International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

Fake News Detection using Text Similarity Approach

S D Samantaray[1], Geetika Jodhani[2]

[1]Professor, Department of Computer Engineering, GBPUAT Pantnagar,Uttarakhand, India

[2]Department of Computer Engineering, GBPUAT Pantnagar, Uttarakhand, India

Abstract: ***In this era of digitization, most of the peopleget news from internet and often it can be difficult to tell whether stories are credible ornot. Information overload and a general lack of understanding about how the internetworks by people has also contributed to an increase in fake news or hoax stories.Traditionally we got our news from trusted sources, journalists and media outlets that arerequired to follow strict codes of practice. However, the internet has enabled a completelynew way to publish, share and consume information and news with very little regulation oreditorial standards.Our aim is to develop an automatic fake news detection system foranalyzing the credibility of online news. So that the reader become aware about the newsthat is factually incorrect and optimized for sharing. News articles are nothing but a pieceof text. Hence, the proposed work divided into two subtasks; Text Analysis andPerformance Evaluation. Text analysis is for the transformation of text into numericalfeatures. These numerical features used for matching the similarity betweenqueried article and other articles. For articles similarity we have used hybrid of three textsimilarity approaches namely N gram (Character Based Similarity), TF*IDF (Term BasedSimilarity) and Cosine Similarity (Corpus Based Similarity). System tested for 100 news articles and analyzed that if more than three articles found to be similar with ≥0.70 matching value will result to truthiness of the input article.***

Keywords: Fake News N-Grams, TF*IDF, cosine similarity, character Based Similarity, corpus based Similarity, term Based Similarity, matching value

1. Introduction

In the Digital world, everything is going online. News media is also one of theexample, which is undergoing a sea change and moving drastically toward Digitization.Today, when we have Facebook, Twitter, and others, tons of social platforms, mediahouses are also getting online. Each house has its own website, Facebook pages, andTwitter accounts etc.In the recent years, online content has been playing a significant role in swayinguser's decisions and opinions. Opinions such as online reviews are the main source ofinformation for e-commerce customers to help with gaining insight into the products theyare planning to buy. Recently it has become apparent that opinion spam does not only existin product reviews and customers' feedback.Fake news and misleading articles is another form of opinion spam, which hasgained traction. Some of the biggest sources of spreading fake news or rumors are onlinemedia websites such as *The Times of India, The Hindu,* and other media outlet.The problem of fake news is not a new issue, detecting fake news believed to be a complextask given that humans tend to believe misleading information and the lack of control ofthe spread of fake content.Fake news has been getting more attention in the last couple of years, especiallysince the U. S. election in 2016. It is tough for humans to detect fake news. There is anargument that the only way for a person to identify fake news is to have a vast knowledgeof the covered topic. Even with the knowledge, it is considerably hard to identify if theinformation in the article is real or fake.The open nature of the web and social media in addition to the recent advance incomputer science simplify the process of creating and spreading fake news. Trend Micro, acyber-security company, analyzed hundreds of fake news services provider around theglobe. They reported that it is effortless to purchase one of those services. In fact,according to the report, it is much cheaper for politicians and political parties to use thoseservices to manipulate election outcomes and people opinions about certain topics.Detecting fake news believed to be a complex task given that they spread easilyusing social and online media and word of mouth. Many things you read online may appear to be true, often is not**.** Usually, these stories created to either influence people's views, push a political agenda or cause confusion and can oftenbe a profitable business for online publishers. Fake news stories can deceive people bylooking like trusted websites or using similar names and web addresses to reputable newsorganizations.This false information, mainly distributed by social media, but periodicallycirculated through mainstream media. Fake news is written and published with the intent to mislead in order to damage an agency, entity, or person, and/or gain financially orpolitically, often using sensationalist, dishonest, or outright fabricated news to increasereadership, online sharing, and Internet click revenue. Websites with fake contentfrequently use a combination of website spoofing and authentic news styling techniques tomislead and manipulate readers, creating their sites to closely resemble and operate likeauthentic news sites. Therefore, there should be a way to detect this type of online news in detecting false, misleading, and sensationalist news is necessary to preserve the spirit oftrue journalism. All details about the proposed framework are provided in Section 2, while the implementation details of the proposed system are reported in Section 3. Section 4 provides the results of the experimental analysis, verifying the effectiveness of the proposed approach. Finally, in Section 5 we draw some concluding remarks, as well as a discussion on the open challenges in the field.

