PROJECT REPORT ON VIRUDH A FAKE NEWS DETECTION SOLUTION

PREPARED BY
Akarshan Gandotra
(14CSU016)

PREPARED FOR
Department of Computer Science
School of Engineering and Technology
The NorthCap University
Gurugram, Haryana

APRIL 2018

ACKNOWLEDGEMENT

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.We are highly indebted to Dr. Shilpa Mahajan for her guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.We would like to express my gratitude towards member of Faculty of University for their kind co- operation and encouragement which help me in completion of this project.We would like to express my special gratitude and thanks to industry persons for giving us such attention and time.

Akarshan Gandotra

CHAPTER 1

INTRODUCTION

We are living in the second decade of 21st century, and in this era the technology has enabled us to share an information instantly and to a great audience very conveniently. This lead to the issue of increase circulation of the Fake news.

Fake news is a type of yellow journalism or propaganda that consists of deliberate misinformation or hoaxes spread via traditional print and broadcast news media or online social media. It has the potential to influence the masses and change election results or spark a communal or cause some other event with severe consequences. The cases of fake news are increasing exponentially as shown in figure 1.

Fake news was not a regular term, but now it is now seen as one of the greatest threats to democracy.

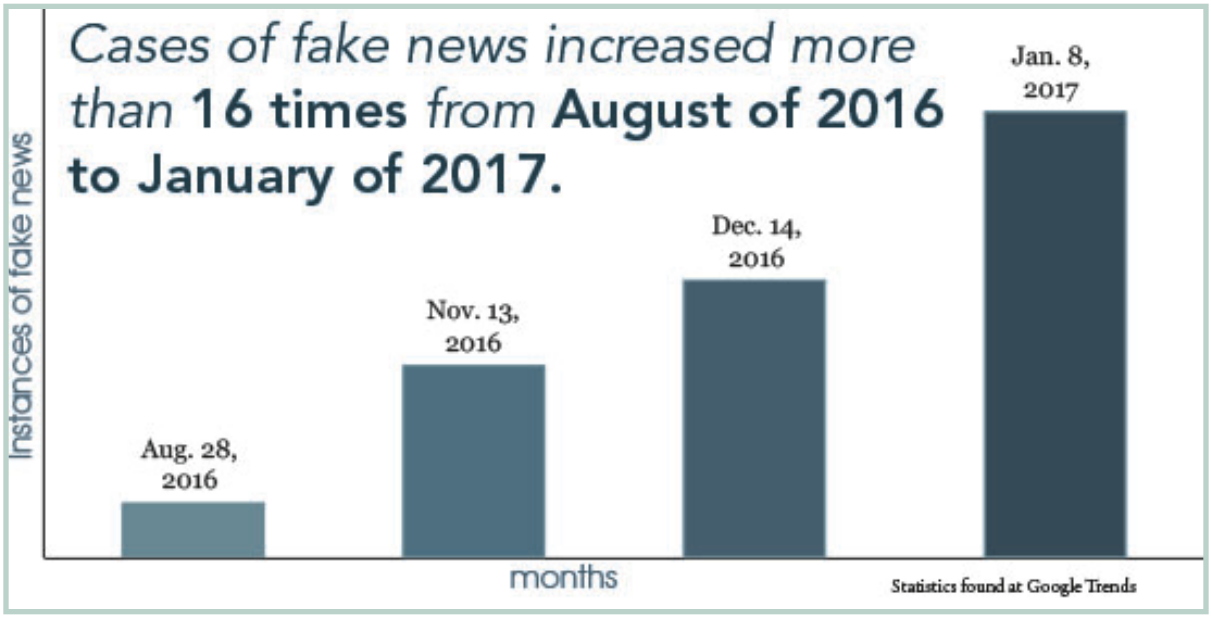
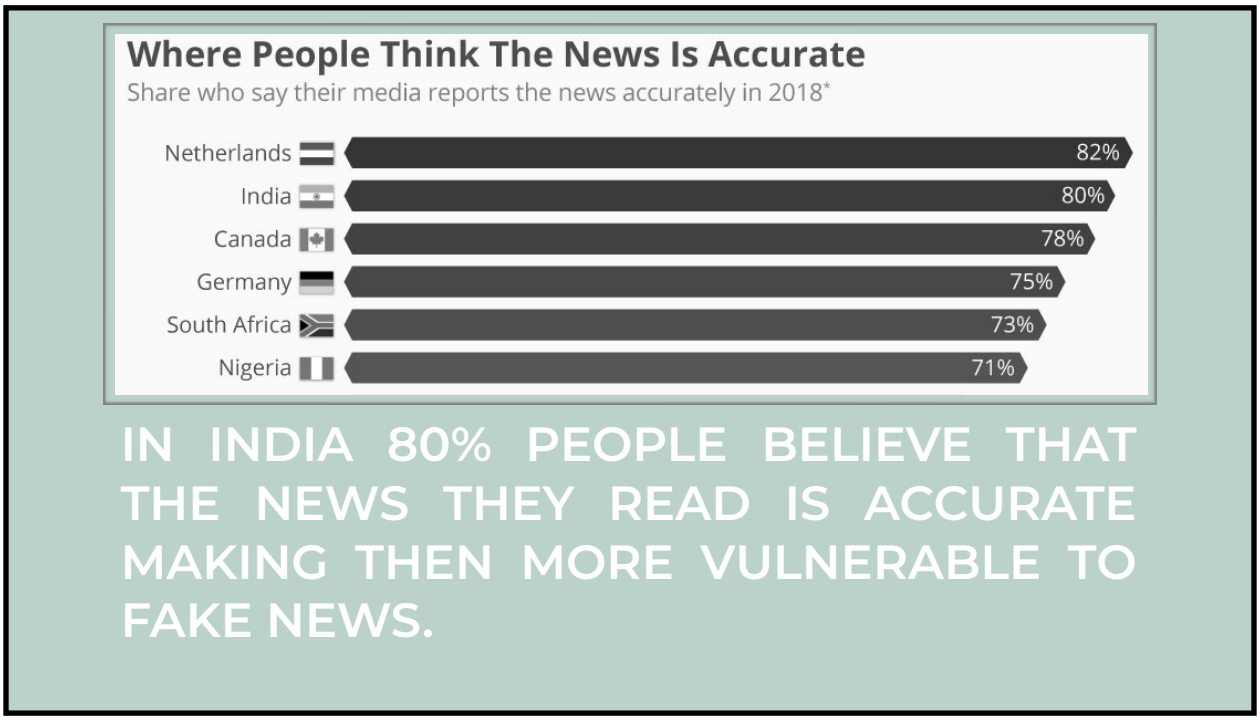


Fig1: Bar-Graph depicting monthly growth of Fake News



Fake news is written and published with the intent to mislead in order to damage an agency, entity, or person, and/or gain financially or politically, often using sensationalist, dishonest, or outright fabricated headlines to increase readership, online sharing, and Internet click revenue.

