Fake News Detection

Prateek Rawal
20MCA1110
Vellore institute of technology Chennai ,India
prateek.rawal2020@vitstudent.ac.in

Varenyam Gurjar
20MCA1107
Vellore institute of technology Chennai, India
varenyam2020@vitstudent.ac.in

***Abstract*—One can easily say in today's world, information aka news to few is more precious than money itself. This news needs to be in authentic form which is usually found in adulterated version. Leading us to have a dire need for an identification of real news from any possible fake news. News, being a form of information can be subjective to the proofs and source for its authenticity. As a human, one can easily identify real news from fake news with the help of one's innate capability to deduce logic and outlandish source of the information piece. Just that one needs few trusted sources to check for the facts and myths. But on a real time basis, there is a dire need for some software which can nip such 'false news' in its bud. Leading it to be one of the most researched area nowadays. Primarily being a part of Information Retrieval, this area is taking up a lot of attention from researchers worldwide to come up with a real-time solution for such an issue.**

In this article we have checked and analysed many research articles along with many survey articles and summed up this paper so as to provide the readers with a short idea of what fake news is, it's different flavours in the news spectrum, its characteristics and identification basic. We also included the different methods used by prior researchers in the same field. Using few researches as examples we learned about the basics of those methods used in fake news identification. The future aspects are also included in this article along with the challenges one faces while doing research in this very field.
Keywords—Fake News, survey, identification, real news, dataset, types of fake news,

I. INTRODUCTION

Fake News being the hype of world nowadays, in layman terms, refers to the intentional or unintentional spread of false information on public platform. One prime example for that can be given as, 'Indian 2000Rupee currency Bill came equipped with spying technology that tracked bills 120 meters below the earth', every person aware of the demonetization issue is well aware of this hoax spreading in early 2017s which actually took over the Indian Public. Digital Media users going in a panic about the height of digitization with a slight doubt of possibility that the news may be fake. At the time, this news seemed to be a plausible set of information considering the sudden demonetization along with Prime Minister Narendra Modiji's claims of ceasing the Black money from Indian Market and the height of digitization being pressurized in Indian environment, one could have been easily fooled by the ongoing news which later turned out to be a big hoax going around in the world of social media. Not only this, the problem of hoax news or fake news have engulfed almost all the spheres in this world. The main idea is to manipulate the emotions and thinking of humans to make them believe something that isn't true. And, the sources of such fake news include mainstream social media platforms including Facebook, WhatsApp, Twitter, Instagram etc . As per the reports of TechCrunch, WhatsApp has hit the 400 million monthly active users mark While earlier in 2017, Facebook had already crossed the 200 million users mark in India alone. The most mindboggling fact being that the mere people who are worried about the increasing growth of fake news in this society, are the majority who do not think twice about the authenticity of information before sharing it forward hence being a part of that influential group who create and spread fake news.

It needs to be pointed out that this is a universal problem which affects all the people around the world. And, it has also been there since the start of time. Though in earlier times as there was no access to worldwide information, the detection of fake news was relatively difficult and cost inefficient. But nowadays it is very easy, feasible and worthy to identify whether a news snippet is fake or real. Which brings us to the actual problem of analysing and identifying the fake news from real news in Gigabytes of information. It may be a mystifying fact that There are 2.5 quintillion bytes of data created each day at our current pace and this pace is nondecreasing overtime . Hence, to identify and eradicate fake news from so much data won't be a child's play. This needs real time, quick, feasible and cost friendly method. In today's world of digital era, Internet of Things and most importantly Artificial Intelligence, there can surely be much easier ways that can do the job, had it been just simple matching and comparing but as already mentioned this task requires much deeper knowledge of Literature, human behaviour, human speech, logics, possibility etc. This makes it a much worse situation to encounter. To summarise the existing concepts, we have wrapped those in section III after the types of fake news in section II, with some standard tools and datasets for use in section IV, the applications and future aspects for

fake news detection field in section V along with a conclusion in section VI to wrap up the paper.

II. Different Flavors of Fake News

One may wonder the emphasis paid on Fake news is far greater than the actual identification of the same. Well, its identification is its own poison. As one identifies fake news, one can contribute a part in clearing the world of that same news. For that one needs to understand the types of fake news. As specified earlier, there are many unknown number of fake news but coarsely there are two main types of fake news - Partly fake news : With a part of critical news missing from a snippet of information, and Fake news : With a part or full news being fabricated to provide with a whole new meaning to the news or just to spice up the news. The fake news universe is very vast and ephemeral and to some extent its directions are unknown yet there are incidences where many forms of misinformation are weaponized into fake news. Some of them are –

1. Satire – When deceptive packaged as a legitimate news story
2. Propaganda – When containing fabrications and packaged as a legitimate news story
3. Misleading or out-of-context information – When also serving as support for fabrication
4. Clickbait - When containing fabrications and packaged as a legitimate news story ,
5. Conspiracy Theory – When packaged as a legitimate news story .
6. Hoax - a humorous or malicious deception.