2. Literature Survey

Ah-Hwee Tanpresented the text mining framework as consisting of twocomponents namely Text refining and

Knowledge distillation. Text refining transforms the unstructured document into an intermediate form and Knowledge distillation deducespattern or knowledge from the intermediate form. In addition, this paper also illustrated thetext mining products and application based on the text refining and knowledge distillation functions.

Fengxi Song *et al.* considered mainly the text representation factors namely "stop words removal", "word stemming", "indexing", "weighting" and "normalization" and the effectiveness of these factors to text classifier.

Wen Zhang *et al.*studied the comparison of TF*IDF, LSI and multi-word methods for text representation and examined their performance of information retrievaland text categorization on the Chinese and English document collection.The experimental results show that in text categorization, LSI method gave the best performance than other two methods. LSImethod also produced best performance in English information retrieval,but in case of Chinese information retrieval, TF*IDF gave the best performance than two other methods.As an overall outcome, they conclude that LSI method was favorable for both semanticand statistical quality.

ZakariaElberrichi and KarimaAbidi categorize the Arabic texts in three different mode. They are
i) bag of words representation
ii) N-grams representation
iii) Concepts representation

From the experimental result, it is clear that categorization based on concept is the betterway of representing the Arabic texts than the other two representation.

A.K. Abdul sahib *et al.*presented a graph based text representation method,namely dependency graph, in order to reduce sparsity and semantic problem in the textual document.

Rubin *et al.* proposed a model to identify satire and humour news articles. They examined and inspected Satirical news articles in mainly four domains, including civics, science, business, and what they called "soft news" ('entertainment/ gossip articles').

Horne *et al.* illustrated how obvious it is to distinguish between fake and honest articles. to their observations, fake news titles have fewer stop-words and nouns, while having more nouns and verbs.

Wang *et al.*(2017)introduced LIAR, a new dataset that can be used for automaticfake news detection. Though LIAR is considerably bigger in size, unlike other datasets,this dataset does not contain full articles; instead, it contains 12,800 manually labelledshort statements from politicalFact.com

3. Methodology and Proposed Algorithm

In the framework proposed, we target the detection of web pagesreporting a news (a web news) having fakeness. In particular, we consider the text related tosuch news and we want to verify the consistency between them and other textually similar ones related to the same topic. The rationalebehind this choice is that the original version (or a textually similar one) of fake text contained in the news would likely bepublished in another web article on the same topic/event. Thus, inthe proposed approach the web news is submitted to a system, whichis able to provide as output a number of web pages, each of themconcerning the same event *and*containing texts that have some textual relationship with the ones of the original link. In this section, we formulate the general model devised to carry out such analysis, while the actual implementation of each step is discussed in the next section. As a first step, we assume that for each web newsarticle$N$, we can extract a structure $\varphi_N$containing *textual metadata*(the title, the author, the body,the date of publication,list of keywordsand length of article)and a set$\rho_N$of*visual metadata* (theimages contained in the web page, list of movies).

Our general goal is then to identify a set $\boldsymbol{S_N}$of similar web news articles' URLson the web (fromdifferent online news media) using Google News Search Engine, wherefor each$N' \epsilon S_N$, both N *and* $N'$are related by meansof their articles' content (the body of the article),determined in the implementation phase.

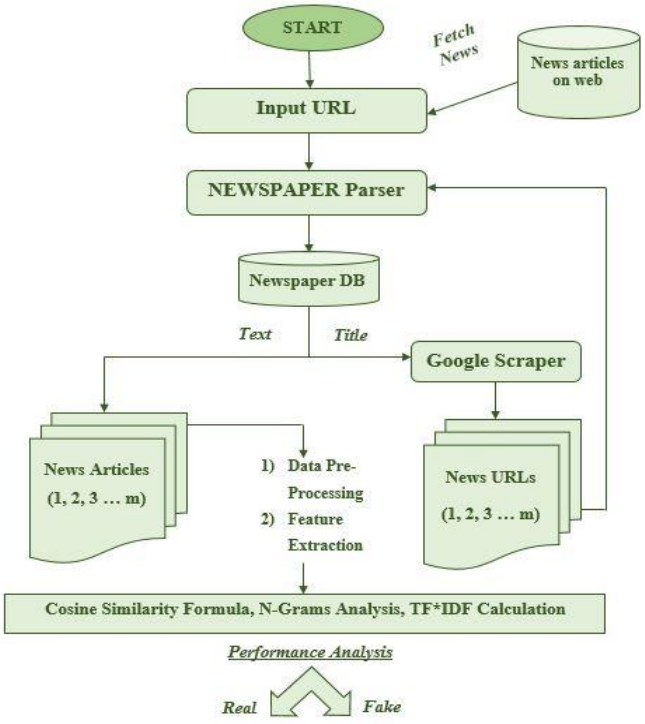

Figure 1: Implementation of the proposed model.