The above image depicts a majority of Indians believe the news to be true. The main sources of the fake news these days are social media platforms like Facebook, Whatsapp and other sources. Any misleading news can easily distributed on social media and become viral.

The underlying idea behind **Virudh/विरुध** is to help the people to allow the people to authenticate daily new they read and project them from believing

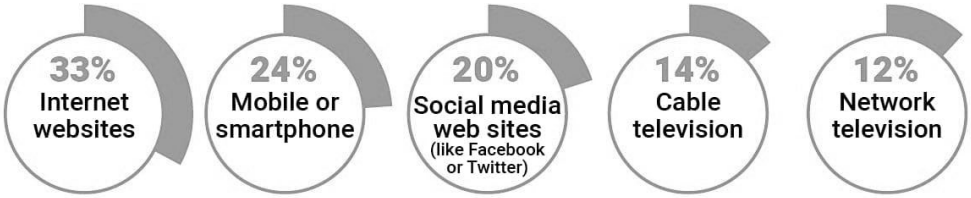


Fig3: Sources of News

any hoax or misinformation. Virudh is a comprehensive solution to classifies the fake news.
Virudh provides APIs which is integrated with the Android Application and performs various checks to ensure authenticity of the news. The APIs can be called from multiple interfaces or can be integrated into other Applications. Using **Industrial-Strength Natural Language Processing library, Spacy, micro web framework, Flask, Web Scraping techniques and Machine Learning,** Virudh is able to distinguish the news.

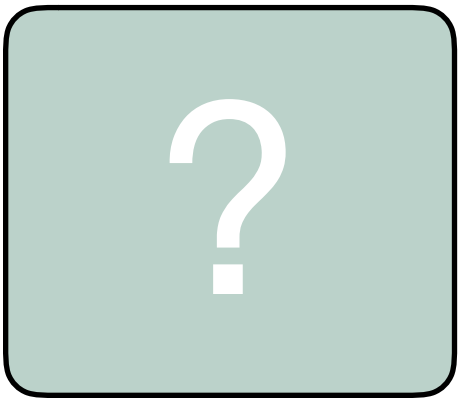
WE LIVE IN A TIME OF FAKE NEWS - THINGS ARE MADE UP AND MANUFACTURED.

- NEIL PARTNOV

CHAPTER 2

THE APPROACH

In order to understand the approach we need to discuss how we can spot a fake news. The following are the step help us to spot one.



Grammar Check

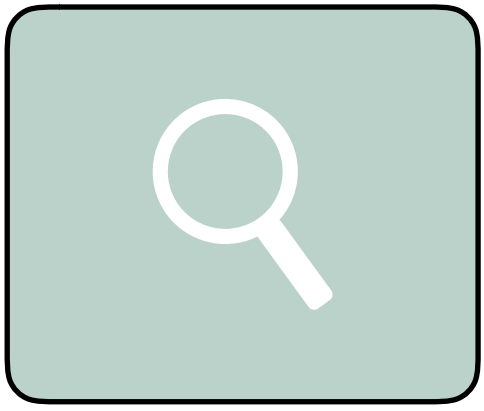
Most of the fake news have grammatical errors and spelling mistakes. News having many of such errors or mistakes can be fake.

Abnormality Check

Some times fake news can be unusual like written in all caps or have excess special characters. Hence, we can find a fake news based on such features.

Check the Source

The source of the news tells a lot about it's authenticity. A credible source ensures real news with accurate facts. If it's source is questionable or unknown, it is difficult to guarantee it's authenticity. And if the news originates from social media platforms puts a lot of question on it's originality and an investigation is a must. Also, checking author can also help to identify.



Check the Date

Reposting old news doesn't mean they are relevant to current events. Date can be a major parameter to determine the authenticity of the news. An irrelevant news from the past can still be damaging. Hence, date of the news also aid in the investigation of the fake news.

Search Engine Results

Simply by searching the news on web can tell a lot about it. If it is published by an authentic source, the results reflects it. In the later section we will discuss more about it. Also clicking on links or supporting sources mention in the news help to determine if information given actually supports the story.

Check the Biases

Fake news are often biased. In recent elections of the USA, the fake news circulated are often biased and this helped to manipulate the election results. Checking if the biased can help us to identify the authenticity of the news.

Past Experiences

Sometimes the news similar to the news we need to check has been circulated in the past later proved to be fake. Past experiences can make us alert and question the authenticity of the news.

Ask the Expert

Consulting the news with friends, experts and other sources can aid in spotting a fake news and also generate awareness.

