A Scalable Approach for Linking News to Tweets

Kinjal Prakash Patel
Department Of Information Technology
Vidyavardhini's College of Engineering and Technology
Maharashtra, India

Laxa Devda
Department Of Information Technology
Vidyavardhini's College of Engineering and Technology
Maharashtra, India

Sejal Amipara
Department Of Information Technology
Vidyavardhini's College of Engineering and Technology
Maharashtra, India

Prof. Maryam Jawadwala
Asst. Professor
Department Of Information Technology
Vidyavardhini's College of Engineering and Technology
Maharashtra, India

***Abstract*— In today's world, people are more attached to Online Social Media. Online Social Media plays an important role in today's world for fetching the news about what is happening in the world. People know what fake news or biased news can do to someone or something and how some people take advantage of such situations and spread fake news which might instigate people and form wrong opinions. Twitter is an Online Social Media which has become popular in the recent years. People access Twitter widely around the world. According to recent observations, print media take up the news stories later after they are first broken into Twitter space as tweets. Many users post multiple tweets on Twitter which spread so fast and easy. These tweets can be accessed by any user who has registered themselves in Twitter. These tweets/stories posted may not always be right or truthful to make sense. Due to this reason, rumors have also been spreading like wildfire and people have believed them to be true. The proposed approach is an attempt to bring a halt to rumor-spreading by providing news links from trusted news sites related to that particular tweet. In this approach, first the text of the tweet is extracted by accessing Twitter API. From this text, keywords are found. For finding keywords, an algorithm is implemented using Natural Language Processing tool. Depending on these keywords, the news articles are collected from a trusted source and Natural Language Processing is done to find the most relevant news articles based on the semantic score. Later these articles are mapped to the tweet. This approach will help people broaden their perspective and understand the proper context and verify by themselves what is right and what is wrong.**

Keywords*— *Twitter, News Websites, Semantic analysis.

I. INTRODUCTION

Now-a-days, Online Social Media is easily accessible and faster than the Traditional media because of wider users on these platforms. Thus, Online Social Media has replaced Traditional Media like television channels, radio channels, etc. An advantage of Online Social Media is that all the people can share information and also gives their opinions on that platform[1].

Due to this reason, Twitter as an Online Social Media has evolved into a source of news for many users. It has become one of the most popular platform where common people share their information and views. People, especially youngsters, are turning to Twitter to seek information about emergency situations and daily events[3]. As many users post information on Twitter as soon as it happens, the information may not always be true. It becomes difficult to find out which information is correct and which is not.

In online platforms like Twitter, a large number of users are exposed to news instantaneously. They generally get swayed by the information and tend to believe the posted information to be true. As a consequence, people have believed many rumors to be true and also have retweeted them. One such example is about the rumor that Burj Khalifa was lit in Indian Tricolor when Prime Minister Narendra Modi visited UAE. Another example of such rumor left millions of users to believe that the US President Barack Obama was severely injured at The Boston Marathon Bombing[3].

Thus, it is very essential to differentiate between correct and false information. It is people's right to know whether the information they are seeing is trustworthy or not. It is of prime importance to search for faster ways to tackle this problem in order to cut down the wastage of revenue and resources spent by the government.[3].

The rest of the paper is organized as follows: The related work is discussed in Section II. Section III describes the comparative analysis of literature we have referred in section II. In Section IV, we describe our proposed approach. Implementation and results presented in Section V. Finally, we conclude in Section VI.

II. LITERATURE REVIEW

Many researches are being carried out to address the problem of false information circling in Social Media sites such as Twitter. In this section, some of the most prominent works in the field are discussed.

Dr. Dinesh B. Vaghela and Divya M. Patel in their paper *Rumor Detection with Twitter and News Channel Data Using Sentiment Analysis and Classification*[1] has proposed a method for detecting rumors on one of the social media i.e. Twitter. Their detection approach is divided into three parts: Preprocessing, Sentiment Analysis and Classification. First the real-time tweets are preprocessed to determine the topic of the tweet posted. Then tweets sentiment polarity is calculated using sentiment score which is then applied to different classification algorithms as an input. Using Twitter streaming API tweets are collected from twitter and are preprocessed to decide the features for classification.