The system performs the following steps, which are summarizedin the scheme reported in Fig. 1:

Step 1: given the Input news N, φ_N and $\rho_N$are extracted.
Step 2: a first set $\tau_N$of links isidentified according to their similarity in terms of title of the article. The rationale behind thischoice is the fact that title based similarity will likely identify links that are strictly related to the very fact reported in the input news N and usually published the very same day and contributes to broaden and diversify the results by identifying the topic of thenews and providing links published in a wider temporalrange.

Step 3: a text similarity analysis isperformed between the body of web newsarticle N *and* N', where$N' \epsilon S_N$. This is done by means of functionMatchValue(.,.)such that two web news N *and* $N'$are considered exactly similar ifMatchValue(.,.) = 1and not similar if MatchValue(.,.) = 0 bymeans of suitable metrics described in the next section.
Once the similarity check is performed on Input article$N$ and other articlesN', the final set S_N is defined as follows,

$$S_N = \{N' \epsilon \tau_N , s.t. \text{MatchValue}(.,.)(N, N') \geq 0.70\}$$

Thus, $S_N$identifies all the web news textually related with the original news N. Finally, the user can visualize the original news article and,the set of similar articles are detectedamong the title related web news in τ_N. At this stage, for text similarity between queried news article and other similar articles suggesting the use of different text similarity approaches for a comprehensive analysis aimed at determining queried news article is fake or real. Moreover, we stress thatthe algorithm might also be applied recursively to each article in τ_N, in order to broaden the search and collect a higher number of text articlematches.

4. Implementation

Each of the phases described in Section 2could be implemented indifferent ways according to a specific rationale and technical needs.In this section, we describe tools and strategies we chose to usein this work in order to build an automatic system supporting theverification of online web news article. The whole process of detecting fake news detection can be decomposed into several phases.
In ***phase I***, we have used a Python module indicated as ***NewsParser,*** which takes as input the URL of a web news N and provides as output all information related to a newspaper like title, author, body, date of publication, keywords, all visual metadata (top image, other images and all movies) and filtered data. All these information are extracted via the *newspaper* Python module (availableonline[1])which are obtained by means of a simple tagger function based on the *Natural Language Toolkit (NLTK)* open source platform[2], (*referred in figure 2 and 3*).

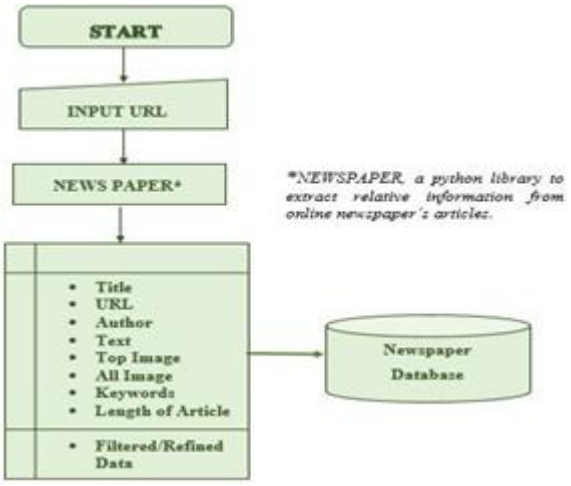

Figure 2: Flowchart of the Proposed Work (*Phase I*)

Figure 3: Detailed view of Filtered and Refined data

In ***phase II,***τ_N is considered as the set of links obtained by searching on Google N, respectively. This is done by means of *GoogleScraper*[3], a Pythonmodule available online which parses Google news search engine resultsand allow users to extract all the links found. *NewsParser*is then applied to each URL to obtain textual and visual datacontained in the webpage. All the data are stored and managed bymeans of a Python interface to SQLite3, (*referred in figure 4*).