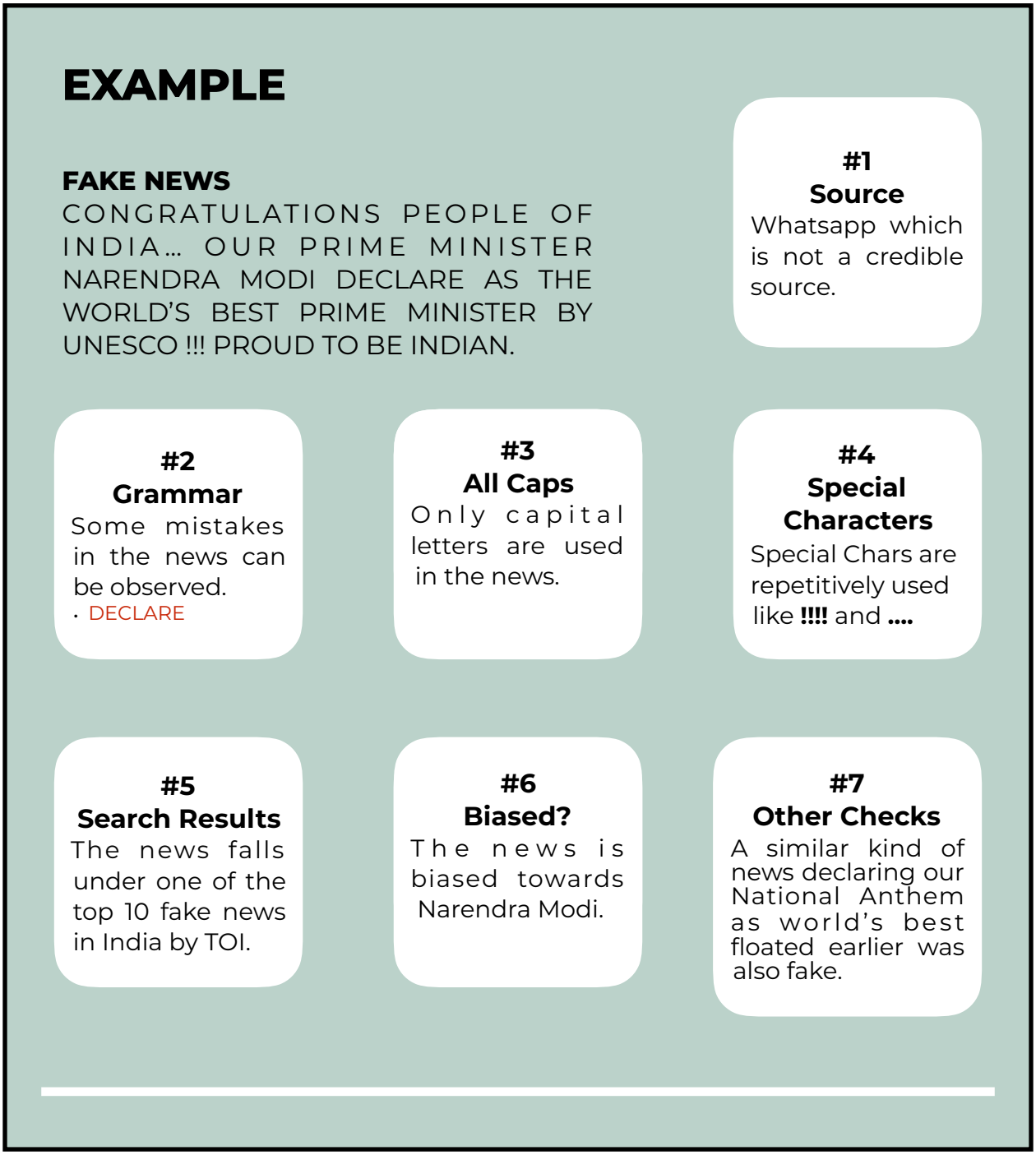


Fig 4: Example of various checks

We can follow these tests/checks to check a news. Virudh automates these tests and checks and provide results against each check just with a single click. These tests/checks are implemented in backend of Virudh. So, let's discuss more about the approach in which Virudh works.

In Virudh we have categorize the tests into two separate categories that are **Soft tests and Hard tests.**

Soft Tests

These tests are some basic level tests that highlight various characteristics associated with the news. Soft tests combine the results from the following:

- Grammar Test
- Important Words Extraction
- Capital Letter Percentage
- Special Character Percentage
- Repetitive Characters

Hard Tests

These are some advance level tests that aim to search the source and other parameters associated. Hard tests involve web scraping and other robust techniques to extract sources, date and the match percentage. Future sections will cover how both these tests are integrated.

The results of these tests are then saved in a database. An analysis on results and user feedbacks are recorded along with the results. Later when we get enough data, machine learning algorithms can be used to give predictive results and even more accurate results. The following figure is a flow of the used approach.

Fig5: The flow of the approach

CHAPTER 3

PROJECT IMPLEMENTATION

3.1 PROJECT INITIALIZATION

Before starting the project, the technologies, tools, techniques and skills that are required to develop Virudh are analyzed. Among various options available a detailed research was conducted in order to make a selection. So, we have decided to go with the following:

BACKEND

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language.

MongoDB is a free and open-source cross-platform document-oriented database program. It is NO SQL database

spaCy

spaCy is an open-source software library for advanced **Natural Language Processing**. It is used for **tokenization** and **Part-of-speech tagging**.

LanguageTool to performs the Grammar Checks. It is written in JAVA.

BS4

Beautiful Soup 4 is a Python package for parsing HTML and XML documents. It is used for web scrapping.

Flask is a micro web framework written in Python and based on the **Werkzeug** toolkit and **Jinja2** template engine. It is BSD licensed.

INTERFACE

Android is a mobile operating system developed by Google. An Android App provides an interface to interact with backend.

*** APIs can be integrated on a number of platforms and interfaces providing access to a larger audience.**

HOSTING

DigitalOcean calls its cloud servers **Droplets,** each Droplet you create is a new server for the use. Droplets are a scalable compute platform with add-on storage, security, and monitoring capabilities to easily run production applications.

After all the tools, techniques and skills are decided, the project was divided into modules and each module was developed independently. The modules are as follow:

#1
SOFT TESTS MODULE
Module dealing with development of Soft Tests.

#2
HARD TESTS MODULE
Module dealing with development of Hard Tests.

#3
API MODULE
Module dealing with the development of APIs in Flask Frame work.

#4
ANROID MODULE
Module dealing with development of Android Application.

Each module mentioned above was developed using an Iterative Waterfall model. A modular system is always convenient to develop.

3.2 PROJECT ARCHITECTURE

An architecture for the project is designed to have a clearer picture of interaction between different modules. Virudh's architecture is as follow:

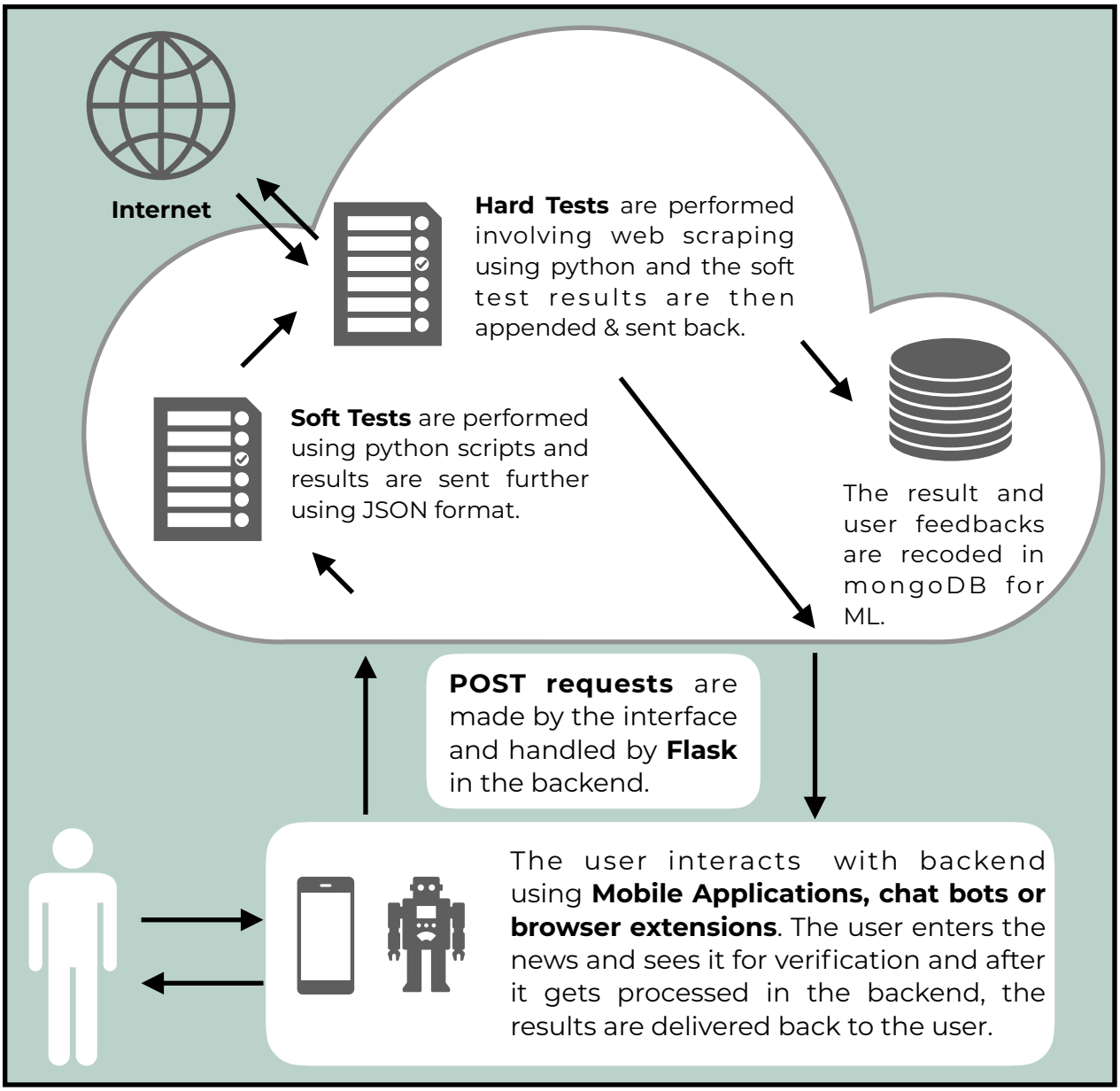


Fig 6: The Project Architecture

The user initially enters the news in the Android Application and sends to the server on a click of the button. A post request sending user input in JSON format is made (or an API is called). Once the server receives the data and methods related to tests or checks are called and results a re accumulated in JSON format and sent back to the from where the requests are made. The results are parsed and shown to the user. This is how Virudh works.

Let's discuss how the modules were developed.

3.3 SOFT TEST MODULE

Soft test modules consist of few elementary checks or tests. The scripts performing soft tests is written in python, The following shows the function called during soft tests associated with each news.

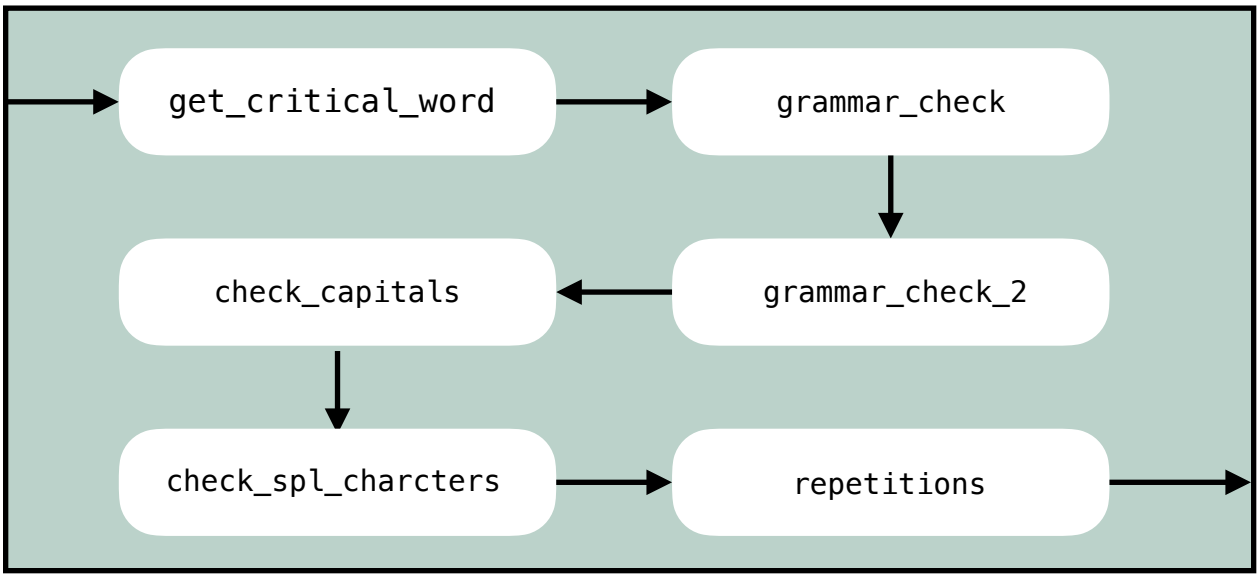


Fig 7: Functions of the Soft Tests Module

Every function is given an input and the output OR results are consolidated in the form of JSON. There are two functions for the checking the grammar. Both use different tools for the same. The thought behind it is that different tools have different rules of grammar, if one tool fails to notice an error, it can be picked up by another. Let's discuss and analyze each function one by one.

