Multimedia Tools and Applications (2022) 81:40675–40707
https://doi.org/10.1007/s11042-022-13129-y

# Disinformation detection on social media: An integrated approach

Shubhangi Rastogi[1] · Divya Bansal[1]

Received: 12 November 2020 / Revised: 10 March 2021 / Accepted: 10 April 2022 /
Published online: 12 May 2022

## Abstract

The emergence of social media platforms has amplified the dissemination of false information in various forms. Social media gives rise to virtual societies by providing freedom of expression to users in a democracy. Due to the presence of echo chambers on social media, social science studies play a vital role in the spread of false news. To this aim, we provide a comprehensive framework that is adapted from several scholarly studies. The framework is capable of detecting information into various types, namely real, disinformation and satire based on authenticity as well as intention. The process highlights the use of interdisciplinary approaches derived from fundamental theories of social science and integrating them with modern computational tools and techniques. Few of these theories claim that malicious users suggest writing fabricated content in a different style to attract the audience. Style-based methods evaluate the intention i.e., the content is written with an intent to mislead the audience or not. However, the writing style can be deceptive. Thus, it is important to involve user-oriented social information to improve model strength. Therefore, the paper used an integrated approach by combining style based and propagation-based features with a total of thirty-one features. The extracted features are divided into ten categories: relative frequency, quantity, complexity, uncertainty, sentiment, subjectivity, diversity, informality, additional, and popularity. The features have been iteratively utilized by supervised classifiers and then selected the best-correlated ones using the ANOVA test. Our experimental results have shown that the selected features are able to distinguish real from disinformation and satirical news. It has been observed that the Ensemble machine learning model outperformed other models over the developed multi-labelled corpus.

**Keywords** Covid-19 · Fake · Disinformation · Satire · Machine learning · Neural network · Ensemble

✉ Shubhangi Rastogi
shubhangi05.rastogi@gmail.com

[1] Department of Computer Science and Engineering, Punjab Engineering College (Deemed to be University), Chandigarh, India

## 1 Introduction

In recent years, the prevalence of social media platforms has acted as a catalyst and resulted in an exponential spike in popularity [29]. Thus, societies and social behavior can now be mapped on these online social networks (OSNs) which in turn give rise to virtual societies. By studying these virtual societies, human behavior patterns can be analyzed to gauge a better understanding of societies by leveraging computational tools [52]. However, the absence of fact-checking of the posts and the unregulated nature of the internet makes social media a fertile ground for the spread of unverified and false information. Since the amount of data posted on social media is huge, manual fact-checking is not a feasible solution. Therefore, computational tools can be very effective in developing techniques to counter false news in an automated manner. To this aim, we provide a comprehensive framework that has been adapted from several scholarly studies. The framework is capable of detecting and classifying information into various types: real, disinformation and satire. The process highlights the use of interdisciplinary approaches deriving from fundamental theories of social sciences and integrating them with modern computational tools and techniques. The spreading of disinformation is an old problem and with the use of OSNs, the spread has become exponential. There do exist theories in social science that have the solution but have not been considered while developing tools and techniques to fight these issues. This study combines the two concepts and generalizes them into four perspectives, namely knowledge-based, style-based, propagation-based and source-based. For instance, as per psychology, a fabricated text with an intent to harm the public is written differently as compared to real text (Undeutsch hypothesis) [5]. Throughout history, false news has been used deliberately to manipulate the beliefs and opinions of people. Ancient Indian mythology, ’Mahabharata’ has the earliest reference of fake news wherein false news was spread to kill Dronacharya, the guru of the Pandavas and Kauravas [7]. Also, during World War II, the allied forces planned ’Operation Mincemeat’ which was a successful British deception operation. Earlier fake news had limited impact but due to online sources of information like OSNs, the reach of fake news has become global. This reach of fake news was best highlighted during the 2016 US presidential election which remains under investigation. During the Covid-19 pandemic, social media platforms have become a key forum for the dissemination of information rapidly. At the same time, the huge amount of health-threatening false information is spreading faster than the virus itself. Moreover, when the whole world was suffering from the pandemic, India went through communal riots that seemed to have been caused by disinformation. This paper highlights this recent event (as described in Section 6.1) which was portrayed as propaganda by a community in order to harm public interests. Clearly, the instances are evident that the motive of these platforms is to get users engaged to earn business revenues rather than providing factual information. Also, social media platforms suffer from echo chambers due to which users see their point of interests without dwelling on facts [20]. Thus, users get trapped in propaganda rather than following authentic news which many a time can also lead to a national crisis. The issue is complex, serious and multi-faceted. To this end, some of the highlights of this paper are: (1) Mainly, studies have considered politics as a domain to detect fake news. Our work introduces a multi-labelled corpus related to an event that happened due to disinformation. This can be used for the development of a cross-domain fake news detection model. Also, a complete annotation guideline has been provided to consider the authenticity and intention of the news; (2) Our work provides an integrated approach by combining two perspectives (style and social context-based); (3) Existing studies highlight various features used to differentiate disinformation from real news, but features inspired from fundamental theories are more explainable; (4) ANOVA statistical technique has been used to select significant features in order to distinguish disinformation from real and satire news. The selected features have performed well on our dataset. Figure 1 summarizes the approach followed in the paper.