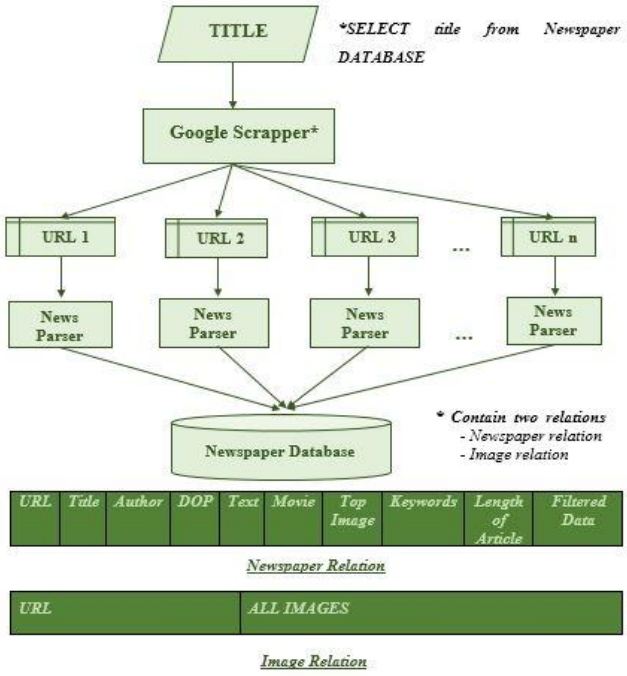

Figure 4: Flowchart of the Proposed Work (*Phase II*)

[1]https://github.com/codelucas/newspaper
[2]http://www.nltk.org
[3]https://github.com/NikolaiT/GoogleScraper

In ***phase III***, Then a set of title based similar news is identified (using *GoogleScraper*), the next task isdetermining whether two news articles can be considered as textually similar or not, which is a widely studied issue for which many optionsare available in the literature. However, for fake news detection problem, we decided to employ hybrid of three text similarity approaches: N-Gram (Character based similarity), TF*IDF (Term based similarity) and Cosine similarity (Corpus based similarity). In this choice, we were driven by the need of identifying a set of news articles, which are matched with input article. We exploited, Python 3.5.2 in *PyCharm Community 2018.1* Edition, which provides different packages and library tools for Text Based Similarity purpose.The outputvalues are properly combined and used todefine the function MatchValue(.,.)determining the similarity oftwo news articles that will be discussed in the next section, (*referred in figure 5*).

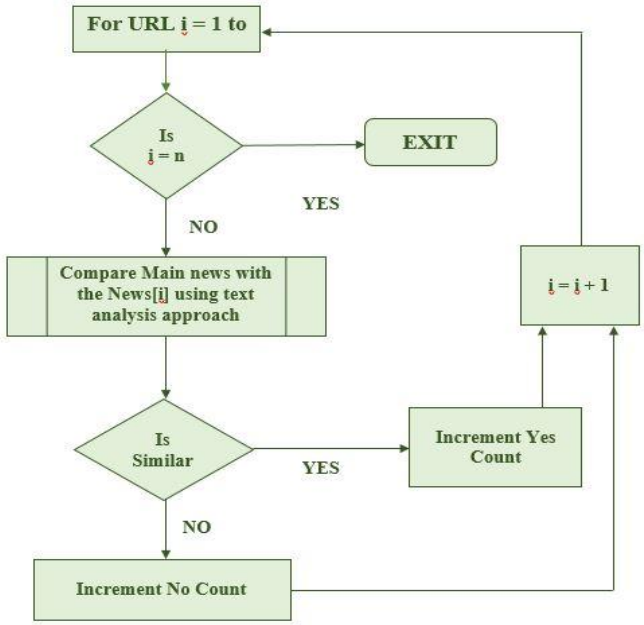

Figure 5: Flowchart of the Proposed Work (*Phase III*)

Measuring Similarity between Texts in Python

Example: Suppose there are three article with the text/body are,

Article 1: The game of life is a game of everlasting learning
Article 2: The unexamined life is not worth living
Article 3: Never stop learning

Let us imagine that you want to determine fakeness about the article with the text with following query: "**life learning".**
Let us go over each step in detail to see how it all works.

Step 1: Term Frequency (TF)

Term Frequency also known as TF measures the number of times a term (word) occurs in an article. Given below (*inTable 1)*are the terms and their frequency on each of the article considered above.

Table 1: Term Frequency Measures for All Articles

Article 1	The	Game	of	Life	is	A	everlasting	Learning
Term Frequency	1	2	2	1	1	1	1	1

Article 2	The	Unexamined	Life	is	not	Worth	Living
Term Frequency	1	1	1	1	1	1	1

Article 3	Never	Stop	Learning
Term Frequency	1	1	1

Step 2: Normalized Term Frequency

In reality, each article will be of different size. In large article, the frequency of the terms will be much higher than the smaller ones. Hence, we need to **normalize** the article based on its size. A simple trick is to divide the term frequency by the total number of terms. For example in Article 1, the term **game** occurs **two** times. The total number of terms in the article is **10**. Hence, the normalized term frequency is **2 / 10 = 0.2**. Given below*(in Table 2)* are the normalized term frequency for all the articles.

