RUMOUR DETECTION ON SOCIAL MEDIA USING MACHINE LEARNING

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
SOFTWARE ENGINEERING

Submitted by:

Saurabh Raj Sangwan
2K16/SWE/13

Under the Supervision of

Dr. AKSHI KUMAR

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

JUNE, 2018

Chapter 1 Introduction and Outline

This chapter briefly introduces the research work proposed in the thesis. Section 1.1 gives an overview of the research undertaken. Section 1.2 sets out the research objectives. Section 1.3 illustrates the proposed framework and the main contributions arising from the work undertaken. Finally, Section 1.4 presents an outline of this thesis describing the organization of the remaining chapters.

1.1. Introduction

With the inception of Web 2.0 [1] and the increasing ease of access methods and devices, more and more people are getting online, making Web indispensable for everyone. The globally accepted new technology paradigm, SMAC (Social media, Mobile, Analytics and Cloud) generates an infinite ocean of data spreading faster and larger than earlier [2]. Active participation is a key element that builds the social web media. Numerous social networking sites like Twitter, YouTube and Facebook have become popular among the masses. It allows people to build connection networks with other people & share various kinds of information in a simple and timely manner. Today, anyone, anywhere with the Internet connection can post information on the Web. But like every coin has its two sides, this technological innovation of social media also has some good as well as bad aspects. We are really benefited by social media but we cannot oversee its negative effects in society. Most people admire it as a revolutionary invention and some seem to take it as a negative impact on the society. As a positive case, these online communities facilitate communication with people around the globe regardless your physical location. The perks include building connection in society, eliminating communication barriers and helping as effective tools for promotion whereas on the flip side privacy is no more private when sharing on social media.

Due to the ubiquitous and over dependence of users on social media for information, the recent trend is to look and gather information from online social media rather than traditional sources. But there are no means to verify the authenticity of information available & spreading on these social media platforms thus making them rumour breeding sources. A rumour is defined as any piece of information put out in public without sufficient knowledge and/or evidence to support it thus putting a question on its authenticity. It may be true, false or unverified and is generated intentionally (attention seeking, self-ambitions, finger-pointing someone, prank, to spread fear & hatred) or unintentionally (error). Further, these can be personal as well as professional. Knapp [3] classified Rumours into three categories, namely, pipe dream, bogy and wedge driving for describing intentional rumours.

Rumours are circulated and believed overtly. And due to the increasing reliance of people on social media, it is inevitable to detect and stop rumours from spreading to reduce their impact. It takes only little time for a single tweet or post to go viral and affect millions.

Fig. 1.1: Cascading effect on Rumour

Rumour detection and mitigation has evolved as a recent research practice where the rumour has to be recognized and its source has to be identified to limit its diffusion. It is essential not just to detect and deter, but to track down the rumour to its source of origin. Various primary studies with promising results and secondary studies [4, 5] have been reported in this direction. A typical rumour analysis task consists of four components:

(1) *Rumour Detection:* where potential rumours are recognized

(2) *Rumour Tracking:* monitors the tweet, filters and captures related posts

(3) *Stance Classification:* determines the orientation of user’s view as “in favour”/ “against” and

(4) *Veracity Classification:* knowledge is garnered based on the selection of significant features and subsequent classification is done to determine the actual truth value of the rumour.

In this work we propose a rumour analytics model for the first and the fourth components that is, the recognition of potential rumours and veracity classification. The remainder of this chapter sets out the research objectives, describes the main contributions of the research work, and presents an outline of this thesis.

1.2. Research Objectives

Statement of Research Question

“Can we detect rumour and automate a predictive model which classifies the questionable veracity of rumour?”

Pertinent psychological studies convey that humans are intrinsically not very good at differentiating conflicting information and to gauge the veracity, classification relies immensely on the extraction of stance or sentiment from relevant posts. Thus, this unifying research question can be broken down into the following six questions, each of which will be addressed by this research:

- How can rumour be detected on social media?
- How can virality be related to rumour detection on social media?
- How can we predict the veracity (true, false, unverified) of a rumour?
- What features are to be investigated for capturing truthfulness of a post with questionable veracity?
- Which supervised machine learning technique is the best for the veracity prediction task?

Consequently, the three main research objectives of the work undertaken are:

i. **Research Objective I** – To seek the correlation of virality on social media and rumour detection
ii. **Research Objective II** – To propose a feature-based predictive model for veracity classification
iii. **Research Objective III** – To find out the best classifier for prediction on benchmark datasets

The objective of this thesis is to find the list of potentially rumourous tweets and then use a predictive technique to automatically determine its actual truth value with accuracy & without delay.

1.3. Proposed Model

The proposed rumour detection model, VRV Model (**V**irality-**R**umour-**V**eracity Model), consists of two modules, that is, firstly the virality prediction module and secondly the veracity classification module. The virality prediction module determines the likelihood of tweet going viral based on the strength of emotion in tweets and its no. of retweets. Once a tweet is identified as viral, tools and techniques that authenticate its source and veracity can be employed to mitigate any intentional and wrongful circulation. That is, a list of tweets with high virality score forms the list of potential rumours and this way the virality prediction module can be considered as a preliminary step to detect a rumour for which the actual truth value needs to be determined. Next, the veracity classification module performs the task of verifying the accuracy of a rumourous post. This veracity classification task aims to determine whether a given rumour can be confirmed as true, debunked as false, or its truth value is still to be resolved. Three different varieties of features (content-based, pragmatic & network-specific) are used to automate the identification of rumour dexterously using six supervised learning techniques namely, Support Vector Machine, Decision Trees, Logistic Regression, Random Forest, K-Nearest Neighbors and Neural Networks on random tweets pertaining to social and political issues. The veracity classification module has also been evaluated on the benchmark Twitter SemEval-2017 Task 8: RumourEval dataset [6].

1.4. Organization of Thesis

This thesis is structured into 5 chapters followed by references.

Chapter 1 presents the research problem, research objectives, justifies the need for and outlines the main contributions arising from the work undertaken.

Chapter 2 provides the essential background and context for this thesis and provides a complete justification for the research work described in this thesis.

Chapter 3 provides the details of the methodology employed and outlines the Rumour Detection Model (**V**irality-**R**umour-**V**eracity Model → VRV Model) that constitutes the proposed approach of the research.

Chapter 4 describes the experimental results obtained from a tweet illustration. It also presents the analysis to account for the tests performed.

Chapter 5 presents future research avenues and conclusions based on the contributions made by this thesis.

Literature Review

In this chapter discusses the background work in the research domains of rumour detection and virality prediction on social media. We present a state-of-art review of rumor detection on online social media. The research gaps have been identified as issues and challenges within the domain which make it an active and dynamic area of research.

2.1 Rumor Detection on Social Media

Social media has the power to make any information, be it true or false, go viral and reach and affect millions. Social networks have been witness to the self-reinforcing Echo Chambers which steers a confirmation bias (false sense of affirmation that we are right in our beliefs) and relevance paradox (readers only consume information that is relevant to them, kind of one-sided). Thus due to the speed of information spread, rumours are cascaded. Recently, the journal '*Science'* published a study which analyzed millions of tweets sent between 2006 and 2017 and came to this chilling conclusion that "Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information" [7]. In fact, the largest and the most popular user-base, Twitter, itself, has been a constant subject of mostly groundless acquisition rumors. In January 2018, a similar flurry of unconfirmed takeover rumors appeared too. Hence, it is necessary to detect and restraint these rumours before they have a serious impact on people's lives.

Fig. 2.1. Twitter Buyout Rumour

2.1.1 Types of Rumour

Formally, a rumor is defined as information whose veracity is doubtful. Some rumors may turn out to be true, some false and others may remain unverified. Not all false information can be classified as a rumor. Some are honest mistakes by people and are referred to as misinformation. On the other hand, there may be intentional rumors put to mislead people into believing them. These are labeled as disinformation and are further classified based on the intent of the originator. The following figure 1 depicts the classification of rumors.

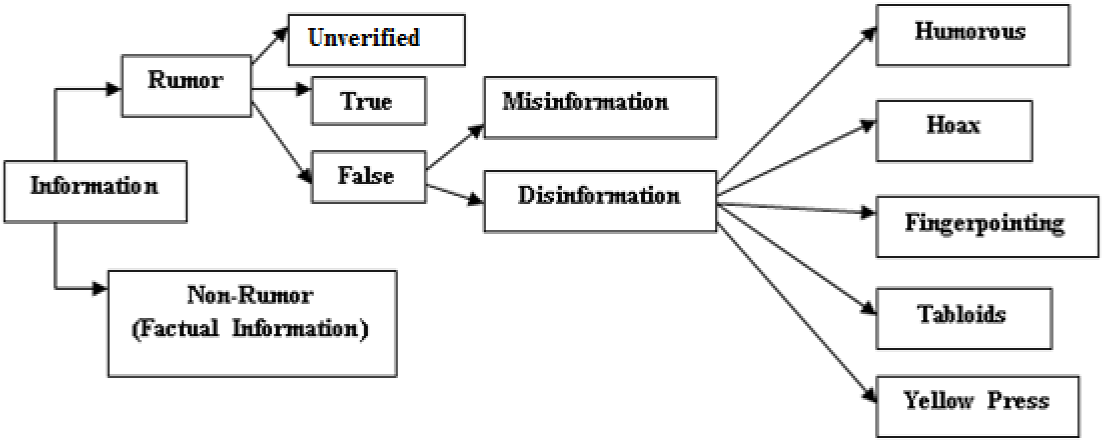


Fig. 2.2. Classification of Rumors

A rumour as any information put out in public without sufficient knowledge and/or evidence to support it. It is misleading, either intentionally or unintentionally. If some information has been put out in public erroneously without authentic or complete information with no ulterior motive of hurting or causing any disturbance to anyone whatsoever, it is called misinformation. It is an honest mistake. Disinformation, on the other hand is information that is intentionally put out in public view to mislead people and start a false rumour. Disinformation depending on the motive of the writer and nature of the post can be classified as humorous, hoax, finger pointing, tabloids, and yellow press. The most harmless type of rumour is the humorous ones. Sources spreading this type of information fabricate news and stories to give it an amusing side. The motive is usually to entertain people. The information is pre-declared to be false and intended only for comical purposes. The best examples of such sources include news satires, and news game shows.