3.3.1 Critical Words Extraction

Spacy is an open-source software library for advanced **Natural Language Processing**. It is used for **tokenization** and **Part-of-speech tagging**.

Spacy is mainly used in this function. The news in String format is passed as the parameter. The space tokenizes the news and then **deep learning** algorithms perform **Parts of Speech Tagging**. After space loads en model, the tokenized words are iterated and **proper nouns** are extracted and appended in a list. The list is then sent back to the main.

The critical words have it's own importance. It tells the user about the context of the news or what the news is talking about. The maintained database of fake news also store critical words against each and every news it's store. Later when classification of the news is done, we can consolidate the critical words of all the stated news and find the most common word occurring in the fake nes.

A notification can be later sent making people aware of news containing those words. In a dataset from **Kraggle**, we have found that during 2017 America's election, J Donald Trump and Hilary Clinton were most commonly occurring words or nouns.

Hence, critical words can tell a lot about the news and track what are the most fake news is all about and get_critical_word function helps to get them.

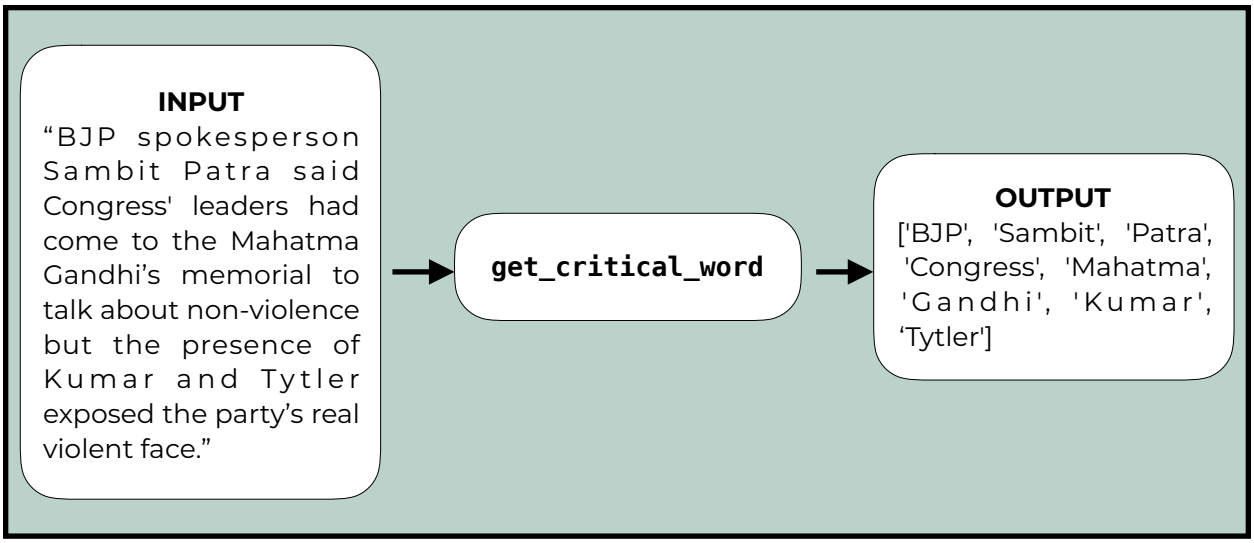


Fig 8: get_critical_word method

3.3.2 Grammar or Language check

A news from a genuine source or a known media house is always free of grammatical errors and spelling mistakes as the news is checked a number of times by highly professional and experienced people before getting published. A fake news distributor is generally more prone to make grammatical errors and spelling mistakes than a reputed media house.

This stage has two functions, **`grammar_check`** and **`grammar_check_2.`** For checking grammar, language spelling mistakes we need a **Grammarly** like tool. Unfortunately no library in python supports it. In JAVA, **LanguageTool** is available for similar kind of functionality. So, we developed a **python wrapper** for the same. For this we needed to setup JDK and JRE. A python wrapper is a binding to a JAVA library so that developers can use JAVA library using python code.

The other function **`grammar_check_2,`** on the other hand calls languagetool.org's check api for the same. A GET request is made (https://languagetool.org/api/v2/check?text=enterthetext&language=en-GB&enabledOnly=false) and response is generated by the server. The response returns incorrect results for the same. The results are in JSON. The JSON is the parsed and the results are extracted. The results are mapped and returned to the main.

The no. of grammar mistakes and spelling mistakes can be a powerful feature to classify fake news.

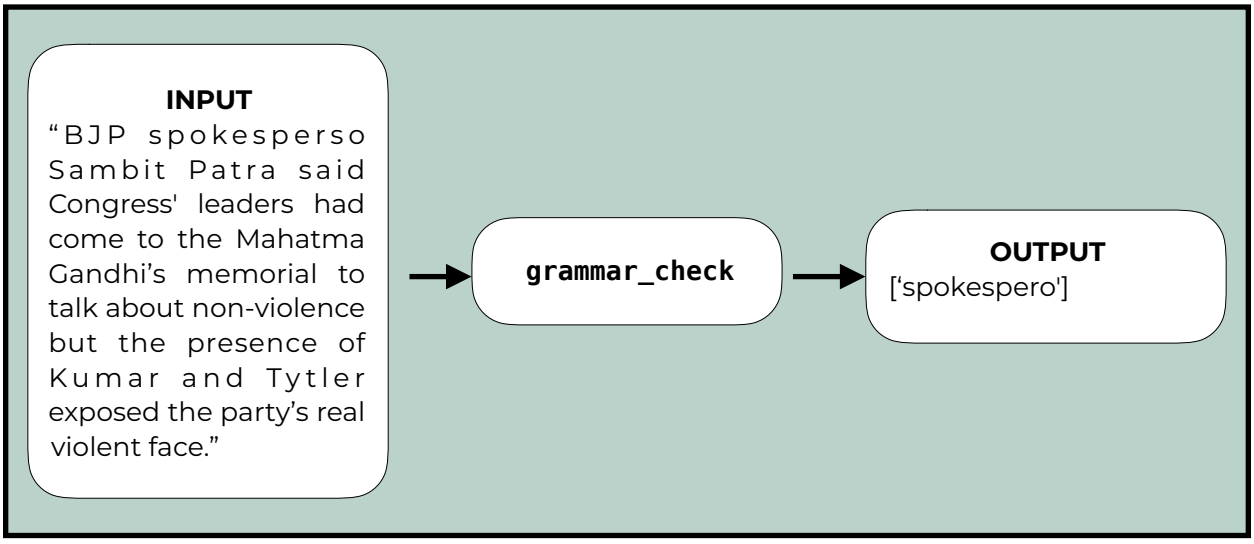


Fig 9: grammar_check method

3.3.3 Capital Letter check

Capital letters can be considered a third form of emphasis, among Italics and Bold text. They are used to denote a louder, almost shouting pronunciation. For example,
THIS TEXT HAS MORE IMPACT THAN this text.

