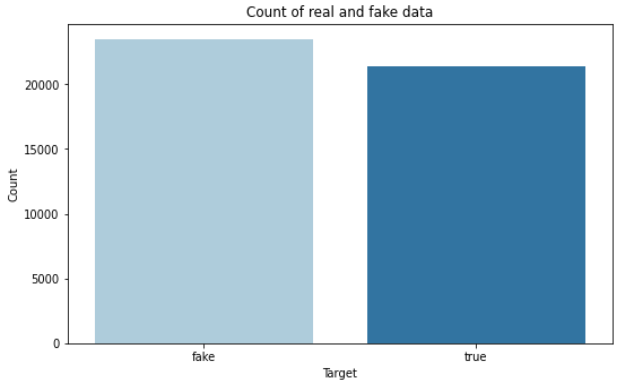
Number of Articles Published Daily

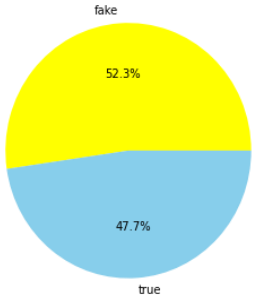


Number of Real & Fake Articles



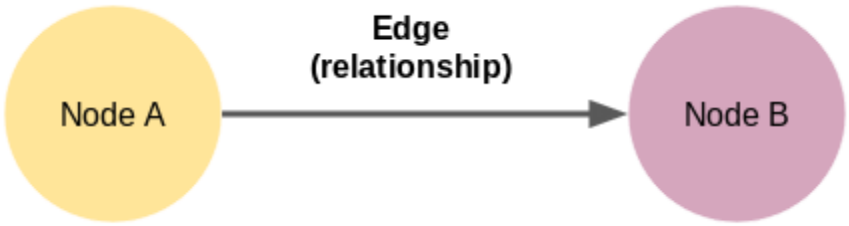
COUNT OF REAL AND FAKE DATA





KNOWLEDGE GRAPH

We can define a graph as a set of nodes and edges.



Node A and Node B here are two different entities. These nodes are connected by an edge that represents the relationship between the two nodes. Now, this is the smallest knowledge graph we can build – it is also known as a triple.

Knowledge Graph's come in a variety of shapes and sizes. For example, the knowledge graph of Wikidata had 59,910,568 nodes by October 2019.

How to represent a knowledge graph

Identifying the entities and the relation between them is not a difficult task for us. However, manually building a knowledge graph is not scalable. Nobody is going to go through thousands of documents and extract all the entities and the relations between them!

That's why machines are more suitable to perform this task as going through even hundreds or thousands of documents is child's play for them. But then there is another challenge – machines do not understand natural language. This is where Natural Language Processing (NLP) comes into the picture.

To build a knowledge graph from the text, it is important to make our machine understand natural language. This can be done by using NLP

techniques such as sentence segmentation, dependency parsing, parts of speech tagging, and entity recognition.

Entities Extraction Implementation

```
def get_entities(sent):
  ent1=""
  ent2=""

  prev_tok_dep=""
  prev_tok_text=""

  prefix=""
  modifier=""


  for tok in nlp(sent):

    if tok.dep_ !="punct":
      if tok.dep_=="compound":
        prefix=tok.text
        if prev_tok_dep=="compound":
          prefix=prev_tok_text + " "+ tok.text

      if tok.dep_.endswith("mod") == True:
        modifier = tok.text
        if prev_tok_dep == "compound":
          modifier = prev_tok_text + " "+ tok.text

      if tok.dep_.find("subj") == True:
        ent1 = modifier +" "+ prefix + " "+ tok.text
        prefix = ""
        modifier = ""
        prev_tok_dep = ""
        prev_tok_text = ""

      if tok.dep_.find("obj") == True:
        ent2 = modifier +" "+ prefix +" "+ tok.text

      prev_tok_dep = tok.dep_
```