The next form of disinformation is a hoax. A hoax is intentional fake news spread to cause panic among people and cause trouble to people at whom it is aimed. A hoax can also be an imposter. Examples include fabricated stories, false threats etc. In 2013, a hoax stating Hollywood actor ‘Tom Cruise to be dead’ started doing the rounds. Social messaging apps like WhatsApp worsen the situations when it comes to hoaxes. Currency ban of Indian rupees 500 and 1000 was done in November 2016. Soon after a hoax message went viral on WhatsApp stating that the government will release a new 2000 rupee denomination that would contain a GPS trackable nano-chip that would enable to locate the notes even 390 feet buried underground. The government and bank spokespersons had to finally issue an official statement stating it was false. Still, many people found the official statement hard to believe as they were so brainwashed by the hoax message.

Another form of disinformation is finger pointing. Finger pointing always has an associated malicious intent and personal vested interest. It blames a person or an organization for some bad event that is happening or happened in the past. It aims at political or financial gain by tarnishing the image of the target person/organization/party/group etc. Tabloids have a bad name for spreading rumours from since when they started. It is type of journalism that accentuates sensational stories and gossips about celebrities that would amount to spicy page 3 stories. Yellow press journalism is a degraded form of journalism which reports news with little or no research at all. Journalists’ only aim is to catch attention using catchy headlines with no regards whatsoever to the authenticity of news. They do not bother to delve deep into a story but just publish it to sell as many stories as possible and make money. It is the most unprofessional and unethical form of journalism.

2.1.2 Typical Architecture of Rumour Analytics

There are four basic components for a complete rumour analytics process. These include, Rumour Detection; Rumour Tracking and Monitoring; Rumour Stance Classification and Rumour Veracity Classification. The analytical process typically begins with a identifying a piece of information (social media post) which constitutes a rumour and ends by determining the truth value (veracity) of the post. The following figure 2.3 illustrates these components:

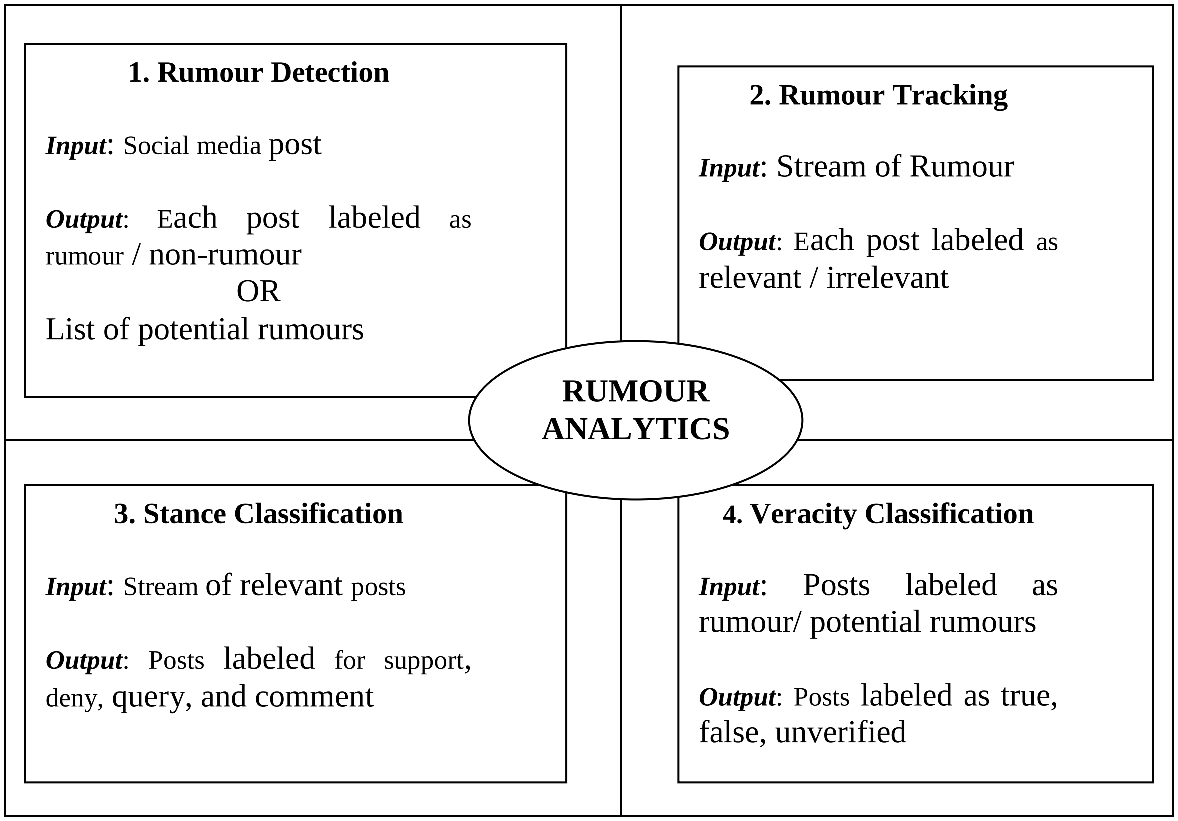

1. Rumour Detection
Input: Social media post
Output: Each post labeled as rumour / non-rumour
OR
List of potential rumours
2. Rumour Tracking
Input: Stream of Rumour
Output: Each post labeled as relevant / irrelevant
RUMOUR ANALYTICS
3. Stance Classification
Input: Stream of relevant posts
Output: Posts labeled for support, deny, query, and comment
4. Veracity Classification
Input: Posts labeled as rumour/ potential rumours
Output: Posts labeled as true, false, unverified

Fig. 2.3. Architecture of Rumour Analytics

2.1.3. Related Work in Rumour Veracity Classification

Machine learning based techniques have emerged as promising viable approach for detecting rumours on social media. The majority of research which deals with rumours on social media is centered towards veracity classification, which is fundamental to determine the truthfulness of rumour. The work carried in this research is post the initial step of rumour detection, which has already been carried either manually or automatically. Hence, we the discussion here specifically focuses on the state-of-art of veracity classification system and skips other components. The following table 1.1 presents a brief literature review of veracity classification:

Table 1.1: State-of-Art of Veracity Classification

Year of publication	Author	Data set	Technique applied	Feature set	Conclusion
2011	Castillo et al. [8]	Twitter	DT, NB, SVM.	Message-based, user-based, topic based, and propagation-based.	DT was superior
2013	Kwon et al. [9]	Twitter	DT, RF, SVM, LR	Temporal, structural, and linguistic	RF classifier performed best.
2017	Kwon et al. [10]	Twitter	RF	User-level, structural, temporal , linguistic	Cumulative spreading pattern of rumors prediction
2012	Yang et al. [11]	SinaWeibo	SVM with the RBF kernel	User based, linguistic based, client-based and location-based features	An increase in accuracy was obtained i.e. from 72.3% to 78.6%
2015	Yang et al. [12]	SinaWeibo	SVM with the RBF kernel	Network based features (created a social network based on the reviews or comments attached to the source tweet)	Enhanced results were obtained using network based features along with traditional features.
2015	Liu et al. [13]	Author's own dataset	SVM with the RBF kernel DT, NB	Verification features like source credibility, source identification, source diversity, source and witness location, event propagation, and belief identification	SVM gave the best results
2015	Ma et al. [14]	Twitter and SinaWeibo	DT, RF, SVM with the RBF kernel	Modeling features over time	The proposed approach (SVM) had produced the best accuracy of around 84.6%
2015	Wu et al. [15]	Used rumours with at least 100	SVM with a hybrid kernel technique consisting of	Message-based, user-based, and report-based	The proposed hybrid approach had shown improved

		reposts	random walk kernel, SVM with a hybrid kernel technique consisting of an RBF kernel.		accuracy
2015	Wang and Terano [16]	Twitter	Social graphs with linear model	Features measured by the number of contacts such as RTs, replies, and comments between two users, activeness measured by the number of days a user has sent out messages, similarity measured by gender and location, similarity between two users, and trustworthiness measured by whether the user is verified or not	Using a new proposed metric, influential spreaders were identified and were used to determine rumours.
2015	Vosoughi [17]	Twitter	Dynamic time wrapping (DTW) and hidden Markov models (HMMs)	Linguistic, user oriented, and temporal propagation related	The results showed that HMMs were superior to DTWs.
2016	Giasemidis et al. [18]	Twitter	SVM	Message-based, user-based, and report-based	The authors reported very good results using decision trees.
2016	Chang et al. [19]	Twitter	Clustering approach	Characteristics of users	Author had applied a simple clustering heuristic and based on it had categorized the

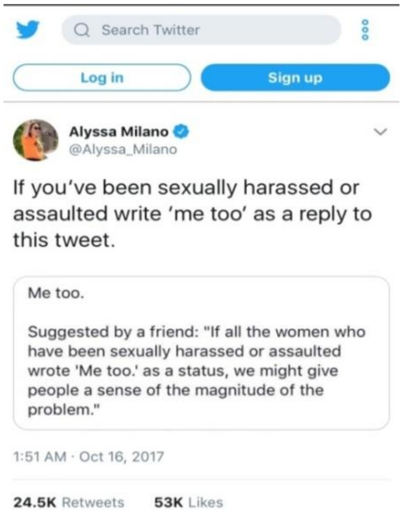
					rumour sets as false or true rumour clusters. The results show improved performances.
2016	Chua and Banerjee [20]	Twitter	Binomial LR	Comprehensibility, sentiment, time orientation, quantitative details, writing style, and topic, negation words; (comprehensibility category), past, present, future POS in the tweets (time-orientation category); discrepancy, sweat and exclusion features (writing style category); and, finally, home, leisure, religion, and sex topic features (topic category).	Author had obtained improved results.
2017	Ma et al. [21]	Twitter	Linear SVM, SVM-Time Series, DT using Ranking method, RF and RNN	Bag-of-words (BoW) and word-embedding	BoW representation was superior to the embedding variant.
2015	Zhang et al. [22]	Liuyanbaike.com (a Chinese rumour-debunking platform	LR	Mention of numbers, the source the rumour originated from, and hyperlinks	Author obtained better results.
2017	Enayet and El-Beltagy [23]	SemEval Rumour Eval	Linear SVC	User and Content based	Linear SVC proved to be the best in terms of accuracy

NB: Naïve Bayes; SVM: Support Vector Machine; LR: Logistic Regression; DT: Decision Tree; RF: Random Forest; RBF Kernel: Radial Basis Function Kernel SVC: support vector classifier Kernel

The detection of new rumors from real-time data is a challenging task. It is easier to detect old posts regarding a rumor that we know because of the keywords. But with emerging rumors we are in a fix as we do not know what to look out for. Also some rumors remain un-specified and there is no conformation or debunking for them. Hence, detecting rumors and resolving their veracity is very tricky. Concurrently, social media virality inherently carries the potential to reach out to a vast majority as it simultaneously affects the social life both positively and negatively. Viral rumours are major carriers of panic. Thus, virality detection can be an initial step to identify and highlight information with questionable veracity. The following sub-section background work in this direction of virality prediction and specifically the use of emotions in viral posts on Twitter.