Saumya Ahuja has discussed about obtaining valid news source via Twitter in her paper *Discovering Significant News Source in Twitter*[2]. Her proposed framework includes News concept generation using online news scrapping followed by extraction of tweets related to the news concept generated using REST and Streaming API of twitter. This relevant tweets are analyzed further for URL extraction to determine significant sources using a predefined criteria. In this research, the tweet collection has been performed over fresh keywords generated every half an hour using the news crawler. Hence, a dynamic approach has been implemented. Also, domains of URLs and not complete URLs have been used for analysis. Hence, the significant sources can be found out.

In *Towards Automated Real-Time Detection of Misinformation on Twitter by Suchita Jain, Vanya Sharma and Rishabh Kaushal*[3], misinformation (rumors) are detected in real-time. They have defined rumor as any information in Twitter space which is not in agreement with information from a credible source. According to them, a verified news channel account would publish more credible information rather than general users. Their approach is based on semantic and sentiment analysis to detect rumors. In order to validate their proposed algorithm, they have also implemented a prototype called The Twitter Grapevine which targets rumor detection in the Indian domain. The prototype shows how a user can leverage this implementation to monitor the detected rumors using activity timeline, maps and tweet feed. User can also report the rumor as incorrect which can then be updated after manual inspection.

Sercan and Erdogan, in their *'A Scalable Approach for Sentiment Analysis of Turkish Tweets and Linking Tweets to News'* [4] introduce and analyze method for linking news to tweets from twitter. News are collected from trusted news site. Data preprocessing steps are applied to the data before the data is sent for further processing by applying algorithms on it. In this model, system is developed using three major steps: Collecting news and tweets (2) mapping tweets to news (3) sentiment analysis on tweets. For mapping news to tweets, Bag of word strategy is used. Bag of word is a Natural Language Based algorithm in which text is considered as collection of words. News are collected from any mainstream media. This news are collected using RSS feed of any mainstream media which provide link to the news article in XML format. For collecting tweets, the keyword which are generated from the news are considered. For mapping tweets to news, methods are applied to remove stop word and only noun phrases are considered. Natural Language Processing tool is used to handle misspelled words, spell check, and for stemming. Sentiment score is calculated by comparing similarities between single tweet and all the news from the source.

Haewoon Kwak, Changhyun Lee,Hosung Park and Sue Moon in their paper *What is Twitter, a Social Network or a News Media?*[5] have obtained the certain twitter account, trending topics ,and certain tweets .In order to identified influential on twitter they have ranked user by number of follower and by Pagerank and found two ranking to be similar. Analysis of the collected tweet are done. Tweets are classified based on trending topics. A retweet is to reach 1000 user no matter how much the follower is once retweeted the tweet flows from 1st hop, 2nd and so on.

Marcelo Mendoza, Barbara Poblete and Carlos Castillo in their paper *Twitter Under Crisis: Can we trust what we RT?*[6] have discussed about the behavior of Twitter users under crisis. Particularly they analyzed the activity of Twitter users during 2010 earthquake in Chile. They characterized and analyzed the Twitter in hours and days following this disaster like number of tweets posted per hour in a day after the disaster or number of retweets based on the disaster. Further they performed the study on segregating false rumors and confirmed news. Their analysis showed that the tweets propagated corresponding to rumors are different from the tweets that spread news.

Soroush Vosoughi in his paper, *Automatic Detection and Verification of Rumors on Twitter*[7], have discussed about detecting the rumors on Twitter. As the misinformation spreads through any social media like Twitter very easily and rapidly, he proposed the approach for detecting and verifying the rumors on Twitter using classification and clustering technique. The rumors are verified based on 3 aspects like network propagation dynamics, group of people involved in propagating and most important the usage of linguistic style for expressing the rumor. Almost 209 rumors from more than 9,38,600 tweets that were collected from Twitter were tested using the verification algorithm including the rumors reported on popular websites. The algorithm works with an accuracy of 70%. These approach can be useful to minimize the spread of false information on Twitter.

In *"An introduction to twitter Data Analysis in Python"*[8] by *Vivek Wisdom and Rajat Gupta*, have approached the way to create twitter data which is ready to be used in various domain. Twitter data are extracted by Registering App and accessing data using python library. Accessed tweet are stored in JSON file that is used for further processing. Stop word which are not useful words are removed. Further tokenization of tweets using python library are done. Stopwords are removed using NLTK library which provide default stopwords for English language.