Table 2: Normalized Term Frequency Measures for All Articles

Article 1	The	Game	of	Life	Is	A	everlasting	Learning
Normalized TF	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.1

Article 2	The	Unexamined	life	is	Not	worth	Living
Normalized TF	0.01429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429

Article 3	Never	Stop	Learning
Normalized TF	0.333333	0.333333	0.333333

Step 3: Inverse Document Frequency (IDF)

The main purpose of doing a search is to find out relevantarticles matching with input query. In the first step, all terms are considered equally important. In fact, certain terms that occur too frequently have little power in determining the relevance. We need a way to **weigh down** the effects of too frequently occurring terms. In addition, the terms that occur less in the article can be more relevant. We need a way to **weigh up** the effects of less frequently occurring terms. Logarithmshelps us to solve this problem.

Let us compute IDF for the term **game**

$$IDF(\boldsymbol{game}) = 1 + log_e \left(\frac{Total\ Number\ of\ Articles}{Number\ of\ Articles\ with\ term\ \boldsymbol{game}\ in\ it}\right)$$

Total number of articles = 3
The term game appears in Article 1

$$IDF(\boldsymbol{game}) = 1 + log_e \left({}^{3}/_{1}\right) = 1 + 1.098726209 = 2.098726209$$

Given below*(in Table 3)* is the IDF for terms occurring in all the articles. Since the terms: **the, life, is, learning** occurs in 2 out of 3 articles they have a lower score compared to the other terms that appear in only one article.

Table 3: Inverse Document Frequency Measures for All Articles

Term	IDF
The	1.405507153
Game	2.098726209
Of	2.098726209
Life	1.405507153
Is	1.405507153
A	2.098726209
Everlasting	2.098726209
Learning	1.405507153
Unexamined	2.098726209
Not	2.098726209
Worth	2.098726209
Living	2.098726209
Never	2.098726209
Stop	2.098726209

Step 4: TF * IDF

Remember we are trying to find out relevant articles for the query: **"life learning"** For each term in the query, multiply its normalized term frequency with its IDF on each article. In Article 1 for the term **life,** the normalized term frequency is 0.1 and its IDF is 1.405507153. Multiplying them together, we get **0.140550715** (0.1 * 1.405507153).

Given below*(in Table 4)* is TF * IDF calculations for **life** and **learning** in all the articles.

Table 4: Term Frequency*Inverted Document Frequency Measures for terms in queried article in all the articles

	Article 1	Article 2	Article 3
Life	0.140550715	0.200786736	0
Learning	0.140550715	0	0.468502384

Step 5: Vector Space Model – Cosine Similarity
From each article, we derive a vector. The set of articles in a collection then is viewed as a set of vectors in a vector space. Each term will have its own axis. Using the formula given below we can find out the similarity between any two articles.

$$\textbf{\textit{Cosine Similarity}}\ (\boldsymbol{a1}, \boldsymbol{a2}) = \frac{Dot\ Product\ (a1, a2)}{\|a1\| \ * \ \|a2\|}$$

$$\textbf{\textit{Dot product}}\ (\boldsymbol{a1}, \boldsymbol{a2}) = a1[0] * a2[0] + a1[1] * a2[1] + a1[2] * \ldots + a1[n] * a2[n]$$

$$\|\boldsymbol{a1}\| = squre\ root\ (a1[0]^2 + a1[1]^2 + \cdots + a1[n]^2)$$

$$\|\boldsymbol{a2}\| = squre\ root\ (a2[0]^2 + a2[1]^2 + \cdots + a2[n]^2)$$

Vectors deals only with numbers. In this example, we are dealing with text articles. This was the reason why we used TF and IDF to convert text into numbers so that it can be represented by a vector.

The query entered by the user can also be represented as a vector. We have calculated the TF*IDF for the query*(in Table 5)*.