### 1.1 The problem

The overarching goal of this research is to automate the process of finding the probability of a particular tweet being disinformation, satire or real news using an integrated approach on a check worthy and undiscovered domain. We define the null hypothesis and alternate hypothesis as follows: Null Hypothesis, $H_o$: There is no significant difference between means of features for disinformation, satire and real news.

Alternate Hypothesis, $H_a$: There is a significant difference between means of features for disinformation, satire and real news.

Since this is a multiclassification problem, ANOVA (Analysis of Variance) [15] statistical test has been performed on each feature as shown in Table 2. The p-value obtained for most of the features is less than or equal to 0.05, which indicates that we may reject the null hypothesis and accept the alternate hypothesis. This has been further explained in detail in Section 4.3. Therefore, we defined the problem statement as:

*"Given a tweet feature matrix, popularity matrix, semantic matrix, multiclass partial label vector, and a generated secondary matrix, we aim to predict remaining unlabeled tweets vector."*

### 1.2 Key contributions

The literature highlights various research gaps which have been explored in this paper. Hence, following are the key contributions based on the potential research tasks in the future scope of various studies to improve the efficiency of the current fake news detection model:

- C1. Intention-based detection of fake content The study considers authenticity as well as intention to measure fake content. Style-based features help to capture the intention. Also, intention depends on data labels. To the best of our knowledge, current studies have not provided clear guidelines of how annotators have manually evaluated text to find out the intention behind sharing political information. To consider the intent of social media posts, this paper describes the complete guidelines for manual annotation in Section 3.1.
- C2. Integrated Approach The four perspectives described in the introduction of this paper are not independent of each other and it is highly desirable to predict fake news using features from multiple perspectives jointly. The style-based approaches capture intention but heavily depend on the writing style which thus varies with the domain, language and time. Thus, the paper has described the formulation of an integrated approach by combining social features and style-based features derived from fundamental social science theories. The features inspired by well-established theories are more explainable and helps to detect disinformation accurately.
- C3. Cross-domain fake news analysis Current studies to detect fake news primarily consider politics as a domain. This paper presents an analysis of a recent incident related to communal national riots which got amplified due to disinformation spread on social media platforms. The work will form the basis to build a comprehensive fake news detection model as a part of ongoing research work.