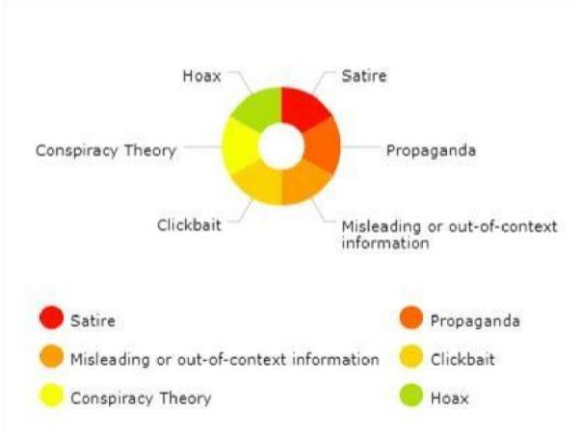

Fig. 1. Types of Fake News

One often confuses among these types of fake news, yet all types are harmful for human harmony and peace. Many researches have been conducted and many are still ongoing to identify and stop the spread of the same and balance the authenticity of information spread in real life. There have been researches that focus on one or two types of fake news, yet it won't be wrong to say that their research has been limited so far.

III. I dentification/Characteristics Of Fake News

A. Manual Fake News Detection The basic idea of doing anything comes from the realworld examples. Those ideas are enhanced and/ or modified to work efficiently in the area required to achieve solution to some problem. Similarly, the basic idea of getting a ang of fake news identification comes from simple human mind. There are many things that the mysterious and efficient human mind do. One of those is to catch the hidden emphasis in a set of information. That would work as the major criteria to confuse reak and fake news. There are many characteristics of fake news like -

1. Comes from an unfamiliar website
2. URL is strange or doesn't match the news outlet
3. Headline is outrageous or doesn't match article
4. May have an old date or doesn't have coverage
5. Doesn't list an author so one can't investigate them
6. Fails to provide proof of claim As seen from generic plausible examples, Fake news are created so that human brain can believe it. They are made such believable that one rarely doubts the news and go to look it up. For Example – If one says,

"Humans are able to breathe under liquids". It is fake news, though partly yet still fake. The actual news snippet being - "Humans are able to breathe under liquids like perfluorocarbons with certain qualities, which is called liquid breathing". One can be easily misled by the former statement assuming many things, like – one can breathe under water as well, or some more instruments are needed, etc. Though someone who is more interested in this area will research on this news and finally would put it under the fake news tab without some relevant information. If one looks the checking procedure of that someone, one can identify the basic steps to follow for identification of fake news. For the analysis of this news, a person has to check few points. Those are -

1. The news is from authentic source not a random one
2. Whether the headlines are relevant to its matter. Even a difference of question mark can create drastic misunderstandings
3. Check whether the author of the article is credible?
4. The sources supporting the claim of information ought to be valid along with the availability of the same information on a number of authentic sources.
5. One should be well aware that the news isn't any form of satire, sarcastic comment or a joke
6. Whether the logics and beliefs contradict with the judgement of the news

These are the few things to note yet need much involvement of human brain like logics and problemsolving skills that are pretty hard to integrate in an automated system. This leads us to how the automated systems or a software, per say might work if they are designed to retrieve real news from a stack of collective news.

B. A more automated approach

There have been numerous researches going around the world over the unique characteristics of fake news and how can one avoid them easily. More importantly much work is going on in building software that can stop the fake news from spreading or just notify the reader of the authenticity of a particular news. To tackle with different news various researchers have gone deep in literature to identify the statement creation and to divide it to understand whether the claim made by the statement is feasible or not? If one looks at the formation of a sentence one can understand that there are many components of a sentence like – subject, predicate, clauses and phrases each can be confused with another. Not to mention the 8 parts of speech - noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection. Also, the confusing tenses of verbs to signify the timings of the happenings.