Tetsuro Takahashi and Nobuyuki Igata in their paper *Rumors Detection on twitter* [9] describe how the twitter is useful in situation and also has its negative byproduct ,spreading Rumors. This paper describe how the rumors have spread after the disaster and discuss how to deal with them. Investigation of the rumor after the disaster and then disclosed the characteristic of those rumors. A system which detects the rumors with acceptable accuracy from twitter is developed.

Weiwei Guo, Hao Li, Heng Ji and Mona Diab in their paper '*Linking Tweets to News*: *A Framework to Enrich Short Text Data in Social Media'*[10] have discussed about an effective way to apply Natural Language Processing to short texts like Twitter feeds. It is observed that a tweet usually covers a single aspect of an event and for the same they proposed a graph based latent variable model which is in contrast with the previous research based on lexical features. They used tweet specific feature as well as news specific feature including temporal constraints to extract text-to-text correlations that completes the semantic picture of a short text to which the WTMF (Weighted Textual Matrix Factorization) model is applied.

Jingjing Wang,Wenzhu Tong, Hongkun Yu, Min Li ,Xiuli Ma, Haoyan Cai, Tim Hanratty Jiawei Han in their paper "*Mining Multi-Aspect Reflection of News Events in Twitter: Discovering, Linking and Presentation"* [11] has proposed a framework to mine multi-aspect reflection of News event in Twitter. The aspects of event are linked to their reflection in Twitter which handles the challenge of selecting informative tweets under noise and bridging vocabularies of news and tweets. The framework also present the event with entity graph, time spans, news summaries and tweet highlights.

In "*Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web", Fabian Abel, Qi Gao ,Geert-Jan Houben, Ke Tao* [12] have created a semantic user profile for the social web. Method for linking twitter post to news article in order to contextualize twitter article strategies are propose and composed that exploit semantic extracted from tweets and related news article. The framework enrich the semantics of tweets clearly and have strong impact on the construction of semantic user profiles for the Social Web.

III. COMPARATIVE STUDY

Following table shows summary of the 4 most relevant papers which are referred for literature review:

Table 1. Papers referred for literature review

Sr. No.	Title	Technique Used	Observations
1.	Rumor Detection with Twitter and News Channel Data using Sentiment Analysis and Classification.	Sentiment analysis and classification techniques like Decision tree, Naive Bayes and Support Vector Machine.	Rumors are detected using sentiment analysis. Achieves high accuracy of Support Vector Machine than the other two.
2.	Discovering Significant News Sources in Twitter.	Significance score method.	News sources and concepts are analyzed and the most significant one is found.
3.	Towards Automated Real-Time Detection of Misinformation on Twitter.	Sentiment and Semantic analysis.	Used some example of rumors and according to verified news channel and general public tweets ratio it detects rumors by using sentiment and semantic analysis.
4.	A Scalable Approach for Sentiment Analysis of Turkish Tweets and Linking Tweets To News.	Naive Bayes algorithm, Complementary Naive Bayes algorithm, Logistic Regression algorithm, Bag-of-words method and Sentiment analysis.	Naive Bayes algorithm performs better than the other two algorithms. Unigram is a better feature model. Two-class dataset has better performance than three-class dataset.

IV. PROPOSED APPROACH

The system proposed consists of these major steps: (1) Collecting Tweets from Twitter API, (2) Collecting News feed from News Websites using News API and (3) Mapping News articles to Tweets.

A. *Collecting Tweets*

The collection of tweets is done using *tweepy* library which is a python library. Twitter Streaming API provided by twitter is used to access these tweets. The data which is collected is in XML format. An app is created that interacts with the Twitter API in order to have access to Twitter data programmatically. After registration of the app at https://developer.twitter.com, a consumer key and a consumer secret are received. An access token and an access token secret is also required. These strings must be kept private: they provide the application access to Twitter on behalf of user account.

Tweepy is one of the python library that can be installed using pip. Now in order to authorize the app to access Twitter on user's behalf, OAuth Interface is used. A convenient Cursor interface to iterate through different types of objects is provided through Tweepy. A maximum of 3200 tweets for extraction are allowed by Twitter. Tweepy is open-sourced library which is hosted on GitHub and it enables Python to communicate with Twitter platform and use its API. Figure 1 shows the screen after creating an app in Twitter console.