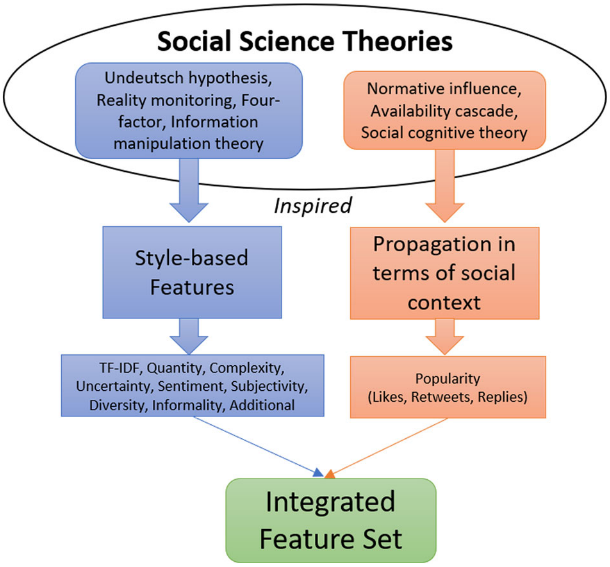

Social Science Theories
Undeutsch hypothesis, Reality monitoring, Four-factor, Information manipulation theory
Normative influence, Availability cascade, Social cognitive theory
Inspired
Style-based Features
Propagation in terms of social context
TF-IDF, Quantity, Complexity, Uncertainty, Sentiment, Subjectivity, Diversity, Informality, Additional
Popularity (Likes, Retweets, Replies)
Integrated Feature Set

**Fig. 1** Background Flow: The figure summarises the approach followed

- C4. Identifying check worthy topics The topic or event is check-worthy if the content causes extensive debates on social media, relates to national affairs and has the historical potentials of being fake. In this paper, a check worthy topic described in Section 6.1 (i.e., case study) has been considered concerning the national crisis.
- C5. Use of multilabel classification to find out the veracity of fake content To detect partially correct news, a multilabel classification is required. In this paper, we are predicting the probability of being fake which has further been used to scale the text from 1 to 5.

The remaining sections are structured as follows. Section 2 outlines the background of different terms related to false information, fundamental social science theories, an overview of various perspectives and the review of related works. Section 3 introduces the methodology and proposed framework. Section 4 presents the research experiments conducted to evaluate features iteratively and respective results with the intent to find the most suitable model. Section 5 states the benchmark studies and visualization. Section 6 gives the case study with limitations. Section 7 makes concluding remarks.

## 2 Background and related works

### 2.1 Important concepts related to fake news

The problem with social media posts leading to the National crisis is not that the information is completely false or certain events never happened. Rather, it is the misleading context presented in the posts with possibly an intent to harm which does most of the actual damage. Hence, a major focus is to find the intention of spreading false information. Literature provides different terms related to fake news such as misinformation, satire and disinformation and many more based on authenticity and intention. However, there is no universal definition available in the literature since it varies with the account of event [38]. Broadly literature has defined these terms based on intention and authenticity. Hence, false information with the intent of causing harm is called disinformation; False information with no intention to harm the audience is called misinformation; whereas information with the intent to entertain the audience and created for fun are called satirical news [33, 57, 60]. Furthermore, to develop a solid foundation for false news analysis, a clear definition is given below for each category which has been used for representation purpose in this paper:

$$label(t_i) = \begin{cases} 0 & \text{if } t_i \text{ is verified and unbiased} \\ 1 & \text{if } t_i \text{ false and intention is to mislead} \\ 2 & \text{if } t_i \text{ is false but intention is to entertain} \end{cases}$$

For authenticity, the government should encourage credible sources of information without compromising on freedom of expression which itself is very challenging to ensure. The credible sources can be maintained by expert domains who check the authenticity of information manually. But it is impractical to manually check this voluminous data on social media against credible sources. Therefore, this paper aims at automating this process to flag data that has a high probability of being fake. Essentially, false information with an intent to harm is written in such a way that it can deceive the targeted audience. Hence, for the analysis of intent, different social science theories have been studied [62]. Although, the intent analysis does require some level of manual annotation and the accuracy of such annotations leads to accurate machine learning models.

### 2.2 Fundamental theories

The problem of detecting false information requires inter-disciplinary approaches derived from areas like psychology, philosophy, economics and others [61]. Therefore, this paper has identified fundamental social science theories which can be potentially used to understand the problem. In our work, we have mapped these theories to important features used in social media as shown in Table 1.