2.2. Virality on Social Media

The widespread activation of information propagation across meta-networks is referred to as the "virality". The magnitude of social media virality cannot be overrated. It can bring fame and prosperity but at the same time can beget notoriety and nuisance. Twitter is one of the most popular social networks worldwide and as per the statistics for the first quarter of 2018, this micro-blogging service averaged at 336 million monthly active users globally [24]. The platform is used as a communication channel by businesses, celebrities and even government. Encouraging vigorous participation in such channels can be intentional or unintentional with the activities ranging from supporting a cause, getting involved, expressing personal feelings or beliefs, attention seeking, self-ambitions, finger-pointing someone, viral marketing, prank or to spread fear & hatred. Information virality refers to the inevitable cascading effect of information spread online which eventually proliferates across meta-networks and affects millions. In October 2017, the *#metoo* movement created a wave of global reckoning for being posted by women who say they've faced sexual harassment and assault [25]. The impact of these two words was so much that it soared across social media including, Facebook and Instagram. It was one seismic activity which demonstrated the fortitude of social platforms and its virality.



Alyssa Milano
@Alyssa_Milano

If you've been sexually harassed or assaulted write 'me too' as a reply to this tweet.

Me too.

Suggested by a friend: "If all the women who have been sexually harassed or assaulted wrote 'Me too.' as a status, we might give people a sense of the magnitude of the problem."

1:51 AM · Oct 16, 2017

24.5K Retweets 53K Likes

Fig. 2.4. Social Media Virality and its effect

Thus, it becomes exceedingly imperative to resolve the authenticity of information and promptly inhibit it from spreading among the Internet users as this can jeopardize the well-being of the citizens. Pertinent psychological studies convey that humans are intrinsically not very good at differentiating conflicting information. Naive Realism and Confirmation Bias further add to the vulnerability. Though the cascading model of tweet-re-tweet captures the virality of a tweet over its lifetime, the likelihood of content going viral has more to do with how activated the person felt after reading it. Crucially, it's just not the volume of tweets that matter, but the "homogeneity" and "irregularities" in the emotion that can make the difference.

The term 'Virality' is originally from the biological sciences where the viruses contagiously spread among organisms. But recently, the term has found a new technological meaning with its social media presence. It is more than the basic person-to-person broadcasting and relies on word-of-mouth. "Going viral" and "Viral marketing" are two buzz terms reigning the online marketing and economics. Primary and secondary studies have been reporting the virality of content (tweets, posts, videos, photos) on social media.

Weng et al. [26] proposed a prediction model for information virality detection on Twitter using data about community structure. They show that, while most memes indeed spread like complex contagions, a few viral memes spread across many communities, like diseases. Using the proposed model the authors also demonstrate the future popularity of a meme by quantifying its early spreading pattern in terms of community concentration. Hoang et al. [27] present a virality model of twitter content to find viral tweets, viral users and viral topics. The highly viral messages, topics and users in GE2011 are extracted and evaluated using the model.

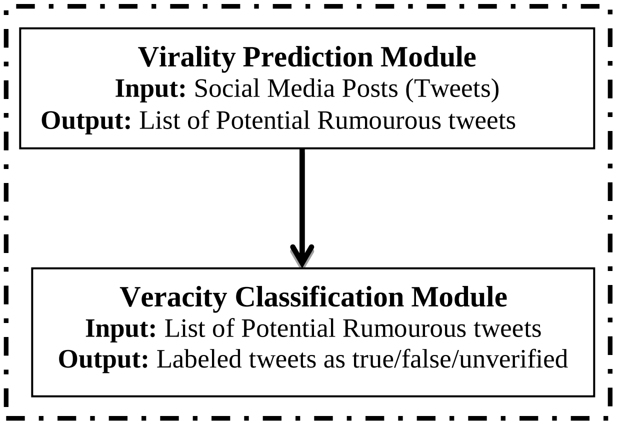
Berger and Milkman [28] were the pioneers to add psychological approach to online content virality. The authors suggest the relationship between emotion and transmission to understand what becomes viral. Hansen et al. [29], study the relation between affect and virality to understand the psychological and sentimental arousal theories. The dataset includes three corpora: tweets about the COP15 climate summit, random tweets, and text corpus including news. The findings also present evidence that negative sentiment enhances virality in the news.

Chapter 3 Proposed Model

Chapter 2 identified issues related to rumour analytics on social media. This chapter illustrates the novel techniques that constitute the proposed model to address those issues presented in Chapter 2. Section 3.1 gives an overview of the research undertaken. Section 3.2 depicts the architectural view of the proposed model. Finally, Section 3.3 illustrates the proposed model, describes each component of the system and shows how each of the proposed technique contributes to the rumour detection process.

3.1. The Proposed VRV Model of Rumour Detection

The intent of the work proposed in this research is to firstly find an approach that will enable predicting a viral tweet by virtue of its public emotion strength. This viral tweet will be a qualified candidate of potential rumour for which the veracity classification will then be done. Thus, as a typical text mining task, the **V**irality-**R**umour-**V**eracity Model (VRV Model) of Rumour Detection consists of two modules, namely, the virality prediction module and the veracity classification module. The following figure 3.1 depicts the proposed model.



Virality Prediction Module
Input: Social Media Posts (Tweets)
Output: List of Potential Rumourous tweets
Veracity Classification Module
Input: List of Potential Rumourous tweets
Output: Labeled tweets as true/false/unverified

Fig. 3.1. The **V**irality-**R**umour-**V**eracity Model (VRV model) of Rumour Detection

The following sections expound the details of the model:

3.2. Virality Prediction Module

The hypothesis laid is that “As unverified information spreads considerably on social media, it works with the same mechanics as that of a large protest where an outsized share of same emotion is representative of the response sensitivity. That is, emotions may be ‘activated’ or ‘deactivated’ to drive people to take action and a dominant emotion of same type across tweets is indicative of a viral spread. Fluctuations in emotions convey uncertainty and may reduce the frequency and intensity of discussion of a trending topic.” Based on this, we propose the use of cognitive behavioural features to assess the virality of information in tweets. The proposed technique detects the emotion quotient (EQ), a measure of emotional intensity associated with five emotions, namely, fear, disgust, sadness, anger, and happiness for the exposed information in tweets to predict its outburst, i.e., virality, pertaining to social and political issues.

The approach is to transform the tweet into an emotional vector representative of the sentimental value for a trending topic. A lexicon based technique is employed to associate the emotional values for the words in the sentence. Parts of speech like adjectives, adverbs and some groups of verbs and nouns have been reported as good indicators of fine-grain sentiment across pertinent literature [30, 31]. In this research, the adjectives, the verbs and the adverbs are considered as the emotion carriers in the sentence for feature-level emotion analysis. In natural language, the adjectives help to express the fundamental feelings and emotions within a tweet. The verbs operate as polarity markers as they convey the tone associated with the emotion. Similarly, the adverbs act as emotion bolsters, which scale the emotion polarity in terms of strength. For example, the occurrence of adverb “not” in “not bad” inverts the emotion value of the next word whereas the occurrence of adverb “ruthlessly” amplifies the emotion value of the next word. The use of emoticons has become a mainstream culture in social content writing and their use cannot be ignored as they suggest adjectives which add tone and clarity to the communication. Basically, the emoticons influence emotional communication. Studies suggest that emoticons, when used in conjunction with a written message, can help to increase the “intensity” of its intended meaning. Thus, the emotion analysis tool works by assigning emotion value to each adjective and emoticon in the sentence and obtaining the polarity value of verbs and the strength of adverbs.

In order to set the benchmark for empirical analysis with the created adjective emotion lexicon base, we apply classifiers. We analyze six supervised learning algorithms namely, Support Vector Machine (SVM), Decision Trees (DT), Logistic Regression (LR), Multi-layer Perceptron (MLP), Random Forest (RF), K-Nearest Neighbors (KNN) for predicting the adjective emotion values for each tweet. The emotion quotient for each tweet is then calculated using a linear equation with scores from all four lexicon base. This patterning of emotions with time along with the number of times a tweet is re-tweeted measures the viral value of a tweet. Thus, the proposed module will enable predicting a viral tweet by virtue of its public emotion strength. The following figure 3.2 depicts the proposed virality prediction module

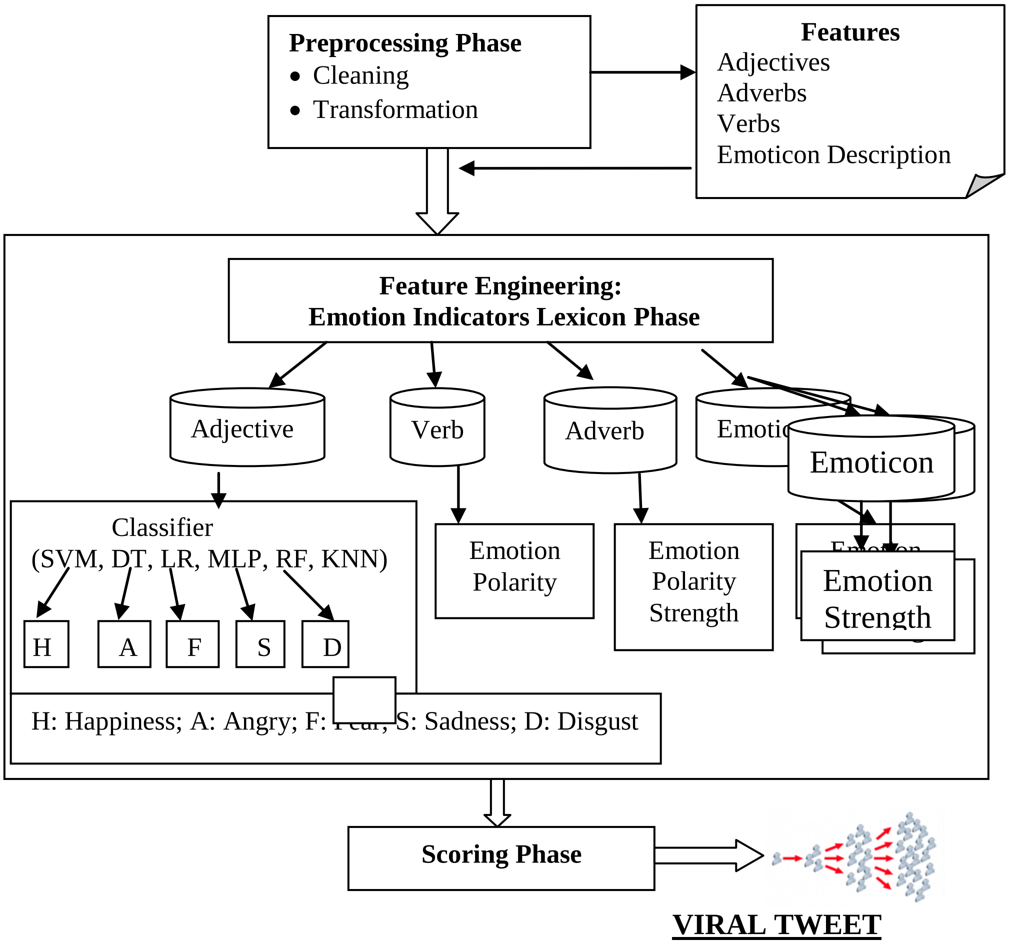


Fig. 3.2: Tweet Virality Prediction Module

As shown, the virality prediction module consists of three phases, namely, the pre-processing phase, the feature engineering: emotion indicator lexicon phase and the virality scoring phase. The following sub-sections expound the details of the module:

3.2.1. Preprocessing Phase

The tweets pertaining to a topic (#topic) are extracted from the publically available Twitter datasets using its API. In order to intelligently mine the text in tweets, preprocessing is done for cleaning and transforming the data for relevant feature extraction.

- Primarily the preprocessing includes cleaning the text by removal of all URLs, hash tags, @username and non-English words followed by the transformation of text for relevant feature extraction. Sometimes people may use hash-tags to convey direct and explicit emotions, for example #sad but we have omitted these as our main aim to predict the strength of emotion and not just the emotion.
- Text transformation firstly replaces the emoticons in text with their descriptive text or phrase. As the name suggests, emoticons are emotion icons and convey the emotions similar to human facial expressions. Their use has become a mainstream culture and so these cannot be omitted as they suggest adjectives which add tone and clarity to the communication. Emoticons influence emotional communication. Researchers found that emoticons, when used in conjunction with a written message, can help to increase the "intensity" of its intended meaning [32]. For example, the emoticon ☹ will be replaced by its description "*sad face*" and will be assigned an emotion strength value of -0.5. Thus we replace all the emoticons with their description and polarity using the values presented in the table 1 below. The list is an updated version of the list used by Kumar & Sebastian [31] to decipher and use emoticons.

Table 3.1: Emoticons

Emoticon	**Description**	**Emotion Strength**
:-D	Big Grin	1
XD	Laughing	1
<3	Heart	1
:), =), :-)	Happy, Smile	0.5
:*	Kiss	0.5
0:)	Angelic	0.5
:|, :-|	Straight Face, Indifferent	0

:\	Undecided	0
:(, =(	Sad	-0.5
</3	Broken Heart	-0.5
=O, :-o	Shocked	-0.5
:’(	Cry	-1
X-(	Angry, Frown	-1
xP	Disgusted	-1

It is important to make note that although the use of emoticons like Winking ;) and Sticking tongue out :P is widespread but it opens up a new avenue of research, as the use of these emoticons is related to a sarcastic, humourous, non-serious, joking tone of the post which may completely reverse the emotion conveyed by the textual indicators. For example, a tweet “We will all be killed then…Lets meet in heaven ;)” is a humourous tone whereas the textual emotion analytics will detect this as a negative one. For the module defined in this research, we have omitted the use of any such emoticons and have only considered the ones defined in table 3.1.

Next, using a POS tagger, only the adjectives, verbs and the adverbs are extracted to build the feature set. The emotion scores are then assigned to these to compute the final emotion quotient for the tweet.

3.2.2. Feature Engineering: Emotion Indicators Lexicon Phase

The adjectives, verbs and adverbs are expressions of sentiments which convey emotions strongly. Adjective is that part-of-speech which describes, qualifies and identifies a noun or pronoun. Verbs express activity in terms of an action, an occurrence, or a state of being. Adverbs are words that change the meaning of a verb, adjective. In unison, these three parts-of-speech and emoticons quantify the emotion strength and will assist in capturing the growing emotional response of online users associated with a topic (an event, a person, a place, an issue). The lexicons for all these three emotion indicators are created and assigned values through a crowdsourcing initiative. Also, supervised learning models have been empiraclly analyzed for prediction of adjective emotion category. The details of each lexicon is explained.

A corpus of most commonly used adjectives created and validated in our earlier research [32] has been used for creating and assigning values to emotion tuples. The sample emotion tuple value for few adjectives is represented in table 3.2. The emotion values are assigned on a scale of 0 to 5 for five emotions in the vector, namely, fear, disgust, sadness, anger, happiness.

Table 3.2: Adjective Emotion Values

Adjective	**Happiness**	**Anger**	**Sad**	**Fear**	**Disgust**
damaging	1.33	3.5	3.06	2.73	2.42
dirty	1.28	2.3	1.94	1.94	3.7
easy	3.92	1.11	1.15	1.19	1.09
easygoing	3.98	1.14	1.14	1.14	1.11
ecstatic	4.08	1.34	1.31	1.8	1.52
elated	3.93	1.21	1.19	1.17	1.12
famous	3.32	1.3	1.21	1.2	1.38
fantastic	4.07	1.19	1.31	1.25	1.22
greedy	1.41	3.14	2.68	2.27	2.94
hard	1.65	2.22	1.75	2.21	1.4
innocent	3.17	1.37	1.49	1.66	1.27
lazy	1.49	2.01	1.83	1.4	2.39
menacing	1.17	2.94	1.78	1.97	2.18
merry	4.38	1.07	1.14	1.08	1.08
noisy	1.39	2.97	1.39	1.41	1.45
nonchalant	1.85	1.4	1.31	1.26	1.47
protected	4.11	1.24	1.33	1.47	1.08
proud	3.18	1.55	1.29	1.58	1.26
quartan	1.39	1.18	1.17	1.17	1.15
rejected	1.05	3.5	3.91	3.47	2
relaxed	4.32	1.12	1.14	1.1	1.04
scared	1.14	2.41	3.02	4.09	1.83
scornful	1.16	3.31	2.13	2.17	1.74
serious	1.45	1.92	1.84	1.97	1.29

Further, we analyze six supervised learning algorithms namely, Support Vector Machine, Decision Trees, Logistic Regression, Multi-layer Perceptron, Random Forest, K-Nearest Neighbors(Fig. 3.3.) for predicting the adjective emotion values for each tweet.

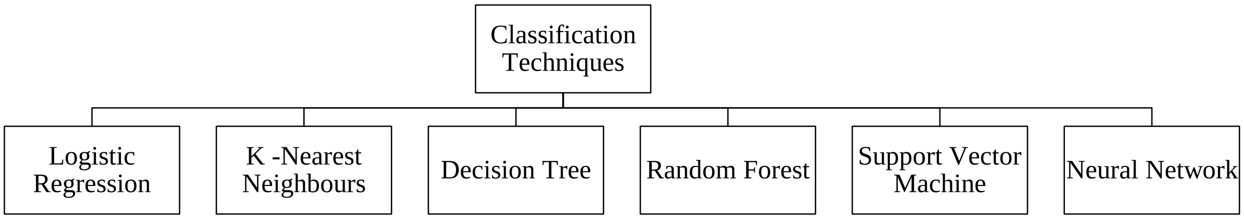


Fig. 3.3: Supervised Classification Techniques

The details about these techniques are given in the table 3.3 below:

Table 3.3: Supervised Learning Techniques

Technique	**Description**
Logistic Regression (LR)	One of the most basic classification techniques, logistic regression utilizes a logistic function, also known as sigmoid function. It associates each input value with a coefficient (ϴ), and trains the given system to adapt to expected output value by modifying these ϴ values
K-Nearest Neighbours (KNN)	K-Nearest Neighbors is a classification algorithm that is based on feature similarity; it focuses on similarities between values in a class. It treats input values as vectors in a feature space, and is based on votes given by its k nearest neighbors. KNN is a lazy learning algorithm; it doesn't generalize through available data, but instead represents the data as it is .
Support Vector Machines (SVM)	Support Vector Machine is a model building mechanism, which helps predict classes through its training algorithms. It represents the dataset as a map in such a way that there's a clear defined gap between various classes it classifies values into. Its approach depends on number of classes it's classifying, and the representation of its mapping. For two classes, it's called a Binary SVM Classifier.
Decision Tree (DT)	Decision trees are powerful tools used for classification. A decision tree symbolizes a set of rules, which help us to determine the class an input belongs to. The decision making process of these trees starts from the root, traverses downwards and ends up at leaves. The leaf nodes of a decision tree represent the values of an attribute. The other nodes are called decision nodes, which test given values and determine factors that help us classify them as we go downwards. Its training involves selecting the appropriate attribute to split the tree at each stage, while keeping the tree compact and organized.
Random Forests (RF)	Random Forest Algorithm overcomes the limitations of Decision tree method, by creating a forest of trees. The higher the number of trees, the greater is the accuracy of the system. It selects random subsets of the training input with replacement and fits decision trees in accordance with those samples, also called Bagging. This technique decreases the variance of the model, by averaging it out across many trees thereby cancelling noise and giving it the ability to generalize again.
Multi-layer Perceptron (MLP)	MLPs is a type of feed-forward artificial neural network which uses back propagation as a supervised learning technique. MLP can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model.

Thus, the adjectives are analyzed and classified for five pre-defined emotion categories namely Happiness, Anger, Sadness, Fear and Disgust. The classification results are evaluated based on precision, recall, accuracy and F-score as the performance measures. We discuss the results in chapter 4.

Out of the five emotion categories considered for this work, happiness is the only emotion which has a positive polarity whereas the other four, namely, anger, sadness, fear and disgust have negative polarity. The natural language words conveying anger, sadness, fear and disgust are often related to anxiety and depression in humans. These are the "trigger" emotions which drive people to take action which makes it more likely to pass things as a chain reaction. Thus, to identify the category of emotions we determine the polarity (positive or negative) of the verbs. An emotion polarity lexicon base for 100 most commonly used verbs is created and the polarity values are assigned within the range of +1 to -1. Further the strength of this polarity is assessed using an adverb emotion polarity strength lexicon base created for this research. The respective emotion polarities & strengths within both the lexicon-base have been congregated through a crowdsourcing task. The polarity strength value and emotion polarity for few adverbs and verbs is shown in table 3.4 and table 3.5 respectively.