Table 5: Term Frequency*Inverted Document Frequency Measures for query article

	TF	IDF	TF*IDF
Life	0.5	1.405507153	0.702753576
Learning	0.5	1.405507153	0.702753576

Let us now calculate the cosine similarity of the query and Article 1.

$$\textbf{\textit{Cosine Similarity}}\ (\textbf{\textit{Query}}, \textbf{\textit{Article 1}}) = \frac{Dot\ product\ (Query, Article\ 1)}{\|Query\| * \|Article\ 1\|}$$

$$\textbf{\textit{Dot product}}\ (\textbf{\textit{Query}}, \textbf{\textit{Article 1}}) = ((0.702753576) * (0.140550715) + (0.702753576) * (0.140550715)) = 0.197545035151$$

$$\|\textbf{\textit{Query}}\| = squre\ root\ (0.702753576^2 + 0.702753576^2) = 0.99384368185$$

$$\|\textbf{\textit{Article 1}}\| = squre\ root\ (0.140550715^2 + 0.140550715^2) = 0.198768727354$$

$$\textbf{\textit{Cosine Similarity}}\ (\textbf{\textit{Query}}, \textbf{\textit{Article}}) = \frac{0.197545035151}{0.993843638185 * 0.198768727354} = \frac{0.197545035151}{0.197545035151} = 1$$

Given below *(in Table 3)* is the similarity scores for all the articles and the query **"life learning".**

Table 6: Cosine Similarity Angle for All Articles

	Article 1	Article 2	Article 3
Cosine Similarity	1	0.707106781	0.707106781

Article 1 has the highest score of 1. Because it contain both the terms life and learning in it.

5. Results and Discussion

In order to evaluate the effectiveness of our approach in revealing fakeness of queried web news article, we used three text similarity approaches.

At first, we carried out a preliminary testing phase aimed at properly defining the criteria for the selection of the web news collectedin τ_N, based on the similarity of their title with the oneof the original input link$N$. This step led to the identification of textually similar news articles among all the ones contained ineach element of τ_N. Subsequently, several web news were submitted to the system and the automatic retrieval was performed recursively on the elements of S_N, thus obtaining a more comprehensiveset of URLs. Finally, we considered we have taken some case studiesfor testing submitted to the proposed system,thus deeply investigating its capability of predicting fakeness of the Input news article. In this preliminary phase, we focused on the selection of news in$\tau_N$based on their textual content (body of the article) by defining the textsimilarity function MatchValue(.,.), so to derive the final set S_N. In particular, we run the algorithm for some web news articles obtaining$\tau_N$and our proposed system evaluate the performance of system by predicting the percentage of similarity between two news articles. Proposed system is tested for 100 news articles from different categoriesTo determine the fakeness of the Input article we have analyzed all case studies by taking different threshold Matching Value and analyzed that if more than three news articles are matched with input news article with Matching Value ≥ Threshold Matching Value(= 0.70), then input news article is considered to be Real otherwise Fake.
If three or less than three articles are matched with queried news article with threshold matching value then our queried article may be or may not be true because only two or three news media houses are updated the news. These news may be very new and no one is updated about this news except two or three media houses and it is real or it can also be possible that this news is false news and made intentionally to mislead reader and optimized for sharing. Therefore, we do not consider these news.
As illustrative examples, we considered case studiesspecified in Table 7.

Table 7: Web news considered and related URLs.

Case study	*Input News*
1.	***At Time Stamp:*** *{Mon Aug 13 13:12:29 2018}* ***"700 Pilgrims Stranded After Landslide In Uttarakhand Highway"*** *(https://www.ndtv.com/cities/700-pilgrims-stranded-after-landslide-in-uttarakhand-highway-1898360)*
2.	***At Time Stamp:*** *{Mon Jul 16 17:47:11 2018}* ***"U.S. firm Mercury LLC tried to enlist ambassadors to help Russian company"*** *(https://www.cbsnews.com/news/u-s-firm-mercury-llc-tried-to-enlist-ambassadors-to-help-russian-company/)*
3.	***At Time Stamp:*** *{Mon Jul 16 20:31:48 2018}* ***"CBI charge sheets Farooq Abdullah, 3 others in Rs. 43-crore cricket scam"*** *(https://www.thestatesman.com/india/cbi-chargesheets-*

	farooq-abdullah-3-others-rs-43-crore-cricket-scam-1502661988.html)
4.	***At Time Stamp:*** *{Mon Aug 13 10:33:56 2018}* ***"Kerala Floods: Centre announces Rs 100 cr immediate relief"*** *(http://www.dnaindia.com/india/report-kerala-floods-centre-announces-rs-100-cr-immediate-relief-2648971)*

For each news, a number of correlated links (title based similar), is retrieved by the system. Table 8 represents the matching value and matching angle *(in Radian)* between Input news *N*and other articles in τ_N obtained by the system for all the examples mentioned above *(in Table 7).*