### 2.3 The four perspectives for detecting fake news

Zhou et al. [60] have specified the four perspectives to detect fake news (and related terms): Knowledge-based; Style-based; Propagation-based, and Source-based. Knowledge-based methods use fact-checking (manual or automatic) to check the authenticity of text. To check the intent of spreading false information, style-based approaches are considered. Style based methods work on an assumption that malicious users practice a different style of writing to attract an audience to trust them. For example, writing text with extreme emotions. The writing style parameters such as text written with extreme emotions, certainty words etc. are focused to distinguish fake text from real. Largely, textual features are categorized into two categories: (1) General features: lexicon, syntax and semantic language-based features [12]; (2) Latent features: text embeddings at the word level, sentence level to give vectors such as word2vec, doc2vec [37]. Recently, Zhou et al. [62] have claimed that general features outperform latent features and also combined features outperform single level features over machine learning classification models. Style based features are useful in early fake news detection before the news gets spread deeply on the social media platform. Though

**Table 1** Theories in Social Sciences helpful in deterring the spread of false information

| Theory | Description | Features | |
|---|---|---|---|
| News-related theories | Undeutsch hypothesis [5] | A statement derived from real life experiences differ in content and quality from that of fabricated. | Writing style |
| | Reality monitoring [28] | Facts contain more detailed sensory information | More unique words |
| | Four-factor theory [63] | Deception are stated differently in terms of arousal, guilt, emotion and try to appear as real. | Sentiments |
| | Information manipulation theory [36] | Deception contains extensive quantity of information | Word counts |
| Social Impact theories | Normative influence theory [13] | People's behavior gets influenced by the society to be more accepted and liked. | Likes |
| | Availability cascade [34] | People accept others acuities when such acuities are gaining more popularity within their social circles. | Retweets |
| | Social cognitive theory [8] | People get actively molded by their surroundings. | Popularity |

there are challenges to this approach such as different domains have different stylistic patterns, malicious users can use different writing style to hide the deception (domain, language and time dependence). To overcome these challenges, social features are required to be combined with propagation-based methods. Propagation based methods are network-based approaches that can be used to check how differently fake news is propagating in comparison to real news. Different additional information such as page rank, stance scores have also been included to check the variety in propagation patterns. It has been observed that fake news (especially political fake news) spread faster, farther and becomes more popular than facts. However, propagation features are insufficient for the early detection of fake news since limited information is available before the news spreads. Nevertheless, Zhao et al. [59] and other researchers have observed that fake news spreads differently than real news at an early stage of propagation. Moreover, some studies have also considered user engagements to detect compromised accounts on Twitter [47]. Finally, source-based methods check the credibility of creator, publisher and spreader (social media accounts) in order to detect fake news indirectly [10]. Accounts can either belong to human users or non-human users (i.e., bots). Source-based approaches work on an assumption that articles posted by unauthenticated users are considered fake. To check the authenticity of an account different features like registration age, the number of friends and followers, history of the user to spread fake content, post frequency and anomalous analysis of cyborgs and bots have been explored in the literature. The major challenge with this approach is the presence of echo chambers in social media platforms. The source-based approaches are useful in fake news mitigation. Notably, the four perspectives are not independent and should be used together for effective detection of fake content. In this paper, we propose an integrated approach using multi-perspective features listed in Table 2. Some of the features are selected from Zhou's ten-dimensional feature set, while some novel additional features have been added. Also, propagation-based features depicting popularity in terms of user engagement have been integrated to predict fake news from multiple perspectives. In our proposed approach, we have made an assumption that sources have been chosen in a balanced form to avoid biases and echo chambers. Hence, the analysis has not considered source-based features.