Many researchers have used different methods in their research to come up with a certain solution for the problem. Some have used the n-grams, while some have used character based or word-based information retrieval [9]. While few of the researchers have took a completely different path and instead of spending their times in article, they have paid more attention to the images being fabricated to support the fake news mentioned in the articles [10]. As one can see, fake news is not only in the text of an article or news, but it is embedded there in every aspect of the articles. It is a mere gamble to decide the fakeness of an article just based on one aspect of the article. It is like a puzzle that one has to identify its each part. Every part of news needs to be verified to consider it as a real news while only one evidence of negative judgement is enough to declare it as a fake news. And by the virtue of Google one can easily check for the written statements in the news.

This can be taken in two parts – 1) To check whether the words used, the statement as a whole signifies all that information that is real and authentic with proof (content based) .

2) To check for any hidden agenda in the statement that might hinder the reader to acknowledge the real meaning of the news rendering it as fake (context based) .

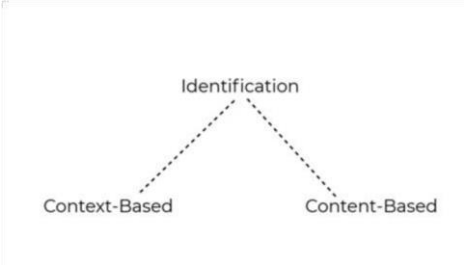

Fig. 2. Identification parts

The first point can be checked via comparing the existing news from some trusted and verified news sites. This uses information retrieval to fetch the news data remotely related to the news article headline and compare the news in consideration with that found on trusted sites. These trusted sites have built their ethics value after many years of honestly and perseverance keeping the customer's benefit in mind so they can be trusted to provide us with some worthwhile authentic information. One can use different techniques to compare the various parts of a sentence to verify the authenticity of news article.

The second part seems to be bit tricky where there should be some method to check the news meaning as a simple question can be turned into a well said sentence just by an altered punctuation. For Example – The news "Osama Bin Laden is Dead?" can be well changed as "Osama Bin Laden is Dead!!". In both the instances, the news meaning has changed even though their comparison after information retrieval from trusted sites may result in a positive feedback. This is the power of punctuation that can alter a statement's meaning.

In accordance for better understanding of existing methods for automated fake news detection, we have divided the many successful researches into different categories:

1. **Deep Learning –** The basic idea of using this method is taking use of the many aspects of deep learning concepts like Neural networks and Artificial intelligence. Though machine learning being an integrated part of deep learning, we have taken it as different method for better comparison. In one of the major researches that we checked for understanding this method by Gaurav Bhatt et al. defined how they have used deep learning to deduce the stance of any news to identify whether the news is real or fake by its stance. They have reckoned the neural embedding from the deep recurrent model, statistical features from the weighted ngram bag-of-words model and hand-crafted external features with the help of feature engineering heuristics And then they have combined all such features using deep neural layers and categorized the headline-body pairs in categories like agree, disagree, discuss and unrelated. The main limitation of this research being 3. that they have focused on stance than

words hence giving the people with illintentions a loophole.

Another research using similar approach is given by Sarah A. Alkhodair et al. is targeted at rumours in fake news. Their proposed model has used both types of learning from the data mining, that is, unsupervised learning to train a word2vec model and learn the word embedding while supervised learning to train a recurrent neural network model for rumour detection [13]. Here again the lacking has been their focus on only the rumour part of fake news spectrum.

In another research by Aswini Thota et al. have also used the deep learning architectures to devise a solution for fake news checking. They have used the dataset FNC-1 with contains the body and headline of a news article along with the label to signify the relation of the article with its headline. In this proposal, they have not considered the stop words and punctuation that holds as much importance as the article itself to identify any assertion from a question or any such type of statement.

2. **Machine Learning –** Though an integrated part of deep learning, many researches have used machine learning as a standalone method for deducing a solution to identify fake news from real one. The one research we used to understand this concept is by Georgios Gravanis et al. proposed a model for fake news detection using content based features and Machine Learning algorithms. They also tested for any improvements that reached with using concepts like adaboost and bagging.
Introducing a new text corpus, the "UNBiased" (UNB) dataset, with news source integration to avoid biased results in classification task. Their results show that the use of an enhanced linguistic feature set with word embeddings along with ensemble algorithms and Support Vector Machines (SVMs) is capable to classify fake news with high accuracy. The types of fake news targeted by them are - Hoax, Satire and Fake news posting. They have also achieved up to 95% accuracy for 5 datasets. Yet the striking limitations that were observed are that for one, the authors of news articles were not considered for its authenticity. Also, no social media diffusion features were included considering the fact that is where we get most of our fake news. The final result can be given as 95% over all datasets used with the AdaBoost to be first in rank and SVM & Bagging algorithms following them.