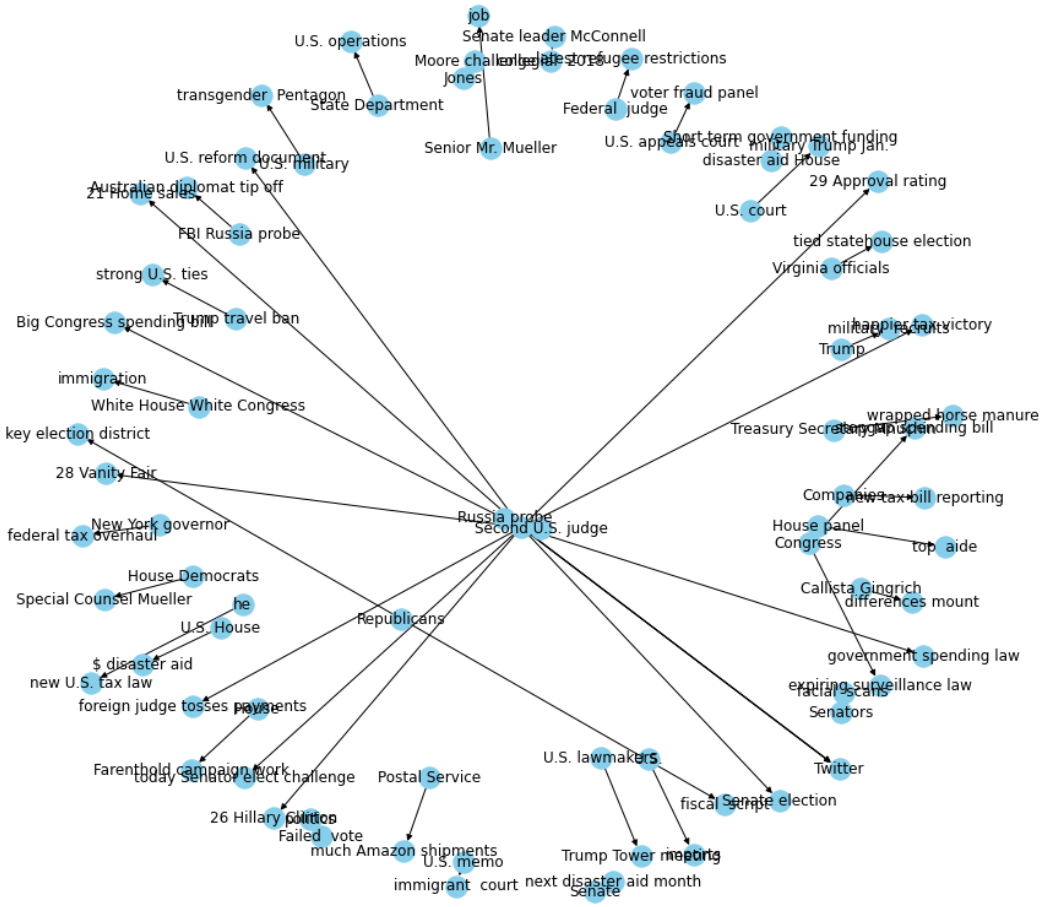
```
        prev_tok_text = tok.text

    return [ent1.strip(), ent2.strip()]
```

Entity Relation Implementation

```
def get_relation(sent):

    doc = nlp(sent)

    matcher = Matcher(nlp.vocab)

    pattern = [{'DEP':'ROOT'},
               {'DEP':'prep','OP':"?"},
               {'DEP':'agent','OP':"?"},
               {'POS':'ADJ','OP':"?"}]

    matcher.add("matching_1", None, pattern)

    matches = matcher(doc)
    k = len(matches) - 1

    span = doc[matches[k][1]:matches[k][2]]

    return(span.text)
```

Text Data Knowledge Graph Implementation

```
G=nx.from_pandas_edgelist(kg_df, "source", "target",
                          edge_attr=True, create_using=nx.MultiDiGraph())
plt.figure(figsize=(12,12))

pos = nx.spring_layout(G)
nx.draw(G, with_labels=True, node_color='skyblue', edge_cmap=plt.cm.Blues,
 pos = pos)
plt.show()
```



Future aspects of this project

- We can create a mobile application or a GUI to find out real time fake news anywhere easily.
- By adding image text recognition we can find out fake news from images (Quite popular these days).
- It can be used by social media developers to detect fake news when that's uploaded and take action accordingly.

SOURCE CODE

- Our GitHub Source Code Link
- Documentation of Project

https://github.com/cybertronayush/-Stop-Hoax-News-A-Fake-News-Detector-Using-Knowledge-Graph

REFERENCES

https://github.com/

https://www.udemy.com/

https://www.youtube.com/

https://medium.com/

https://www.analyticsvidhya.com/blog/

https://www.sciencedirect.com/science/article/abs/pii/S1389041719301020

https://ieeexplore.ieee.org/abstract/document/6103760/

https://dl.acm.org/doi/abs/10.1145/1963405.1963500

https://link.springer.com/chapter/10.1007/978-3-319-49586-6_54

THANK YOU!

THANK YOU SO MUCH FOR READING REPORT AND CHECKING OUT OUR PROJECT!

HAVE A NICE DAY

REGADGS!