Table 3.4: Adverb Emotion Polarity Strength

Adverb	**Emotion Polarity Strength**
Extremely	+1
Terribly	0.9
Seriously	0.8
Totally	0.7
Completely	0.6
Most	0.5
Too	0.4
Very	0.4
Highly	0.4
Pretty	0.3
More	0.2
Much	0.1
Any	-0.1
Quite	-0.2
Just	-0.3
Little	-0.4
Dimly	-0.5
Less	-0.6
Not	-0.8
Never	-0.9
Hardly	-1

Table 3.5: Verb Emotion Polarity

Verb	Emotion Polarity
Love	1
Adore	0.9
Won	0.9
Like	0.8
Enjoy	0.7
Kiss	0.7
Smile	0.6
Impress	0.5
Attract	0.4
Excite	0.3
Relax	0.2
Kill	-1
Shoot	-1
Revenge	-1
Hate	-1
Destruct	-0.9
Harm	-0.9
Hurt	-0.8
Fight	-0.8
Beat	-0.7
Hit	-0.7
Yell	-0.6
Lost	-0.5
End	-0.4
Detest	-0.2
Reject	-0.1

The seed lists of positive and negative adverbs and verbs whose orientation we know is created and then grown using the WordNet [33]. That is, for each Adverb and Verb occurring in a tweet, it is checked for its presence in the seed list. If it is a hit, the values are assigned and returned else in case of a miss, WordNet is used to extract synonym and antonym with known value and assigned the value accordingly.

3.2.3. Scoring Phase

Once the emotion value from all indicators is extracted, the next step is to gauge the emotion quotient of the tweet for subsequently calculating the viral value of a tweet and virality of a topic.

The Emotion Quotient (EQ) of a tweet is calculated using the following equation (3.1)

$$EQ = \frac{1}{a+b+c+d}\left(\frac{\sum_{i=1}^{n}|E_{adj}|_i}{n*5} + \frac{\sum_{i=1}^{m}|E_{vb_i}|}{m} + \frac{\sum_{i=1}^{p}|E_{avb_i}|}{p} + \frac{\sum_{i=1}^{q}|E_{emot_i}|}{q}\right) \qquad (3.1)$$

where,

E_{adj_i},E_{vb_i}, E_{avb_i} , E_{emot_i} are the emotion values of adjective, verb, adverb and emoticon respectively. As these quantify the strength of the emotion, we take the mod of values;

n, m, p and q are the number of adjectives, verbs, adverbs and emoticons present in the tweet;

The parameters a, b, c and d are used to signify the presence of the emotion indicators. For example, if an adjective is absent, the value of will be 0 and if it’s present the value of a will be 1. This has been done to dampen the values of emotion quotient such that they are normalized within the range of 0 to 1. The value of the parameters is assessed as shown in table 3.6 below:

Table 3.6: Parameter Value

Parameter	**Value =0**	**Value=1**
a	n=0	n>0
b	m=0	m>0
c	p=0	p>0
d	q=0	q>0

Next, based on the emotion quotient of a tweet, the viral value of the tweet (VV$_{\text{tweet}}$) is calculated using the following equation (3.2)

$$VV_{tweet} = Polarity\left[\frac{EQ*R}{T}\right] \qquad (3.2)$$

Where, the Polarity is in terms of positive or negative sentiment (indicated by + or -). It determines the emotional factor of the post. Out of the five emotions considered, fear, anger, sadness and disgust are negative emotions whereas happiness is a positive one. But in the absence of an adjective in tweet, this polarity classification is not possible. So, we propose that, as the adverbs qualify adjectives and verbs, the adjective group (adjective*adverb) or the verb group (verb*adverb) polarities will determine the overall polarity of a particular tweet. This is imperative in determining the emotional orientation of the posts as the strength of same emotion type will be a yardstick of virality.

EQ is the emotional quotient of the tweet calculated using equation 1;

R is the no. of retweets, that is, the total no. of times the tweet has been reposted;

T is the time, that is, the life span of the tweet counted in number of days.

A transaction file is maintained for each tweet on the topic storing the emotion quotient, its polarity, no. of retweets, life-span and the viral value of the tweet. The threshold for a tweet being called "viral" has been set to 400. So any value of virality greater than 400 implies that the tweet has a cascading effect and steps to authenticate its accuracy and origin must be taken by agencies (business or government). That is, the information further needs to be checked for veracity and origin to restrict flare-up of rumour. The + and – simply indicate the polarity of the post. As discussed earlier, out of the five emotions considered, fear, anger, sadness and disgust are negative emotions whereas happiness is a positive one.

3.3. Veracity Classification Module

Veracity classification refers to the task of verifying the accuracy of a rumourous story. This veracity classification task aims to determine whether a given rumour can be confirmed as true, debunked as false, or its truth value is still to be resolved. Given a set of posts associated with a rumour and, optionally, additional sources related to the rumour, the task consists of assigning one of the following labels to the rumour, Y ε {true, false, unverified} [34]

The proposed model has primarily four phases,

i. Data Acquisition
ii. Data Pre-processing
iii. Feature Extraction
iv. Model Training

The task begins by collecting data, which in this research are primarily random tweets pertaining to the social and political issues. The list of potential tweets is obtained from the first module, that is, the virality prediction module. The following figure 3.4 illustrates the module:

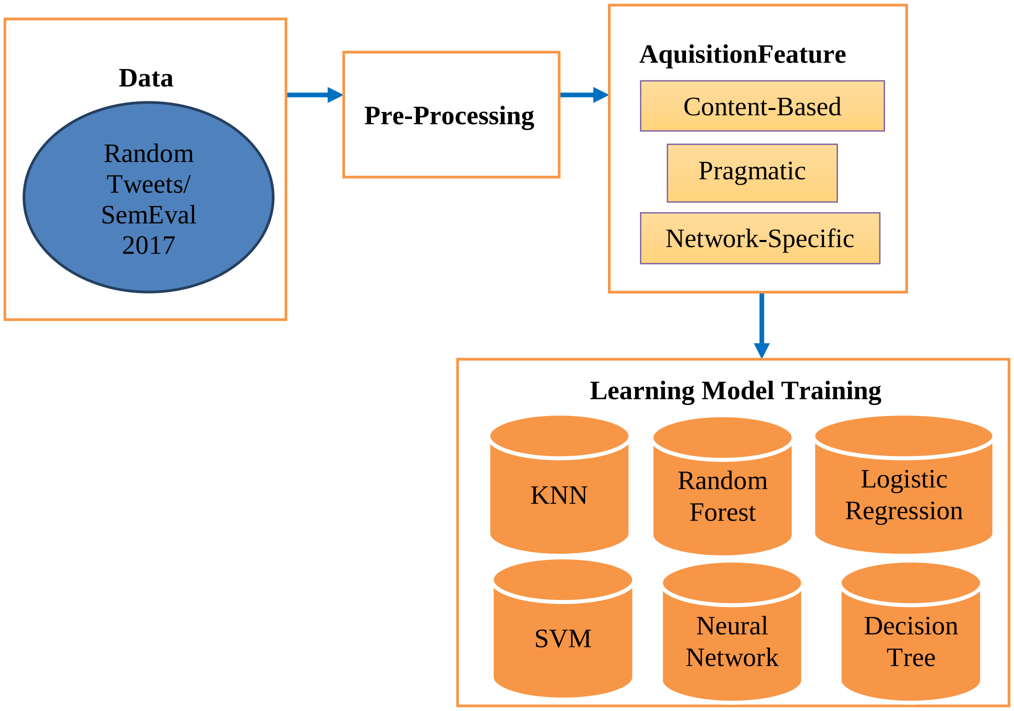


Fig. 3.4: The Veracity Classification Module

This collected data needs to be in a uniform structured format so that relevant features can be extracted. The pre-processing task includes consolidation, cleaning, transformation and reduction. The relevant features (including content based, pragmatic and network specific features) are then extracted and each tweet is then classified as being a true, false or unverified using various supervised learning techniques namely, Support Vector Machine, Decision Trees, Logistic Regression, Random Forest, K-Nearest Neighbors and Neural Network. The details of each sub-task are as follows:

3.3.1. Data Acquisition

To evaluate the system using the aforesaid learning techniques, two datasets have been examined. Firstly, a dataset with random viral tweets obtained from the first module. It includes 300 tweets on social and political issues, annotated for veracity as true, false and unverified. The benchmark corpus, SemEval 2017 Task 8.A- RumourEval dataset [6], has been additionally used to aid an improved critical assessment of the selected supervised techniques used. This dataset consists of 5568 labeled tweets. This dataset contains ten different topics; each of which has several rumorous originating tweets. The tweets are organized in a tree structure based on their reply chain and the originating rumorous tweet is the root of the tree. In total, the dataset contains 325 rumors. Each rumorous conversation contains 15 tweets in average.

3.3.2. Pre-processing

This phase works similar to the pre-processing phase of virality prediction module, presented in section 3.2.1. Pre-processing of the data is done by replacing URLs, mentions, hashtags and numbers in tweets with placeholders in order to capture the presence of URLs but not the specific details offered by these entities. Further, we employ tokenizing and stemming [35]. Tweet tokenizers are especially useful as they have been developed keeping Twitter's Internet "lingo" in mind. Additionally, all non ASCII-English characters are removed, to keep the domain of the data specific to the English language. Initial qualitative analysis of the dataset reveals that social network in cascading rumours is often significant when the users are conversing amongst themselves. Also, they may signify the named entities. Hence, ignoring mentions in the tweets would lead to loss of information.