Depending on the four perspectives described in the introduction, current fake news studies are majorly based on style-based and propagation-based approaches. We have reviewed recent advancements in both approaches with research based on the Latent approach and presented in a tabular form (as shown in Table 3) for better understanding. Many previous studies are based on style-based features and have shown promising results. In the literature, popular style based features are linguistic features such as n-grams [1], psycholinguistic features using LIWC, number of punctuations, stopwords, readability scores (eg., number of complex words, long words, syllables, characters) [41], syntax and dictionary-based features [43]. Psychological features such as sentiment, emotion are strong differentiating factors between fake and real content [2, 25]. Siering et al. proposed a framework based on the verbal cues of the content (e.g., Average sentence length, subjectivity, PoS) to know the deception process, the psychology of fake spreaders and type of cues [51]. Nonetheless, Zhang et al. introduced non-verbal features i.e. the social behaviour of a user (e.g., follower count, photo count, posting rate) which improved the performance of the model [58]. Style-based features are useful in the early detection of fake news but there occur domain, language and time dependence. To overcome this challenge, propagation and user-oriented features have been presented by several studies. Extant models for automated fake news detection rely on user-oriented features on social media platforms such as the number of likes, retweets, shares, replies and comments [21, 35]. Due to the low cost of creating deceitful websites and high volumes of software-controlled profiles (social bots), the problem of misinformation has become more complex. These social bots can post content in bulk and target an audience that is easy to deceive [49]. Therefore, user-oriented features play a vital role in the development of a fake news detection model. Some of the studies have followed a reverse approach of finding users who are more inclined to share fake news in OSNs. Different supervised classification models have been tested over combined features including stylometry, personality, emotion and embeddings [16]. The assembled features have also been employed to detect fake profiles [39]. Furthermore, obtaining a balanced dataset from social media platforms is also an issue since the number of real news is assumed to be larger than fake news. Overflow of posts, comments and other user engagements on social media demonstrates the deployment of big data strategies. Some studies

**Table 2** Extracted Features based on perspectives

| Approach | Features | Description |
|---|---|---|
| Style-based features | TF-IDF (F1) | Relative frequency of words |
| | Quantity (F2) | # Characters |
| | | # Words |
| | | # Noun Phrases |
| | | # Sentences |
| | Complexity (F3) | Average # characters per word |
| | | Average # words per sentence |
| | | Average # punctuations per sentence |
| | Uncertainty (F4) | # Modal verbs |
| | | # Certainty terms |
| | | # Generalizing terms |
| | | # Tentative terms |
| | | # Numbers and quantifiers |
| | | # Question marks |
| | Sentiment (F5) | # Positive words |
| | | # Negative words |
| | | # Anxiety/angry/sadness words (emotion) |
| | | # Exclamation marks |
| | | Content sentiment polarity |
| | Subjectivity (F6) | # Subjective verbs |
| | Diversity (F7) | # Unique words |
| | | # Unique nouns, verbs, adjectives, adverbs |
| | Informality (F8) | # Typos/spellchecks |
| | | # Swear words/ netspeak/assent/fillers |
| | Additional (F9) | # Hashtags |
| | | # Mentions |
| | | # Stopwords |
| | | # URL |
| | | Mean word length |
| User engagement features | Popularity (F10) | # Likes |
| | | # Retweets |
| | | # Replies |

have applied novel approaches to handle veracity in big data such as TF-IDF with temporal Locuvain approach for categorization and clustering for appropriate document gathering [26, 27]. Also, data annotation is a challenging task in any domain. Some studies presented semi-automated tools to reduce data annotation time [48], while few introduced web-based annotation tools such as BRAT using Natural Language Processing (NLP) technique [53]. Few studies in the state-of-the-art have employed feature-based techniques for annotation in interdisciplinary domains [6, 30]. However, the majority of studies for this domain (i.e., fake news) have done manual data annotation through human annotators with domain expertise. Generally, annotated data can be obtained using different ways such as fact-checking websites (e.g., “Snopes” or “PolitiFact”) which are mainly focussed on one domain (mostly