Table 8: The matching value and matching angle obtained by the system for all the examples in Table 7

	Matching Value	Matching Angle *(in Radian)*
News 1	**1.** 0.9695 **2.** 0.9704 **3.** 0.9702 **4.** 0.9618 **5.** 0.9132 **6.** 0.9639	**1.** 14.19 **2.** 13.98 **3.** 14.02 **4.** 15.88 **5.** 24.05 **6.** 15.45
News 2	**1.** 0.9559 **2.** 0.9519 **3.** 0.9519 **4.** 0.9820 **5.** 0.9121	**1.** 17.07 **2.** 17.83 **3.** 16.97 **4.** 10.87 **5.** 24.19
News 3	**1.** 0.6624 **2.** 0.6829 **3.** 0.5313 **4.** 0.6568 **5.** 0.5479 **6.** 0.5868 **7.** 0.7100 **8.** 0.5199	**1.** 48.515 **2.** 46.921 **3.** 57.905 **4.** 48.939 **5.** 56.772 **6.** 54.065 **7.** 44.758 **8.** 58.669
News 4	**1.** 0.6579 **2.** 0.8142 **3.** 0.5718 **4.** 0.7875 **5.** 0.6180 **6.** 0.8860 **7.** 0.6815 **8.** 0.5219 **9.** 0.9048 **10.** 0.8124	**1.** 48.86 **2.** 35.49 **3.** 55.12 **4.** 38.04 **5.** 51.83 **6.** 27.62 **7.** 47.04 **8.** 58.54 **9.** 25.20 **10.** 35.67

Each article isthen matched with the news article of the queried webnews and if more than three articles have matching value greater than threshold matching value $(= 0.70)$ then the input/queried news is classified as real otherwise fake. Shown in Table 9 arethe number of retrieved correlated links, as well as the matched ones, for allthe four examples here reported.

Table 9: For all the four examples here considered, the number of URLs retrieved by the system, as well as the number of URLs matchedwith the original input one based on their similarity are reported.

News	No. of extracted URLs	No. of matched URLs
1.	6	6
2.	5	5
3.	8	1
4.	10	5

From table 9, News 1 when inputted to the proposed system, the number of extracted articles are 6. By applying text similarity approach the number of matched results obtained by proposed system are also 5 *(refer to Table 9)*. Thus, this news is finally predicted as True by proposed system. Similarly, in News 2 number of article extracted by our system is 5 and all the articles have MatchValue$(.,.) \geq$ Threshold Matching Value. So, this news is predicted as true by proposed system. Similarly in news 3, the number of matched article with Matching Value $\geq$ 070 (threshold matching value) is only 1 out of 8 extracted news articles *(refers in Table 9)*, which does not fulfil the matching criteria. So this news is predicted as *fake* by our system. In News 4, number of articles extracted are 10 and 5 of them are matched. So, this news is predicted as real by our system.

The performed matching served to build a reliable baseline needed for the definition and verification of a valuable automaticmetric for text similarity.

The similarity function MatchValue$(.,.)$ is hybrid of three text similarity methods (Character based, Term based and Corpus based similarity)are used which are ruled by a set of thresholds empiricallydetermined based on the preliminary tests on the selected news*(described in section 4)*.Despite the low number of news considered and the simple approachadopted, we verified that this was suitable for our preliminary analysis, although a deeper analysis in this respect would help in gainingaccuracy and will be subject of future work.

6. Summary and Conclusion

In the 21st century, the majority of the tasks are done online.In the digitalized era, growing problem of fakenews makes things more complicated and tries to changethe opinion and attitude of people towards digital technology. When a person is deceived by the real newstwo possible things can happen. People start believing that theiropinion about a particular news are true as assumed.Another problem is that even if there is any news articleavailable which contradicts a supposedly fake one, peoplebelieve in the words, which just support their thinking withouttaking in the measure the facts involved.Thus, in order to curb the situation, the proposed system for identifying the fake news in which each character and word of the input query article is matchedwith other similar news articles using hybrid of three text similarity approaches. It includes N-Grams (character based similarity algorithm), TF-IDF (term based similarity approach) and Cosine similarity (Corpus based similarity algorithm). The current study investigated fake news articles as a factor contributing to thecredibility of online news and found the accuracy of 91.67%.