3.3.3. Feature Extraction

This phase identifies the characteristics of the datasets that are specifically useful in predict the actual truth value of the rumor. The main aim is to find the distinguishing features that can categorize the rumour into true/false/unverified. Three different varieties of features (content-based, pragmatic & network-specific) are used to automate the identification of rumour veracity adeptly:

- ***Content-based features:*** These include the lexical (Part-of-Speech) and syntactical features (Bag of Words, term-frequency); negation relationship (syntactic and diminsher)
- ***Pragmatic features:*** These involve the semantic features such as emoticons, sentiment, anxiety related words and Named Entity.
- ***Network-specific features:*** It involves two kinds of metadata:

  (i) User metadata: Account Verification Status; Follow Ratio; Posts Count

  (ii) Message Metadata: Hashtag, URL link, Quantifiers

The following figure 3.5 illustrates the various features that are extracted:

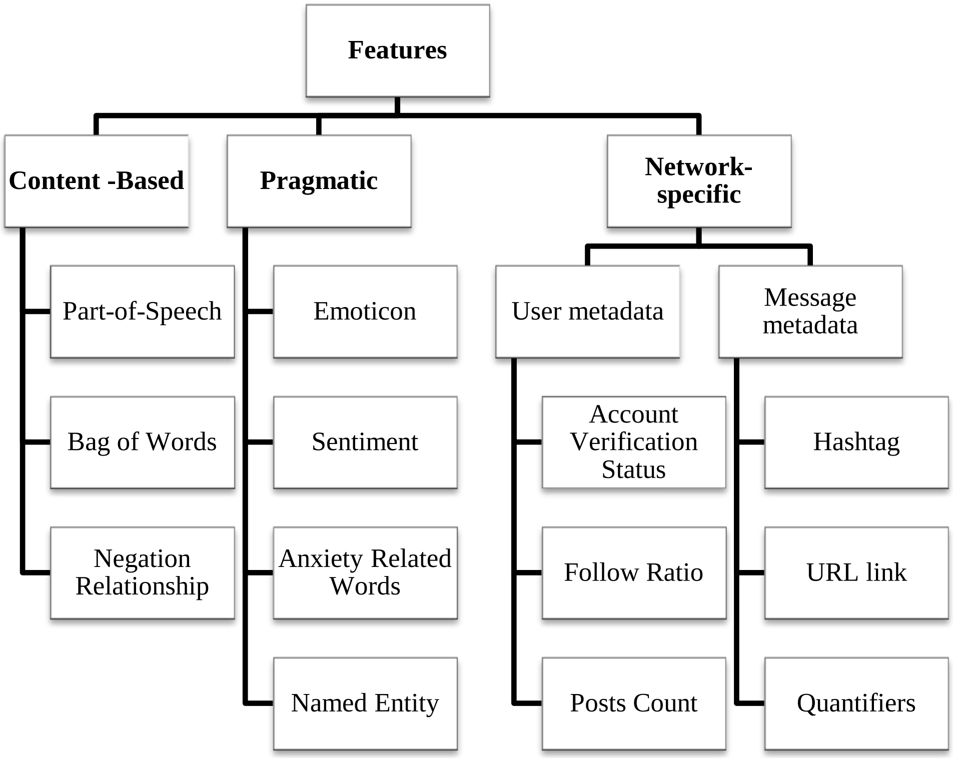


Fig. 3.5. Feature Set for Veracity Classification

The following table 3.7 summarizes the features adopted in this work:

Table 3.7: Feature Set

Feature Type	Feature Name	Description
Content-based features	Part-of-Speech (POS)	Each tweet is parsed using CMU Twitter POS tagger and assigned a binary feature value of either “1”

Content-based features		or “0”representing the presence or absence of a POS tag in tweet. A vector of POS tag counts is also maintained.
	Bag-of-Words (BOW)	Create a lexicon from all the tweets in the dataset where each word in a tweet is assigned as a feature of the lexicon. The term frequency (number of times a word occurs in the tweet) is calculated for the words occurring in the tweet and set as the feature value. For all other words “0” is used.
	Negation Relationship	Build a lexicon of syntactic and diminsher negations [34] *Syntactic:* no, not, rather, couldn’t, wasn’t, didn’t, wouldn’t, shouldn’t, weren’t, don’t, doesn’t, haven’t, hasn’t, won’t, wont, hadn’t, never, none, nobody, nothing, neither, nor, nowhere, isn’t, can’t, cannot, mustn’t, mightn’t, shan’t, without, needn’t, *Diminisher:* hardly, less, little, rarely, scarcely, seldom The feature value is binary for hasNegation is set to “1” or “0” depending on whether the tweet has a negation relationship or not.
Pragmatic features	Emoticons	A lexicon of emoticons using Wikipedia5. The emoticons are grouped by categories and its presence within a category us used as a feature with a value set to “1”. The category value is set to “0” otherwise.
	Sentiment	Sentiment refers to the use of polarities (positive and/or negative) in written text, like rumors. We used the AFINN-111 Sentiment Lexicon, which is a list of 2477 English words labeled with sentiment strength [35]. Each word is assigned with an integer in a range of polarity from -5 up to +5, negative to positive. It includes a number of words

Pragmatic features		frequently used on the Internet such as LOL (Laughing Out Loud), which are indicative of emotions of the user, especially on Twitter.
	Anxiety-related Words	We relate Anxiety to emotion analysis and thus use the adjective emotion value vector [31]. The emotions Fear, disgust and sadness are closely related to anxiety. The presence of these in the tweet, implies a value of “1” otherwise “0”.
	Named Entity	Named entities: Person, Organization, Date, Location and Money are explicitly extracted using Twitter NLP tools. The presence of a named entity tag in a tweet is depicted by a feature value of “1” otherwise “0”.
Network Specific Features	Account Verification Status	User account verification status is represented as a binary feature with values assigned as “1” if the account is verified, or “0” otherwise.
	Follow Ratio	The number of followers of a user contemplates his repute and social presence. However, users fallaciously boost their social reputation with manipulated follower count [36]. To oppress this effect, follow ratio is characterized which is the logarithmically scaled ratio of followers over followees: $\log_e 10$ (#followers/#following)
	Posts Count	This feature is deduced by the activity tracking of the user, that is, the number of tweets that a user has posted. The count is normalized by rounding up the 10-base logarithm of the posts count: $\log_e 10$ (postscount) to compensate for the count difference of posts within users.
	Hashtags	Hashtags serve as brief explanations of the tweet content. The hashtags are extracted and assigned binary value “1” or “0” depending on their presence or

Network Specific Features		absence in the tweet.
	URL link	The presence of a URL within tweet is assigned the value “1” or “0” otherwise.
	Quantifiers	This checks for the presence of statistics in rumors in the form of numerals (e.g., 1, 100), numbers (e.g., first, hundred), or quantifiers (e.g., few, much). Rumors with specific quantitative details are likely to be true. We assign the tweet “1” if it contains any statistics or “0” otherwise.

Each tweet feature described above is extracted along with its class label. These are used by the classifiers either for learning purposes when they are run in training mode or for prediction if they run in testing mode.

3.3.4. Classification

In this pen ultimate phase, the rumour is analyzed and classified for three pre-defined veracity categories namely true, false and unverified. The six supervised learning techniques are the same that were used for the virality prediction model, that is, Support Vector Machine, Decision Trees, Logistic Regression, Multi-layer Perceptron, Random Forest, K-Nearest Neighbors. Figure 3.3. and Table 3.3 expounded the details of these learning techniques.

Chapter 4 Experimental Results and Analysis

This chapter describes the experimental results and the analysis to account for the tests performed.

4.1. Virality Prediction Illustration

Basically, the Virality prediction module encompasses the following:

- Feature Engineering
- Implementation of six supervised learning techniques to empirically analyze a better classifier for adjective emotion value detection
- Quantifying the emotional value of tweet and scoring the virality of a tweet

To clearly illustrate the effectiveness of the proposed method, a case study is presented with a sample set of tweets.

Sample Tweet: Let us consider a sample tweet on trending topic #Texasshootout which has 870 retweets in 1 day and compute its emotion quotient (EQ), Polarity and Viral Value (VV_{tweet})

> After the brutal shootout in school, children harmed…bombs to kill more! I am scared :'(
> #Texasshootout #lifeunderthreat

Pre-processing of Tweets

After downloading tweets using the #topic, the data is cleaned by removing hashtags, usernames, hyperlinks, RT symbol, punctuations and non-English characters. The emoticons are transformed to the description as defined in table 3.1. Stemming and tokenization is also performed for pre-processing the tweets. Stemming is done on text in order to preserve the root of the word, for example it reduces harming to its root word i.e. harm.

> After the brutal shoot in school children harm bomb to kill more I am scare *cry*

POS Tagging

Subsequent to the preprocessing, only the adjectives, adverbs and verbs are extracted from the feature set. Each tweet is parsed using CMU Twitter POS tagger. The resultant file is a list of tweets that only have adjectives, verbs and adverbs (in the original order), which are referred to as emotion indicators.

brutal	shoot	harm	kill	more	scare	*cry*
ADJECTIVE	**VERB**	**VERB**	**VERB**	**ADVERB**	**ADJECTIVE**	**EMOTICON**

Emotion Scoring

Once the POS tagging is done, the words are scored using the crowdsourced lexicon values. The above parsed tweet is thus scored as follows:

- Here we can see that “brutal” & “scare” are adjectives, “shoot”, “kill”, “harm” are verbs, “more” is an adverb and “cry” is the description of emoticon.
- The adjective emotion values of “brutal” and “scare” are represented by the vectors [1.16, 3.65, 2.99, 3.28, 2.86] and [1.14, 3.31, 2.13, 4.09, 1.83] respectively such that the values in vector are representative of [<Happiness>, <Anger>, <Sadness>, <Fear>, <Disgust>] as shown in table 3 2
- Classifier detects the emotion polarity of adjectives as Anger for “brutal” and Fear for “scare”, which are both negative emotions, giving a polarity of -1 to the tweet
- The emotion polarity for the verbs, “shoot”, “kill” and “harm” are assigned as -1, -1. -0.9 respectively (from the table 3.4)
- In the list of adverbs we get the emotion polarity strength values of “more” as 0.2 (from the table 3.5)
- The polarity value of cry from emoticon table 3.1 is -1
- Now using equation (3.1), the EQ of the tweet will be computed as follows:

$$EQ = \frac{1}{1+1+1+1}\left\{\left[\left|\frac{E_{brutal}+E_{scare}}{2*5}\right|\right] + \left[\left|\frac{E_{shoot}+E_{kill}+E_{harm}}{3}\right|\right] + \left[\left|\frac{E_{more}}{1}\right|\right] + \left[\left|\frac{E_{cry}}{1}\right|\right]\right\}$$

$$= \frac{1}{4}\left\{\left|\frac{3.65 + 4.09}{10}\right| + \left[\left|\frac{(-1) + (-1) + (-0.9)}{3}\right|\right] + \left[\left|\frac{0.2}{1}\right|\right] + \left[\left|\frac{-1}{1}\right|\right]\right\}$$

$$= \frac{1}{4}\left\{[0.774] + \left[\frac{0.9667}{3}\right] + [0.2] + [1]\right\} = 0.25\{[0.774] + [0.322] + [0.2] + [1]\}$$

$$= 0.25 \times 2.296 = 0.574$$

Thus, the EQ of the Tweet is 0.574 and the polarity from classifier is negative, -1.