**Table 3** Literature Review

| Approach | Ref. | Type & Dataset | Purpose | Features | Result |
|---|---|---|---|---|---|
| Style-based | [41] | News articles; Crowd-sourced dataset 240 fake & 240 real | Exploratory analysis on identification of linguistic differences in fake and legitimate news content | Linguistic features: ngrams, punctuation, psycholinguistic, readability, and syntax | 74% accuracy with SVM |
| | [43] | News articles; 1627 articles from Buzzfeed corpus | Hyperpartisan news can be differentiated well by its style of writing from mainstream | ngrams, readability scores, dictionary-based features | 90% accuracywith linear classifier |
| | [51] | Crowdfunding platform *kickstarter.com*; suspended and non-suspended projects (652) | To examine the importance of content-based and linguistic cues to detect fraud | Linguistic (diversity, uncertainty, informality etc.) and content-based cues (Bag-ofwords) | 77.45% accuracy using ensemble classifier |
| | [58] | Customer reviews from *Yelp.com*;1,033 authentic reviews and 1,100 fake reviews on restaurants | Compare verbal and non-verbal features in order to detect fake online reviews | Non-verbal (membership length, tips count, photo count) and verbal features (review length, average sentence length, noun ratio) | 87% accuracy over random forest classifier |
| | [25] | Facebook and news sites; Buzzfeed election data (120), real fake satire websites (224), Burfoot and Baldwin data (4233) | Fake and satire news are distinguishable using stylistic features of the title over machine learning | stylistic, complexity and psychological features | 91% accuracy with SVM |
| | [1] | News articles Real from Reuters & fake from Kaggle (12600) | Fake news detection using n-grams analysis | TF-IDF with n-grams | 92% with Linear SVM |
| | [2] | Social media; Twitter API | Comparison of different feature sets in order to detect sarcastic articles | Sentiment, complexity and language based features | 80% accuracy with gradient boosting |
| Social-based | [49] | Social media; 14million messages Twitter | Social bots heavily spread low credibility content by retweeting, replying those posts | Replies, retweets, mentions, sentiment etc. | 94% accuracy using binary classifier |

**Table 3** (continued)

| Approach | Ref. | Type & Dataset | Purpose | Features | Result |
|---|---|---|---|---|---|
| | [35] | Social media; 343645 Twitter messages | Sentiment analysis of Twitter discussions on US presidential elections and also, to distinguish misinformation from negative information | Retweets, likes, replies, sentiments, number of followers | Observations |
| | [21] | Social media; 10.8M Twitter posts and 6.2M Reddit comments | To develop neural network for classifying deceptive and trusted news sources in terms of speed and types of reactions | Reactions like answer, appreciation, elaboration, question | Observations |
| Latent-based | [23] | News articles and fact checking websites;1400 true and 2004 fake | To find out the best combination of features and word embeddings to detect deception accurately | Linguistic features | 95% accuracy with adaboost ensemble classifier |
| | [24] | Exisiting datasets; Amazon (21000) & Hotel datasets (800) | To improve the performance of fake customer reviews detection model | n-grams, word embeddings, lexicon-based emotion indicators | 89.56% & 82.80% with proposed DFFNN model |

politics), satirical websites (e.g., “The Onion” or “Faking News”), crowdsourcing services (AMT), industry detectors and expert journalists. Researchers also collected fake and real data by targeting well known fake and real news sources. Horne et al. [25] also listed a few real, fake and satire news sources to avoid manual annotation of the scraped online data. In this paper, annotation guidelines have been carefully designed to consider both authenticity and intention which can be useful in future annotations. As per our knowledge, the complete annotation guidelines have not been mentioned in the previous studies. Typically, the literature showed various research gaps which have been incorporated in this paper as key contributions in Section 1.2. For instance, literature majorly considered political data thus, it is required to explore other domains. Therefore, in this research, we investigated the latest sensitive event that happened in India in which different narratives were presented on media. To the best of our knowledge, this worth attention event has not been investigated yet. Also, no study has adapted social science theories to derive effective features which provide the foundation for our study. Therefore, we proposed a highly accurate model to detect real, fake and satire news using a set of effective features from multiple perspectives.