Clearly, all caps part of the sentence is more impactful than the small characters and the fake news try to take advantage of this. Most fake news use only capital letters to give a more emphasis to the news or the message they are sending. The fake news often carry high percentage of all capital letters while a genuine news use them wherever required. This can turn to be a useful to detect the fake news.

A simple algorithm in python using **Regex** is written to detect the number of capital letter present in the news and gives it's percentage by dividing by total number characters multiplied by 100. It is one of the simplest to find but is very important feature to tell authenticity of the news. This percentage then sent back to main and consolidated in JSON.

Along with Grammar and other checks, this feature can prove to be a important decider to check whether the news is fake or not.

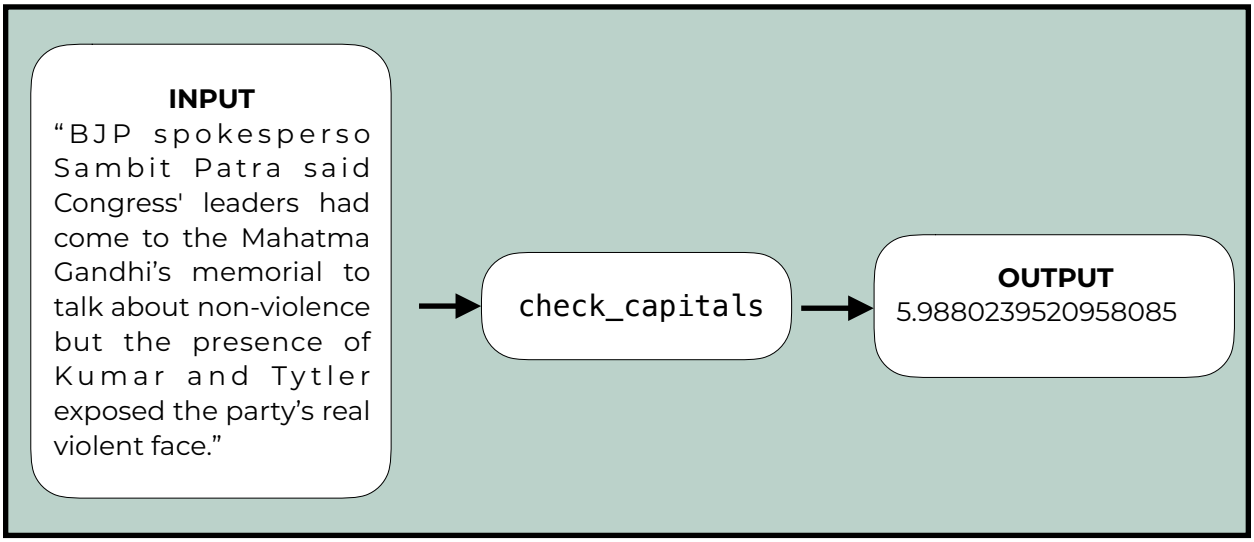


Fig 10: captial_check method

3.3.4 Special Character check

A fake news uses a number of special characters without any significance. As fake news is mostly written by unprofessional or not so educated people, they end up using a lot of special characters. The reason is similar to the use of capital letters, to create emphasis or impact.

The method associated with this test is similar to that of capital check. We used a simple algorithm in python using **Regex** is written to detect the number of special characters present in the news and gives it's percentage by dividing by total number characters multiplied by 100. It is one of the simplest to find but is very important feature to tell authenticity of the news. This percentage then sent back to main and consolidated in JSON.

Similar to capital letter check, this check is also used can prove as a important feature to distinguish the fake news.

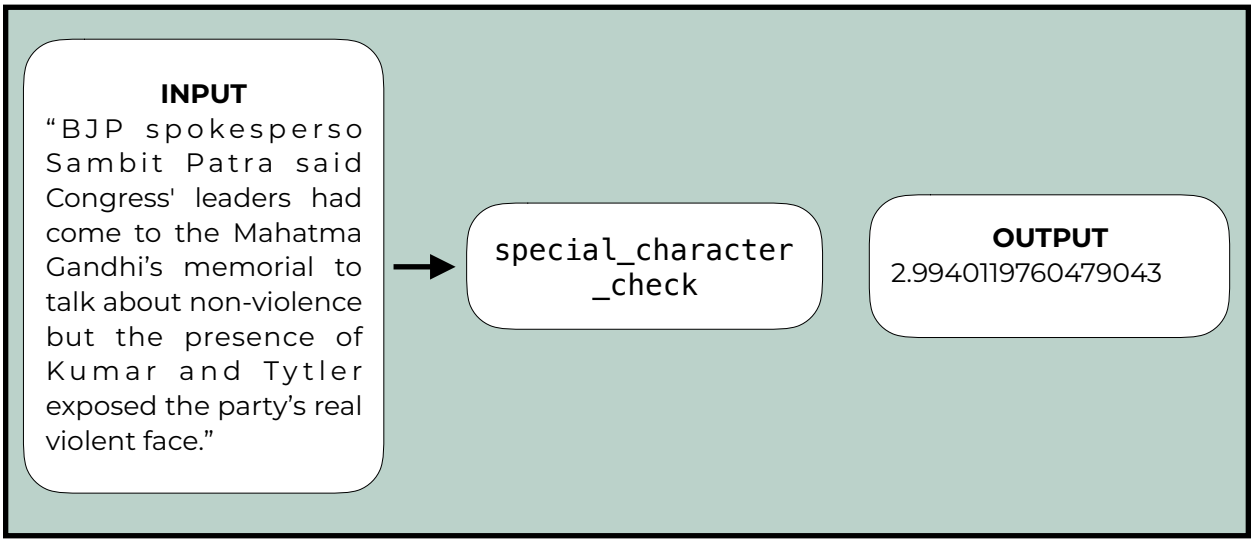


Fig 10: captial_check method

3.3.4 Repetition check

This is an extension of special_character_check method. Most of the fake news don't only use special characters but they tend to have them in repetitions. For example the use of **!!!!, ..., $$$$** is common.