- Now using equation (3.2), the VV_{tweet} is computed as follows

$$VV_{tweet} = (-1)\left[\frac{0.574 * 870}{1}\right] = -499.38$$

- Similarly we calculate the values for the other tweets on the same topic as shown in the following table 4.1:

Table 4.1: Illustration of Scoring Module

Original Tweet	**Features**	**Emotion Quotient$_{Tweet}$**	**Polarity**	**Retweet**	**Life-span**	**Viral Value$_{Tweet}$**
This is pretty serious…We will all be killed </3 :’(#texasshootout #scared	Pretty(Adv) serious (Adj) all (Adv) kill (Vb) *broken heart cry (Emoti)*	0.6485	-1	1105	1	-716.59
More kills! They are terrorist! School children hurt X-(=O #texasshootout #godhelp	More (Adv) kill (Vb) hurt (Vb) *angry shocked (Emoti)*	0.6166	-1	700	1	-431.62
Innocent people & children killed. Are they humans? Terrible it is Xp X-(:-o #texasshootout #rip #inhuman	Innocent (Adj) kill (Vb) hate (Vb) terrible (Adv) *disgusted angry shocked*	0.842	-1	1402	1	-1180.48
Bravo! Great work… the school for rich people! :-D :P #texasshootout	Great (Adj) work (Vb) rich (Adj) *big grin*	0.756	+1	247	1	+186.90

#wedeserveit						
Bombs to kill planted! Highways closed as extreme violence reported. Scared to death ;(=O #texasshootout #disturbed	kill (Vb) plant (Vb) close (Vb) extreme (Adv) scare(Adj) *cry shocked*	0.767	-1	3762	1	-2885.45

- Using equation (3.2), the virality of each tweet is obtained. As discussed in chapter 3, the threshold for a tweet being called "viral" has been set to 400. So any value of virality greater than 400 implies that the tweet has a cascading effect and steps to authenticate its accuracy and origin must be taken by agencies (business or government). The + and – simply indicate the polarity of the post.

Thus, using the proposed virality prediction module the likelihood of tweet going viral can be determined and this can be an initial step to identify and highlight potential rumours with questionable veracity.

As an output of this module, we obtain a list of potential rumours which is passed to the next module for each post's (tweet's) truth value check.

4.2. Veracity Classification Results

This section highlights the results and observations related to performance of various supervised learning techniques used in this study for veracity classification in tweets using performance measures like precision (P), recall(R), accuracy (A) and F-measure (F) [38]. The empirical analysis results demonstrate that the veracity classifier effectively finds the truth value of the tweet. The classifiers are run for both random tweets and RumourEval Datasets. The preliminary results are clearly motivating.

Table 4.2. describes the efficacy measures that are used to quantify the classifier performance:

Table 4.2: Efficacy Measures used

Measure	Description
Accuracy	Accuracy refers to the closeness of a measured value to a standard or known value. It is the proportion of the total number of predictions that were correct.
Precision	Precision denotes the proportion of predicted positive cases that are actually positive. It's the primary measure used to observe effectiveness of computational techniques, and is used widely in information retrieval systems. It's also known as Confidence.
Recall	Also known as Sensitivity, Recall is defined as the ratio of predicted positive cases that were actually positive to the total positive cases. It helps measure coverage of real positive cases[38]. It tends to play second fiddle to Precision in information retrieval. However, it does play a significant role in computational linguistics, where accurate translation of expressions is paramount.
F- Measure	The F measure (F1 score or F score) is defined as the weighted harmonic mean of the precision and recall of the test. It's a combined metric which determines robustness and effectiveness.

The following table 4.3 describes the results of the classifier run on random tweets collected on social and political issues**:**

Table 4.3: Classifier Performance Results for Random Tweets

Measures →	**A**	**P**	**R**	**F**
Techniques				
KNN	70.6	0.71	0.71	0.71
SVM	85.6	0.85	0.86	0.86
DT	84.8	0.85	0.85	0.85
RF	89.1	0.89	0.89	0.89
MLP (NN)	91.2	0.90	0.92	0.91
LR	92.0	0.91	0.92	0.92

It is observed that Logistic Regression and Neural Networks give the highest accuracy scores (92% and 91.2% respectively). As the data was crisp and concise, high values for all four metrics were observed. Next to it are RF and SVM depicting 89.1% and 85.6% accuracy. DT came next with a comparable accuracy of 84.8%. KNN showed the lowest accuracy of around 71%. The following Figures, namely, fig.4.1, 4.2, 4.3 and 4.4 depicts the results shown in table with the help for graphs.

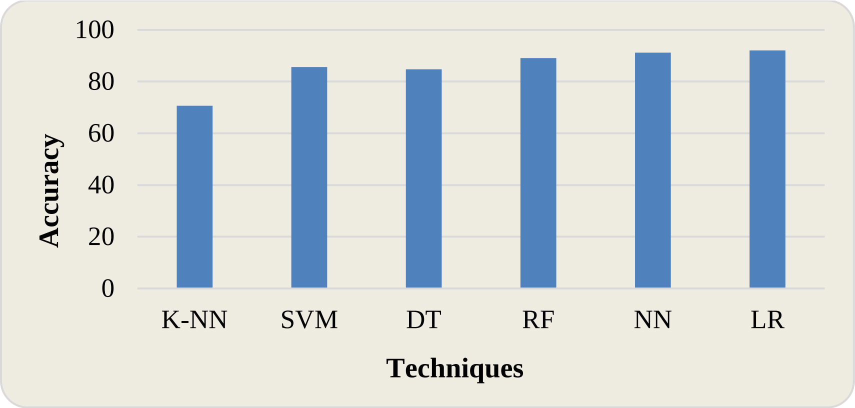


Fig.4.1. Accuracy (Random Tweets)

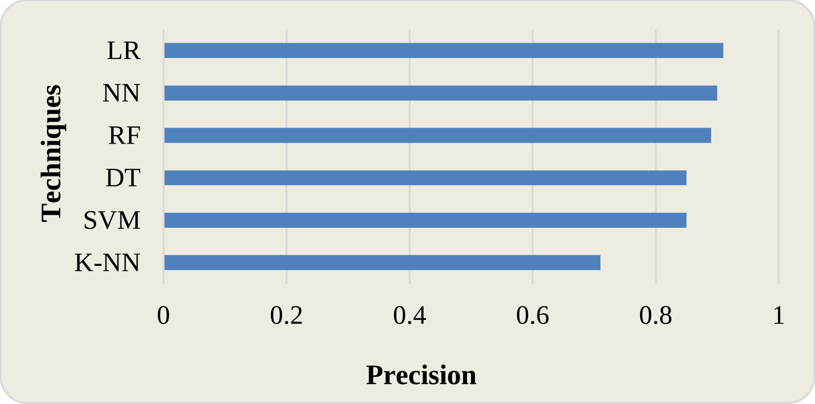


Fig.4.2. Precision (Random Tweets)

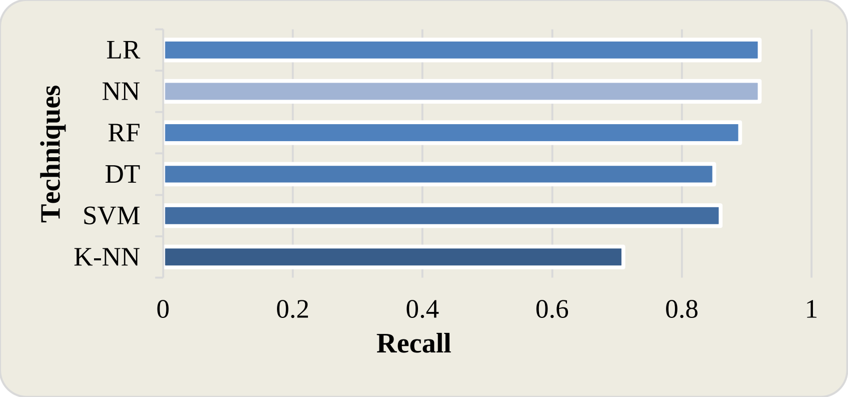


Fig.4.3. Recall (Random Tweets)

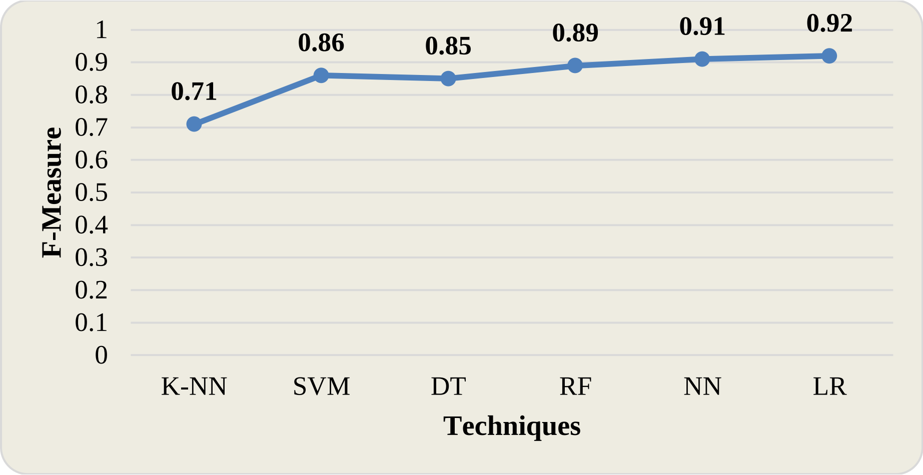


Fig.4.4: F-Measure (Random Tweets)

The classifier performance for the RumourEval benchmark corpus was observed to be quite similar to the results obtained for Random Tweets. The following table 4.4 describes the results of the classifier run on RumourEval benchmark corpus**:**

Table 4.4.: Classifier Performance Results for RumourEval Dataset

Measures →	**A**	**P**	**R**	**F**
Techniques				
KNN	72.1	0.72	0.72	0.72
SVM	85.8	0.86	0.86	0.86
DT	84.9	0.85	0.85	0.85
RF	90.2	0.89	0.89	0.89
MLP (NN)	91.5	0.91	0.92	0.92
LR	92.7	0.92	0.93	0.93

It is again observed that Logistic Regression and Neural Networks give the highest accuracy scores (92.7% and 91.5% respectively). As the data was crisp and concise, high values for all four metrics were observed. Next to it are RF and SVM depicting 90.2% and 85.8% accuracy. DT came next with a comparable accuracy of 84.9%. KNN showed the lowest accuracy of around 72%. The following Figures, namely, fig.4.5, 4.6, 4.7 and 4.8 depicts the results shown in table with the help for graphs.