Others – This part includes all the other methods used for fake news detection by many researchers. Emphasizing on text similarity features or as one says
Natural Language Processing (NLP) , researchers S D Samantaray and Geetika Jodhani [9] have proposed a model for fake news detection by using three text similarity approaches. Of the two parts of the research, one being text analysis to convert the text to numerical features. These numerical features used for calculating the similarity between queried article and other articles. Three types of similarities used - N gram (Character Based Similarity), TF*IDF (Term Based Similarity) and Cosine Similarity (Corpus Based
Similarity). Second part being the analysis of that similarity to identify the truthness of an article. The limitation here would be the need for an intelligent layer that can include intentionbased decision so as to also include the context based checking of articles.

Another approach using analytics to identify fake news by Chaowei Zhang et al. have two phases for the research namely, detecting fake topics and fake events. Proposing a Fake News Detection System (FEND), the researchers have used analytics as their base. The main focus of this research being on distinguishing facts with opinion articles. One can say opinion articles are a kind of twisted truth or another perspective of truth. So, identifying it is also as important as identifying fake news. Hence, one can say few types of fake news not identifiable here like satire etc.

Research done by Eugenio Tacchini et al. in his research 'Some like it Hoax' have proposed two classification techniques - one based on logistic regression and the other on a novel adaptation of Boolean crowdsourcing algorithms. The idea is to target the fake Facebook users for fake news identification. Focusing on social media, mainly Facebook likes made by authentic and non-authentic users, this research has provided with a new perspective that needs to be considered for identification of hoax spreading around us. Though this research has yet managed to identify hoax or not, but it has paved way for another such researches for other types of fake news as well.

TABLE I. COMPARISON AMONG THREE MAIN TYPES OF FAKE NEWS IDENTIFICATION

There are far more researches going on in this field to emphasize the role of fake news detection software or method in the real world. The base line remains same in all those including the challenges one may face during implementation of any such fake news detection mechanisms.

IV. STANDARD TOOLS AND DATA SETS

For analysing the functioning of any tool for achieving any task, some dataset is needed to verify whether the tool

Criteria	Deep Learning	Machine Learning	Others
Concept	Recurrent Neural Networks [20], Artificial Intelligence	SVMs, AdaBoost, Bagging	N-grams, character, word similarities/ analytics/ crowdsourcing algorithms
Identification	Context - Based	Context - Based	Content - Based
Type of Fake News Targeted	Stance related , rumour	Hoax, Satire, Fake News Postings	Opinion articles, hoax

is working fine or not. Similarly, to get some dataset is a must that should contain real as well as fake news to check the accuracy of its findings. The data set should have these characteristics in them:

1. There should be real news in the data set, so the tool doesn't reject the news
2. There should be fake news to check whether the basic target is being hit by the tool.
3. The data set should contain subjectively a variety of news for the whole spectrum of fake to real news.
4. A criterion defining the percentage of fake or real news in a piece of news.
5. It would be an advantage only if the dataset may contain the real version of the fake news provided.

These are the very few yet required qualities in a data set that qualifies to test a tool meant to identify fake news. To analyse few data sets for fake news, some examples are given below:

1. *Kaggle* - Though having neat structured information with enough news pieces for fake news detection, it lacks the spectrum of news and doesn't include the real news pieces along with
   the partial fake news. Kaggle's fake news dataset is based on BSDetector tool which uses a list of "fake news" sources
2. *LIAR* - LIAR contains short political statements, obtained through the website PolitiFact.com. Each statement is annotated with the author, the context, a veracity label and a justification for such label . includes the required spectrum of realness for news having six fine-grained labels for the truthfulness ratings: pants-fire, false, barely true, half-true, mostly-true, and true . This data set can be used for automatic fake news detection. And this corpus can be used for stance classification, argument mining, topic modelling, rumour detection, and political NLP research along with text classification models like LR and SVMs .
3. *FEVER* - Thorne et al. has included his own produced statements by altering Wikipedia sentences, and then providing evidence for or against such claim in Wikipedia articles
4. Emergent - Ferreira and Vlachos have provided a dataset containing both rumoured claims and related news articles, explaining their accuracy. The objective for such dataset is stance classification .
5. *CREDBANK* - CREDBANK is a large-scale dataset, containing 60 million tweets This dataset is twitter directed for identification of the fake tweets. Tweets are grouped into events by means of topic modelling techniques. Each event is annotated for credibility via Mechanical Turk
6. *George McIntire:* His data set includes lots of news snippets to test the tools with having both sides of spectrum for realness of news. Yet it should not be dangerous to mention it might not be much efficient considering the fact that the news categories might not be 100% right and that there is no in between news for the same .
7. *FakeNewsNet:* It is a data repository with News Content, Social Context and Spatial temporal Information . It may contain the different types of news for the reference of researchers all around the world being enticed by the fake news detection issues.
8. *UNBIASED* - "UNBiased" (UNB) dataset, which integrates various news sources and fulfils several standards and rules to avoid biased results in classification task
9. FNC-1 Challenge - The specific dataset of Emergent which is used for FNC-I challenge is a digital journalism project for rumour debunking. Dataset includes body of the news article, the headline of the news article, and the label for relatedness (stance) of an article and headline