7. Discussions and Future Directions

In this era of Facebook, Twitter, users encountered with hundreds of news articles, it will be a very tedious task for a user to copy and paste the news URL to know whether fake or not. A browser extension plugin can be developed which can detect fake news articles automatically from the web page they are visiting, without any extra work by users. It

will be more convenient and robust.The system can be extended by using more than one news scraper to neutralize the biasing of Scrapers available online. In our approach, we are basically focuses only on character/ word similarity in different news articles. Present system find similarity based on keywords, so there is need to develop a new layer of intelligence system to include intension-based decisions.In addition, database for new URL based field shall be created to be used for future researchers.

8. Acknowledgement

The authors would like to thank all for their valuable discussions and guidance. The authors would also like to thank the anonymous reviewers for their valued suggestions in improving the quality of this paper.

References

[1] N. Heise and J. Spangenberg, "News from the crowd: Grassroots and collaborative journalism in the digital age," in ACMWWW Companion, 2014, pp. 765–768.
[2] C. Castillo, M. Mendoza and B. Poblete, "Information Credibility on Twitter," inInternational World WideWeb Conference Committee (IW3C2), 2011, pp. 675-684.
[3] Chen, Y., Conroy, N.J., Rubin, V, "News in an Online World: The Need for anAutomatic Crap Detector," in The Proceedings of the Association for InformationScience and Technology Annual Meeting (ASIST2015), 2015, pp. 6-10.
[4] Chen, Y., Conroy, N.J., Rubin, V., "Misleading Online Content: RecognizingClickbait as "False News"," in ACM WMDD '15.
[5] Chung, C. J., Nam, Y. and Stefanone, M. A., "Exploring Online News Credibility:The Relative Influence of Traditional and Technological Factors" in Journal ofComputer-Mediated Communication 17: 171-186, 2012.
[6] Flanagin, A. J. and Metzger, M. J., "Perceptions of Internet Information Credibility" inJournal of Mass communication Quarterly, *2000*, Vol. 77, No. 3: 515-540.
[7] Fritch, J. W., and Cromwell, R. L., "Evaluating Internet resources: Identity,affiliation, and cognitive authority in a networked world" in Journal of the AmericanSociety for Information science and Technology, *2001*, 52(6): 499-507.
[8] Garrison, B., Driscoll, P., Saiwen M., Abdulla, R. and Casey, D., "The Credibilityof Newspapers, Television, and Online News" in Association for Education inJournalism and Mass Communication, annual convention, *2002*, pp. 1-30.
[9] Gupta, A. and Kumaraguru, P., "Credibility Ranking of Tweets during High ImpactEvents," ACM PSOSM 12.
[10] Horne, B. D. and Adali, S., "This Just In: Fake News Packs a Lot in Title, UsesSimpler, Repetitive Content in Text Body, More Similar to Satire than Real News,"The Second International Workshop on News and Public Opinion at ICWSM, 2017.
[11] Pasquini, C., Brunetta, C., Vinci, A. F., Conotter, V., &Boato, G., "Towards the verification of image integrity in online news," in IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2015, doi:10.1109/icmew.2015.7169801.
[12] Ordway, D. M., "Fake news and the spread of misinformation," journalistsresource.org/studies/society/internet/fake-news-conspiracy-theoriesjournalism-research.
[13] Rapoza, K., "Can 'Fake News' Impact The Stock Market?"forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market.
[14] Rubin, V. L, Chen, Y., and Conroy, N. J., "*Deception Detection for News: ThreeTypes of Fakes," in The Proceedings of the Association for Information Science andTechnology Annual Meeting (ASIST2015).*
[15] Rubin, V. L, Chen, Y., and Conroy, N. J., "Fake News or Truth?Using Satirical Cues to Detect Potentially Misleading News," in the Proceedings ofthe Association for Computational Linguistics: Human Language Technologies(NAACL CADD2016).
[16] Ruger S. M. and Gauch S. E., "Feature Reduction for Document Clustering andClassification," This workis partially supported by the EPSRC, UK, and the NationalScience Foundation CAREER Award 97-03307, USA.
[17] Schweiger, W., "Media Credibility - Experience or Image? A Survey on theCredibility of the World Wide Web in Germany in Comparison to Other Media,"European Journal of Communication, 2000, Vol. 15: 37-59.
[18] Sebastiani, F., "Machine Learning in Automated Text Categorization," in ACMComputing Surveys, 2002 Vol. 34, No. l. pp. 1-47.
[19] Tan A., "Text Mining: The state of the art and the challenges," in Proceedings of thePAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases.
[20] Viner, K., "How technology disrupted the truth,"theguardian.com/media/2016/juI/12/howtechnology-disrupted-the-truth.