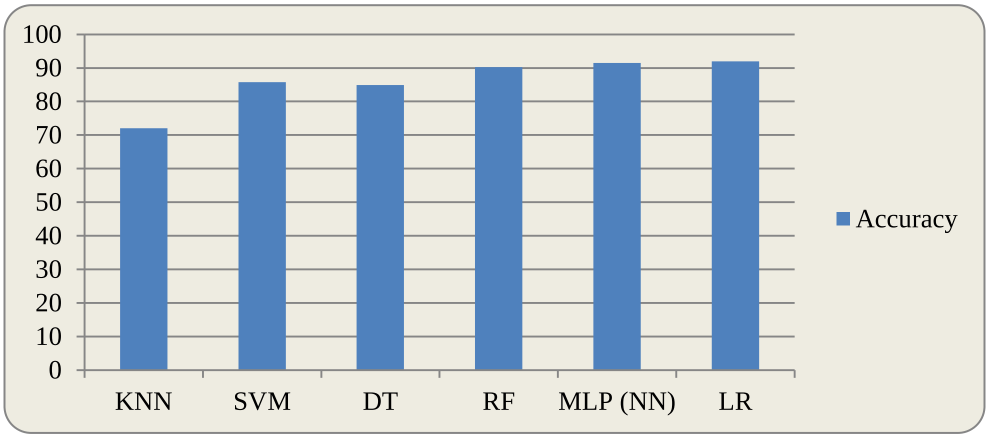
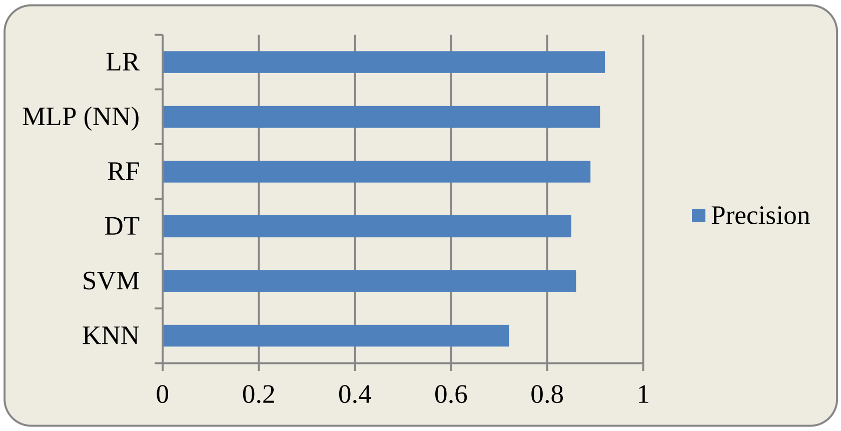


Fig.4.5. Accuracy (RumourEval)



\

Fig.4.6. Precision (RumourEval)

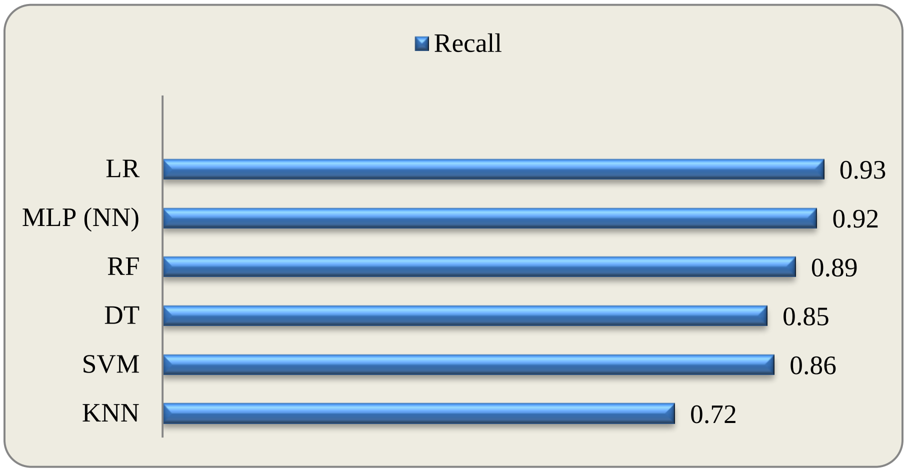


Fig.4.7. Recall (RumourEval)

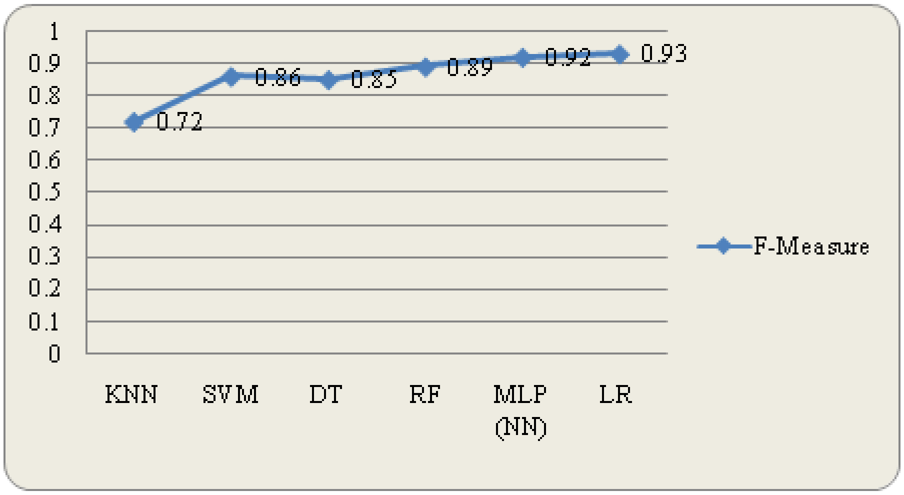


Fig.4.8. F-Measure (RumourEval)

It is interesting to note that using Ensemble methods such as Random Forests gives improved and enhanced results in comparison to the traditional single Decision Tree model for both the datasets considered.

Chapter 5 Conclusion and Future Scope

5.1. Research Summary

Sharing online content is an indispensable part of our contemporary lives. A typical rumour cycle starts with stories and/or statements that are generally circulated without confirmation or certainty as to facts. *Social media* instantly comes alive encouraging, spreading and breeding these rumours. Consequently, it becomes exceedingly imperative to resolve the authenticity of information and promptly inhibit them from spreading among the Internet users as it can jeopardize the well-being of the citizens. To capture the truthfulness of these stories the veracity of rumours needs to be timely and effectively established. The proposed rumour detection model, the VRV model (**V**irality-**R**umour-**V**eracity Model), primarily determined the likelihood of tweet going viral based on the strength of emotion and number of re-tweets. The hybrid approach made use of natural language textual cues of emotions from parts-of-speech like adjectives, verbs, adverbs and emoticons. The virality of social and political tweets was perceived accurately using the scoring module. Once the list of potentially rumourous tweets was prepared using this virality prediction module, these rumours were checked for their truth values to debunk false rumours and mitigate their spread and impact. Three set of feature categories were used to train and test the six classifiers and the results were evaluated on two datasets, that is, Random Tweets and the benchmark RumourEval dataset. Empirical evaluation of supervised learning techniques yields the best results for logistic regression and neural networks (multi-layer perceptron).

Thus, the objective of this research to find the list of potentially rumourous tweets and then use classifier to automatically determine its actual truth value with accuracy & without delay is accomplished. The key contributions of this research are as follows:

- A model to detect rumour and predict its veracity value (truth value) automatically
- Correlation of virality to rumour is characterterized in order to acquire a list of potential rumours in real-time, for which the truth value needs to be determined. Virality detection is applied as an initial step to identify and highlight information with questionable veracity. Precisely, viral rumours are major carriers of panic.

- Veracity Classification based on three variety of feature sets: Content-based, Pragmatic, & Network-based
- Empirical Analysis on two datasets (Random Tweets & RumourEval 2018) to find out the best veracity classifier.

The following figure 5.1 summarizes the research concept of rumour detection exemplified in this thesis.

.

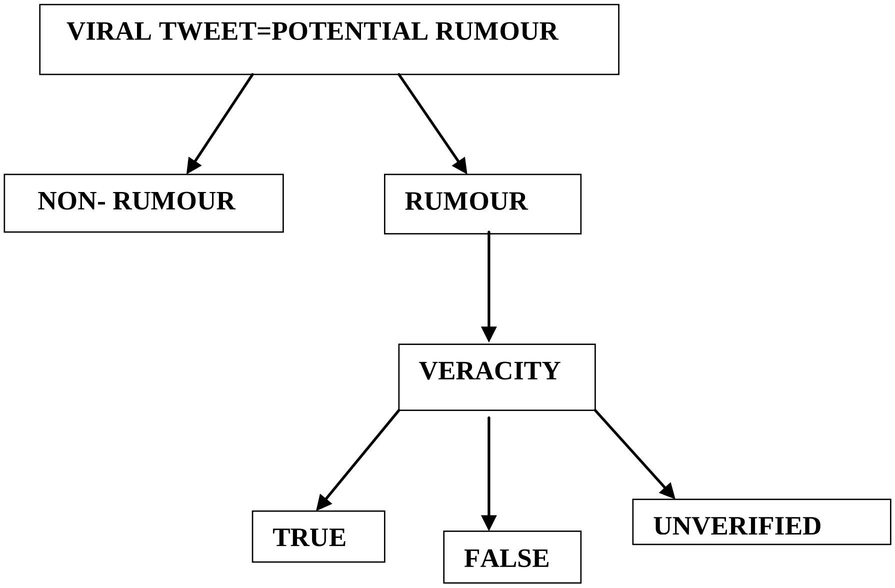


Fig. 5.1. Research concept of Rumour Detection

5.2. Future Research Directions

As a future direction of work, the fluctuations in emotions can be captured as they convey uncertainty towards a topic and may further assist in veracity check or rumour stance detection. Also, contextual information within the post can be assessed for virality prediction. As another promising direction of future research, the use of evolutionary and swarm techniques can be studied and validated for optimal feature selection which can improve classifier's performance. Also, tools for rumour prediction in mash-up tweets (written in mash-up languages, example: Hinglish =Hindi+English) are open for research.