The following sections present the research design and methodology along with the experimental results and model interpretation.

# 3 Research design and methodology

## 3.1 Dataset and annotation

Data has been collected from Twitter using top trending hashtags in India on Nizamuddin Tablighi Jamaat case using most relevant keywords such as Nizamuddin, TablighiJamaat, CoronaJihad, TablighiVirus, Islamophobia etc. We collected tweets starting from 29th March 2020 to 14th April 2020 because it was the time period when the topic was trending the most on Twitter. We have outlined the steps involved in building the dataset in Fig. 2. In this paper, a multiclass corpus named Fak_ov (with three labels: real, fake and satire) has

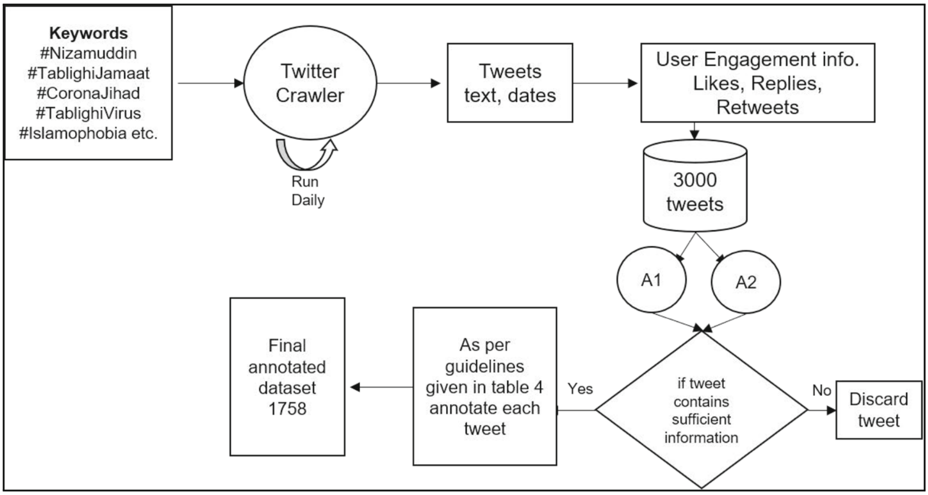

Keywords
#Nizamuddin
#TablighiJamaat
#CoronaJihad
#TablighiVirus
#Islamophobia etc.
Twitter Crawler
Run Daily
Tweets text, dates
User Engagement info. Likes, Replies, Retweets
3000 tweets
A1
A2
if tweet contains sufficient information
Yes
No
Discard tweet
As per guidelines given in table 4 annotate each tweet
Final annotated dataset 1758

**Fig. 2** Data collection flowchart

**Table 4** Annotators Guidelines

| |
|---|
| A URLs list of verified Indian news media like Times of India, India Today, CNN, The Hindu, Hindustan Times and a few Indian fact-checking websites, namely Altnews, Boom Live, WebQoof, and Factly were considered. Rating of each tweet has been done based on the following questions:<br>• Question the source;<br>• Question the motive such as: Why is this narrative amplified or why the similar content has been circulated on multiple platforms?<br>• Is it an Echo chamber: repeating false statements does not make it a fact illusory truth effect?<br>• Is this information inciting hate or doubt against any particular community?<br>• If the content is not factual but the purpose is to entertain? |

been developed for the complete analysis on disinformation during the Covid-19 Nizamuddin Tablighi Jamaat case. Corpus, "Fak_CoV" contains content (tweet text), temporal (date and time) and social (likes, retweets and replies) information. A total of 3000 tweets were extracted and annotated by two human annotators to obtain the ground truth concerning the presence of reliable tweets related to the event. Annotators have given an option of discard-tweet if the text under the tweet doesn't contain sufficient information. The definitions of terms disinformation and satire have been given in the introduction of this paper. Whereas, real news is the verified news from authentic sources. To assess news intention suitably, one count on the training labels often annotated by experts [60]. Most of the current studies have not described clear annotation guidelines to consider intention within a dataset. Table 4 lists the guidelines provided to annotators in order to consider authenticity as well as intention which in turn satisfies our C1 contribution. Figure 3 shows the complete architecture of the framework proposed to automate the process of flagging the data which has a high probability of being fake. Annotators analyzed each question with the help of URLs given and