B. *Collecting News*

While there are many media outlets offer APIs, it is cumbersome to collect them individually. News API allows to search news articles and retrieve those live articles from all over the web. News API is a simple HTTP REST API which allows for searching news articles and retrieving those live articles from all over the web. NewsAPI.org is an easy-to-use API to get news from over 30,000 sources from

all over the world. This API is free for all projects that are non-commercial (including open-source) and in-development commercial projects.

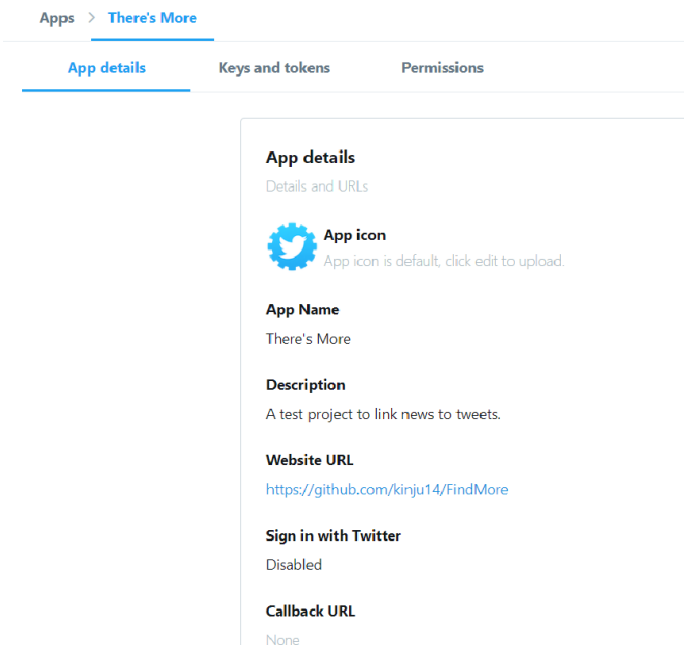

Fig. 1. Creation of App in Twitter Account.

An API key is needed to use the API - this is a unique key that identifies the requests. They're free for development, open-source, and non-commercial use. First the registration is done then the key is obtained. For registration, information has to be added by visiting the NewsAPI site which is newsapi.org. Figure 2 shows creating of API keys.

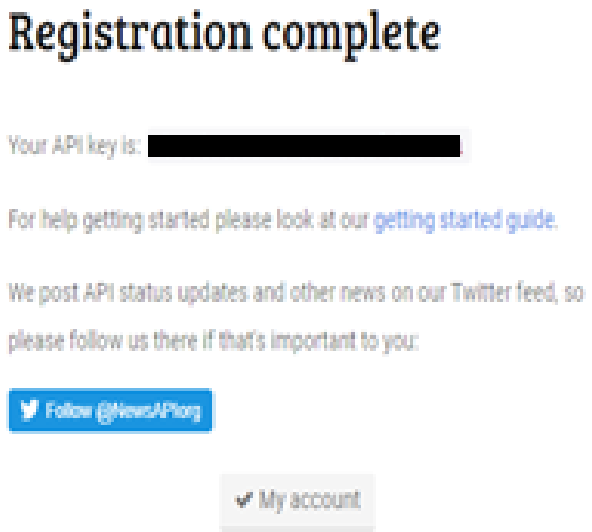

Fig. 2. Creation of API keys.

C. *Mapping of News articles to Tweet*

For performing pre-processing, NLTK is used. The Natural Language Toolkit (NLTK) is a platform or library which helps in building Python programs. These programs work with human language data for applying in statistical natural language processing (NLP). Natural Language Processing is nothing but manipulation of text/speech or understanding text/speech by any software or any machine. There is an analogy which states that humans interact, understand each other views, and respond with the appropriate answer. NLTK contains different text processing libraries that are used for tokenization, parsing, classification, stemming, tagging and semantic reasoning. NLTK also includes graphical demonstrations and sample data sets. It is also accompanied book which explains the principles behind the underlying language processing tasks that are supported by NLTK. NLTK is basically an open source library for the Python programming language. It was originally written by Steven Bird, Edward Loper and Ewan Klein for use in development and education.