So, we have created a method that can figure such findings. A simple **Regex** helped to get subsequent repetitions of any character, if this is greater than 2 then it is counted.

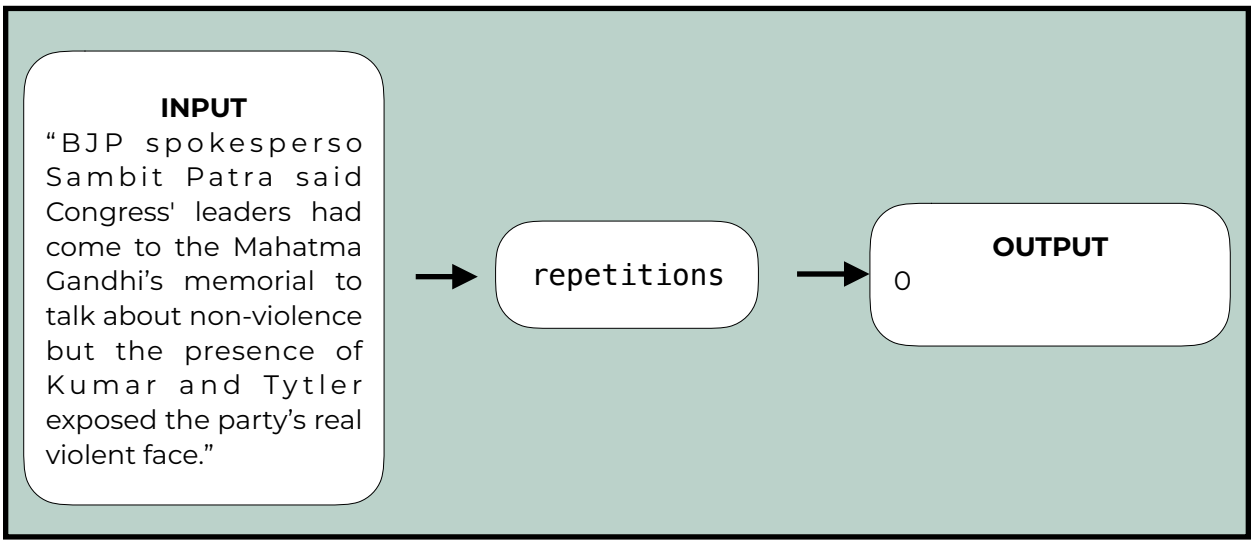


Fig 10: repetition method

3.3.5 Soft Tests Result Consolidation

In computing, JavaScript Object Notation or **JSON** is an open-standard file format that uses human-readable text to transmit data objects consisting of attribute–value pairs and array data types.
When exchanging data between a browser and a server, the data can only be text. JSON is text, and we can convert any JavaScript object into JSON, and send JSON to the server. We can also convert any JSON received from the server into JavaScript objects. This way we can work with the data as JavaScript objects, with no complicated parsing and translations.

Once all checks or tests are done, the results are consolidate in **JSON** format and will be returned as follow:

{'critical_words': ['BJP', 'Sambit', 'Patra', 'Congress', 'Mahatma', 'Gandhi', 'Kumar', 'Tytler'], 'incorrect_text_1': [], 'capital_percentage': 5.9880239520958085, 'repetitive_characters': 0, 'special_percentage': 2.9940119760479043}

This can enable us to share results easily with the applications or used by developers for API integration purposes.

3.4 HARD TEST MODULE

Hard tests are more vigorous test that help to identify the source of the news on the web. Another test it does is sentiment analysis. These test results contain more information and tells more if the news is fake or not. The following shows the function called during hard tests.

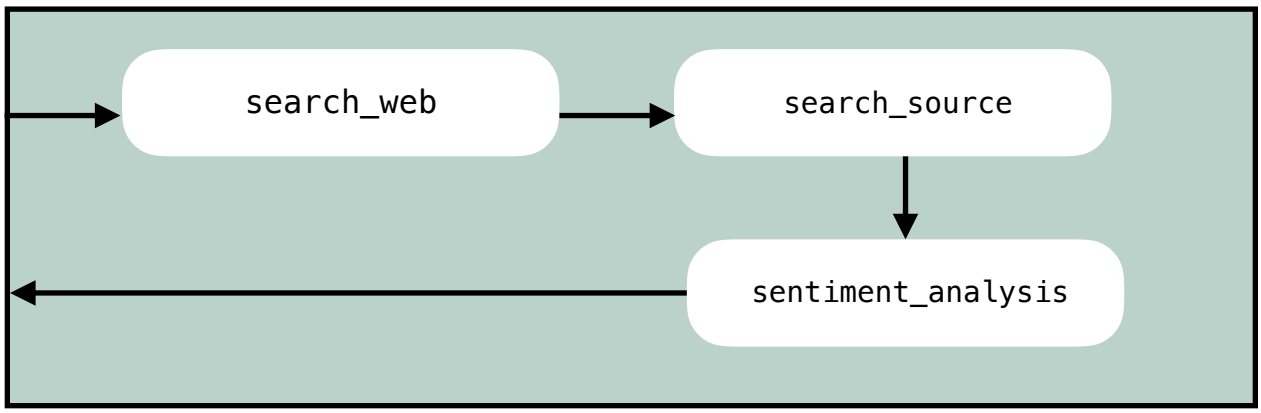


Fig 11: Functions of the Hard Tests Module

Every function is given an input and the output OR results are consolidated in the form of JSON. The search_web and search_source are methods to search the news and scan the source. While sentiment_analysis tells about the tone of the news. Let's discuss and analyze each function one by one.

3.4.1 Search the Source

Today most of the news companies have their presence over the internet. As the smartphone users increased most of them today use internet as their news feed and the users of the newspaper are declined sharply, it is depicted in figure 12. The news is on the internet as soon as it breaks. News sites share them on their websites, social media platforms and other places. This is taken as an advantage .
A genuine new will be published by a number of media sources while fake news only spread on not so popular websites, Facebook or Whatsapp or any other platform.