There are many people and researchers coming up with many types of tools to check whether the news is fake or real. These being mostly the websites with some internal coding that works to decide whether a given link leads to a fake or real news. Some of those are:

1. Classify.news
2. Factmata.com
3. http://www.fakenewsai.com/

V. Applications, Uses, Future Aspects, Challenges

Having more than one application for fake news detection methods or technology claims to be a greedy and much resourceful prospect one can look forward to. Some of those applications and uses are:

1. The straightforward one to check if a news article is fake or real or partially fake .
2. One more future aspect that can be a result of same checks can be a real time application that can differentiate the fake news and real news from a set of a number of news articles.

3. Another one being that there can be a web crawler sort of bot for the checking and tagging of news articles for being fake.
4. Till now the main focus is just identifying whether the news is fake or not, if one could identify the type of fake news it would be helpful for users.
5. One big issue is that out of many methods used for fake news identification, most methods focus on one or other type of fake news. There should be a method that can cater to all types of fake news and identify them.
6. Usually, to pose the articles as real, few fake news contains fabricated images and videos to somewhat cheat the readers of those articles of its authenticity. This mixture of text along with video and image checking is necessary for confirming the credibility.

This leads to the many challenges of all such future aspects. Those being:

1. One ideal challenge that everyone faces is the chances or possibility of a news snippet to be fake today and turn out to be real in near or far future. That is the challenge that one may have a tough time to solve. This possibility itself makes the future aspects of fake news a rocky path.
2. Also, how can one identify the fake news from real news when the news is relatively new and real along with the situation that the incident referred to by it has just taken place.
3. All the methods are highly dependent on credibility of few journalism sources.
4. The data sets need to be working for the techniques used with all variety of news available.
5. The method or application should be able to provide with the resulting tagging within real time.
6. There needs to be few authentic sources with pure real news so as to fact check and compare with their news articles for authenticity.
7. The biggest challenge being the different ways one can fake a news article making it very critical to identify and analyse each and every area signified by the article.

There are many more such challenges that result in problems in research or development of fake news detection method or system. This leads the researchers and developers in continuous situations that makes them realise that the research is still far from finished in this area.

VI. Conclusion

With the current scenario ongoing one can easily tell that in today's world nobody is anything less than a journalist who can just post what they intend to. There are no criteria on which their authenticity can be based. Also, there are some mischievous elements in this society that have successfully developed bots that send continuous spam and fake news to social media for their own benefit or just some fun. Social media has become a hub for fake news that needs to be corrected before this fake news render these social media useless. For such there are many researches going on all over the world emphasizing on various points in a news or information snippet to identify and inform of any fake news. We have hence summed up the different types of existing methods that can be used for fake news identification along with few of researches each. Then we also provided with few of what are the available datasets for fake news detection, some focusing on rumours, some on stance etc. We also summarized the future aspects and challenges one faces during fake news detection research. Therefore, it would be apt to say the work is far than finished anywhere considering many shortcomings. Few of them are:

1. Fake news detection should have some criteria to check the satirical or sarcastically meaning of a statement as well along with similarity with text or image.
2. There should be some logic checking function that is inbuilt in human brain yet very difficult for machine.
3. One big issue that we might face for some more time is that all the methods or software available in market like realorsatire.com websites have a functionality that they can only check for one news to be real or fake. No such functionality is available to randomly check news to be fake or say that the work is far from completed in the area of real on social media. There is no such web crawler or bots' type of service available to normal humans.

To sum up these points, as vast is the domain of Fake News, such vast is the Fake News detection research area. Considering the completed work till now, we can say that the work is far from completed in the area of Fake News Detection.