The method wordtokenize() is used to split a sentence into words. A Data Frame is the converted form of the output of word tokenization. This Data Frame is created for better text understanding in machine learning applications. The Data Frame can also be provided as input for further text cleaning steps. These steps are processes such as punctuation removal, numeric character removal or stemming. Machine learning models need numeric data that can be trained and to make a prediction. Word tokenization becomes an important part of the text to numeric data conversion.

A kind of normalization for words is stemming. Normalization is a technique where a set of words in a sentence are converted into a sequence. This is done to shorten the sentence's lookup. The words which have the same meaning but have some variation according to the context or sentence are normalized. In another word, there is one root word, but there can be many variations of the same root word. Take an example of the root word "sit". It's variations are "sits, sitting, and like so". The root word of any variations can be found with the help of stemming. Stemming removes redundancy in the data and variations in the same word and hence is considered as an important

preprocessing step. As a result, the filtered data will help in better machine training.

An algorithmic process of finding the lemma of a word depending on the meaning of the word is lemmatization. The morphological analysis of words is usually referred to as lemmatization. It aims to remove or discard inectional ending. The base or dictionary form of a word which is known as the lemma of the word can be found with the help of Lemmatization. The NLTK Lemmatization method is based on built-in morph function which is provided by WorldNet. Stemming as well as lemmatization both are included in text preprocessing. Stemming algorithm works by cutting either the beginning or end of the word. In other words, it cuts the suffix from the word. In a broader sense.

D. System Models

The tweets from twitter are extracted and sent to the server side where processing will take place which includes extracting the entities from the tweet, sending back the news articles and performing the semantic analysis of the articles. Below is the flowchart of the proposed approach.

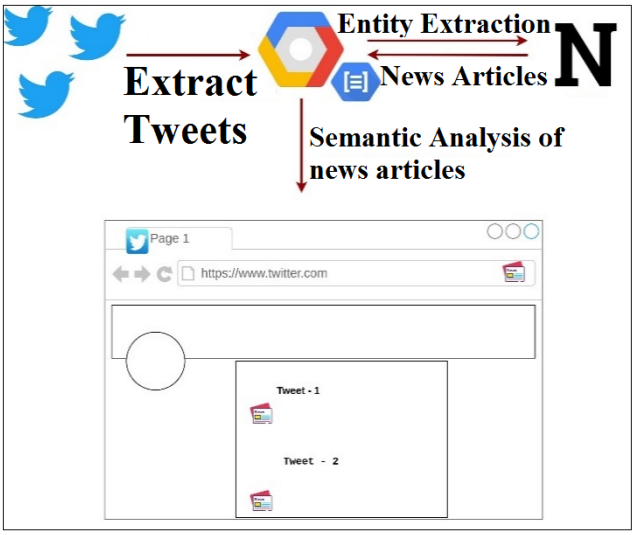

Fig. 3. Flow Diagram

The user selects the tweet and the Chrome extension sends the tweet for natural language processing. The related news are extracted and sentiment analysis of the articles is done. The analysed news sources are returned back. Below is the sequence diagram of the proposed approach.

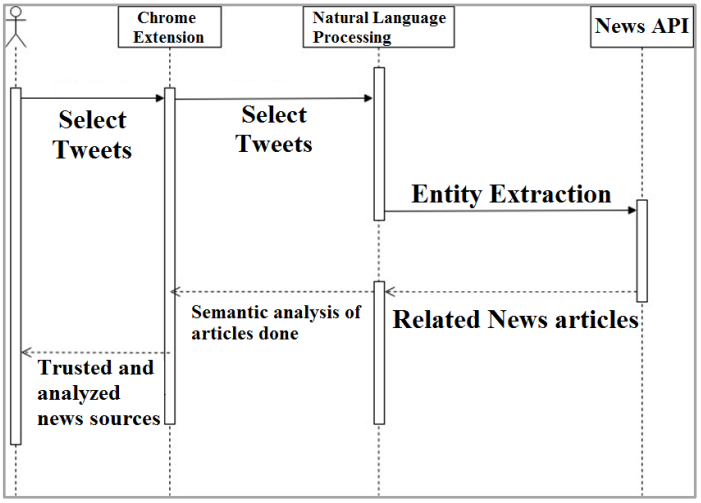

Fig. 4. Sequence Diagram.

V. IMPLEMENTATION AND RESULTS

A. Packages and technologies used

The programming is done in python language with the help of other packages and APIs. The collection of tweets has been implemented primarily using the Tweepy library supported by python. NLTK and Gensim is used for semantic analysis. News articles are collected from news API.