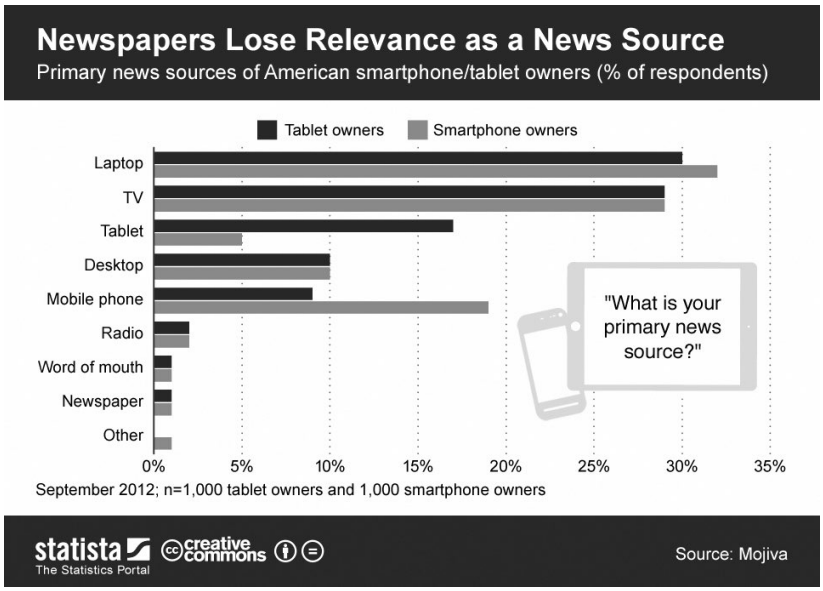


Fig 12: Popularity of news sources

In this method **web scrapping** is done on the search results. Initially is search query is sent to Google. The search query is initially **unicode encoded** as Google accepts unicode string as search query. For example

- **top 10 traveling places in Europe** gets converted into **top+10+traveling+places +in+Europe**
- **(Hello) ! &** gets converted into **%28Hello%29+%21+%26**

This is done by replacing the character of string with their equivalent unicodes.

Once a request is made we get a response. The response is in HTML. Now the html is parsed using **Beautiful Soup** library in python. BS4 is a Python package for parsing HTML and XML documents. After the parsing is done we extract the “a tags”. The a tags contains the link to the of best matches or results of the search query which in this case in news. Once the links are extracted, they can’t be used further. Some links are associated with AMP while others with Google cache pages and can’t be used further. For this we cleaned the links by splitting that parts in the link. The links are consolidated in a list and sent for further process.

This is how scraping of links takes place using BS4.

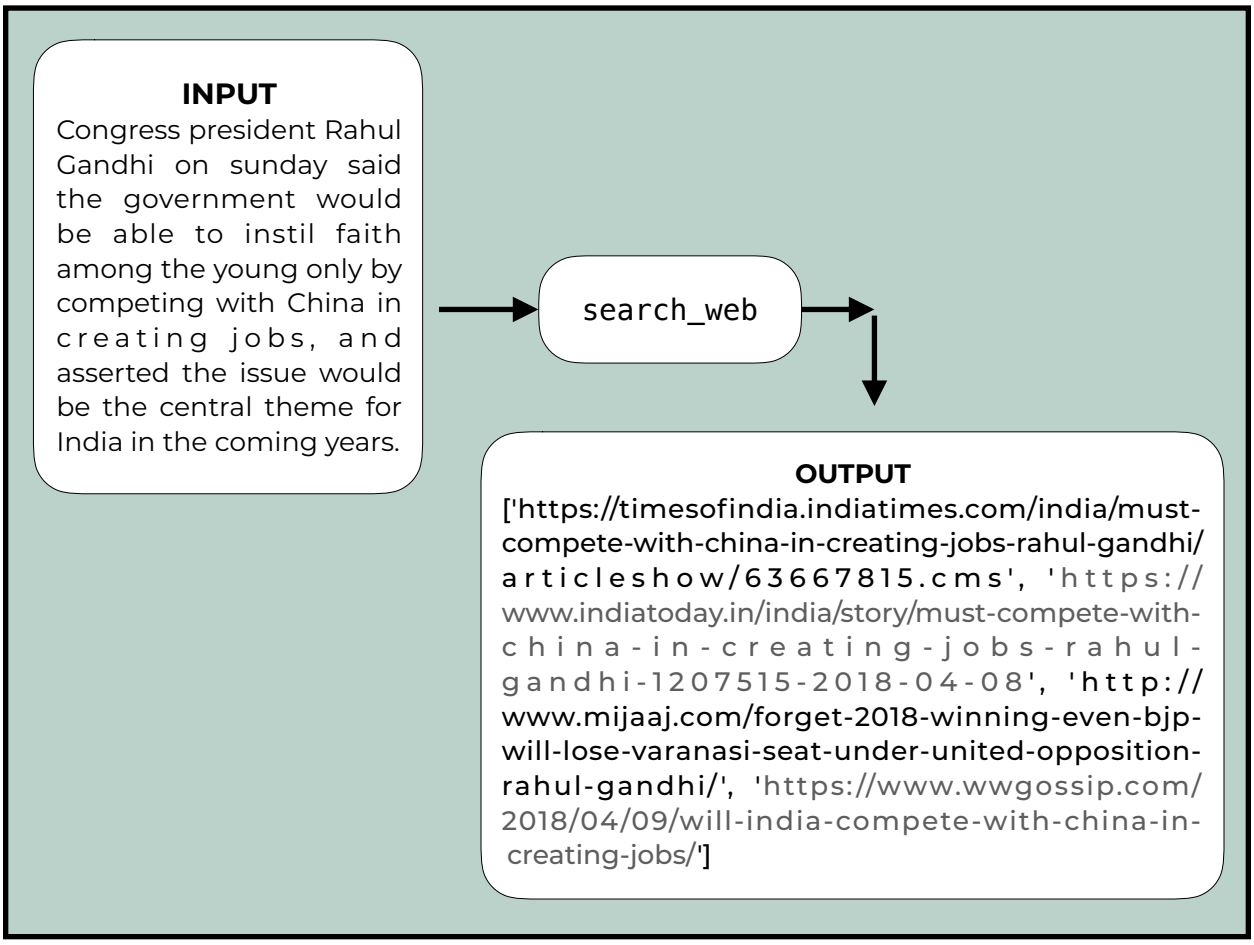


Fig 12: Popularity of news sources