B. Results and Discussions

The implementation was able to retrieve live streaming tweets from Twitter API using the Tweepy library. Out of all the information collected from the tweets, only the ID and TEXT is retrieved and stored in an XML file. Below is the output of the tweets collected.

"id":1233278997816963072"text":"RT @realDonaldTrump: Obama just appointed an Ebola Czar with zero experience in the medical area and zero experience in infectious disease\u2026"
"id":1233278999708594177"text":"RT @kenanfikri: Mayor Pete #Buttigieg won 65% of Iowa's \"flipped\" counties--those that Obama carried twice before they plumped for Trump.\u2026"
"id":1233278999884812288"text":"RT @GravelInstitute: Ben Shapiro knows that Bernie is a committed Marxist
"id":1233279000346316801"text":"@feliperaytyson Pete es un Obama m\u00e1s
"id":1233279002388942850"text":"RT @w_terrence: Michelle Obama said Harvey Weinstein is a wonderful human being and her good friend.\n\nI didn\u2019t know wonderful human beings\u2026"
"id":1233279005861675008"text":"RT @ChrisMurphyCT: President Obama set up anti-pandemic programs in 47 vulnerable countries

Fig. 5. Output for tweets.

The news articles collected contain lots of data such as total number of results fetched from the API, source of the news, id of the news articles collected, author of the article, title or heading of the article, description or text of the article etc. A number of information is retrieved. This information is saved in a document for further analysis. Below diagram shows the output of the news articles retrieved.

{"status": "ok", "totalResults": 20, "articles": [{"source": {"id": "bbc-news", "name": "BBC News"}, "author": "BBC News", "title": "Iowa Caucus: \u2018We voted for Obama, then Trump\u2019", "description": "In 2016, Howard County in Iowa was the single largest pivot county in the country, voting for President Donald Trump after supporting Barack Obama in 2008 and 2012.", "url": "https://www.bbc.co.uk/news/av/world-us-canada-51330442/iowa-caucus-we-voted-for-obama-then-trump", "urlToImage": "https://ichef.bbci.co.uk/news/1024/branded_news/17D93/production/_110738679_p08225xp.jpg", "publishedAt": "2020-02-02T00:09:19Z", "content": null}, {"source": {"id": "bbc-news", "name": "BBC News"}, "author": "https://www.facebook.com/bbcnews", "title": "Russia meddling to help Trump win re-election, US lawmakers hear", "description": "US intelligence agencies reportedly warned Congress of the alleged meddling in a meeting last week.", "url": "https://www.bbc.co.uk/news/world-us-canada-51582025", "urlToImage": "https://ichef.bbci.co.uk/news/1024/branded_news/F62A/production/_110981036_7048da6f-468d-4465-9898-d8a7ae4cf5fb.jpg"

Fig. 6. Output for news.

Further, the news that were retrieved are passed to perform tokenization, stopwords removal and lemmatization. In tokenization, all the sentences are broken down to an individual words. All the stopwords are kept intact and not removed. Only the words in the sentences are separated. After tokenization is performed, stopwords are removed such as 'a', 'an', 'the', 'is', 'of', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', etc. It can be observed that all the stopwords have been removed from the text. After stopwor removal, lemmatization is performed in which the words are obtained in their original form. Below is the output of tokenization, stopword removal and lemmatization.

```
----------Tokenization----------
['``', 'id', "'", ':1233278997816963072', "'", 'text', "'", ':', "'", 'RT', '@', 'realDonaldTrump', ':', 'Obama', 'just', 'appointed', 'an', 'Ebola', 'Czar', 'with', 'zero', 'experience', 'in', 'the', 'medical', 'area', 'and', 'zero', 'experience', 'in', 'infectious', 'disease\\u2026', "'"]
----------Stop_Words_Removal----------
['``', 'id', "'", ':1233278997816963072', "'", 'text', "'", ':', "'", 'RT', '@', 'realDonaldTrump', ':', 'Obama', 'appointed', 'Ebola', 'Czar', 'zero', 'experience', 'medical', 'area', 'zero', 'experience', 'infectious', 'disease\\u2026', "'"]
----------Lemantization----------
[b'id/NN', b'text/NN', b'rt/NN', b'realdonaldtrump/NN', b'obama/NN', b'just/RB', b'appoint/VB', b'ebola/NN', b'czar/NN', b'experience/NN', b'medical/JJ', b'area/NN', b'experience/NN', b'infectious/JJ', b'disease/NN']
----------Tokenization----------
['``', 'id', "'", ':1233278999708594177', "'", 'text', "'", ':', "'", 'RT', '@', 'kenanfikri', ':', 'Mayor', 'Pete', '#', 'Buttigieg', 'won', '65', '%', 'of', 'Iowa', "'s", '\\', "'", 'flipped\\', "'", 'counties', '--', 'those', 'that', 'Obama', 'carried', 'twice', 'before', 'they', 'plumped', 'for', 'Trump.\\u2026', "'"]
----------Stop_Words_Removal----------
['``', 'id', "'", ':1233278999708594177', "'", 'text', "'", ':', "'", 'RT', '@', 'kenanfikri', ':', 'Mayor', 'Pete', '#', 'Buttigieg', '65', '%', 'Iowa', "'s", '\\', "'", 'flipped\\', "'", 'counties', '--', 'Obama', 'carried', 'twice', 'plumped', 'Trump.\\u2026', "'"]
----------Lemantization----------
[b'id/NN', b'text/NN', b'rt/NN', b'kenanfikri/JJ', b'mayor/NN', b'pete/NN', b'buttigieg/NN', b'win/VB', b'iowa/NN', b'flip/VB', b'county/NN', b'obama/NN', b'carry/VB', b'twice/RB', b'plump/VB', b'trump/NN']
```

Fig. 7. Output after pre-processing.

III. CONCLUSION AND FUTURE WORK

The framework presented maps News articles to Tweets for people to know the credibility of the tweets. Live streaming tweets from the Twitter API using the Tweepy library were successfully collected and the text of the tweet was extracted. Also news feed from the news websites were collected.

For future work, the plan is to perform semantic and sentiment analysis to find the most relevant news articles and map them to the tweet. Also, a Chrome extension will be made which is integrated into the Twitter User Interface. When clicked it will extract the entities and provide news articles related to that topic.

REFERENCES

[1] Dr. Dinesh B. Vaghela and Divya M. Patel,"Rumor Detection with Twitter and News Channel Data Using Sentiment Analysis and Classification", Intl. Journal of Advance Engineering and Research Development, vol. 5, no. 2, 2018.

[2] Saumya Ahuja, "Discovering Significant News Sources in Twitter", Intl. Conference on Research in Computational Intelligence and Ccommunication Networks (ICRCICN), IEEE 2015.

[3] Suchita Jain, Vanya Sharma and Rishabh Kaushal, "Towards Automated Real-Time Detection of Misinformation on Twitter, Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE 2016.

[4] Sercan Kulcu and Erdogan Dogdu, "A Scalable Approach for Sentiment Analysis of Turkish Tweets and Linking Tweets to News', Semantic Computing (ICSC), Tenth International Conference on Semantic Computing, IEEE 2016.

[5] Kwak, H., Lee, C., Park, H., Moon, S. What Is Twitter, a Social Network or a News Media? In: 19th ACM International Conference on WWW, pp. 591-600. (2010).

[6] Mendoza, Marcelo, Barbara Poblete, and Carlos Castillo. Twitter Under Crisis: Can we trust what we RT?. Proceedings of the first workshop on social media analytics. ACM, 2010.

[7] Vosoughi Soroush. Automatic detection and verification of rumors on Twitter. Diss. Massachusetts Institute of Technology, 2015.

[8] Vivek Wisdom, Rajat Gupta. An introduction to Twitter data analysis in python published in ResearchGate Publications on September 2016.

[9] Tetsuro Takahashi, Nobuyuki Igata. Rumor detection on Twitter. Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012.

[10] Weiwei Guo, Hao Li, Heng Ji , Mona Diab. 'Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media'published in proceedings of the 50TH annual Meeting of the Association for Computer linguistics.

[11] Jingjing Wang,Wenzhu Tong, Hongkun Yu, Min Li ,Xiuli Ma, Haoyan Cai,Tim Hanratty,Jiawei Han. Mining Multi-Aspect Reflection of News Events in Twitter:Discovering, Linking and Presentation.on November 2015.

[12] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web," in The Semanic Web: Research and Applications. Springer, 2011, pp. 375–389.