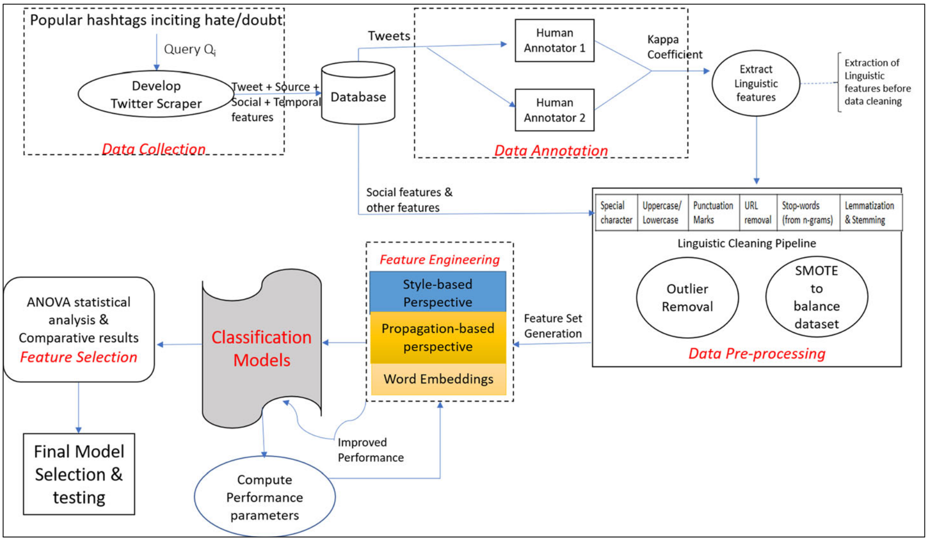


**Fig. 3** Generic framework for detection of fake news

rated each tweet as 0 (real),1 (disinformation) and 2 (satire) accordingly. During our pilot study, we observed options ‘disinformation’, ‘satire’ and ‘real’ were redundant since few tweets seemed mixed and unclear to the annotators. Hence, for the final annotation, we kept only one of the options. Inter annotator agreement (IAA) has been studied to measure how well the two annotators made the same decision. Therefore, Cohen’s kappa coefficient(K) [4] was used as a statistic to measure inter-annotator reliability using the formula:

$$K = (Po - Pe)/(1 - Pe)$$

Where Po is the observed agreement among the annotators and Pe is the hypothetical probability of the annotators indicating a chance agreement. The evaluation resulted in an overall Kappa score of 0.758 (a score greater than 0.7 implies a good agreement between annotators). However, annotation per tweet has been accepted using the majority score and the tweets for which annotators gave different ratings was skipped. After removing tweets with different ratings and insufficient information, in total, we obtained 1758 tweets in the final annotated corpus.

### 3.2 Data preprocessing

To acquire insights from the dataset, it is a general practice to do preliminary and exploratory data analysis.

Imbalanced dataset: It is a vital step related to classification models to check whether a dataset is balanced or not. For example, if the majority of data belongs to only one particular class then the model will classify each record into the class which has the majority and obtain a decent accuracy. In this case, the model did not even consider other classes. But there do exist approaches to deal with the problem of imbalanced dataset like under-sample the majority class, oversample the minority class, using error metrics recall, precision etc. Figure 4 shows the percentage and the length distribution of articles across the categories. It is not a perfectly balanced dataset. Even though the length is evenly distributed and the text contains more than 200 characters. Thus, there exists enough information in most of the tweets to perform the classification. Furthermore, to handle the imbalanced dataset, oversampling techniques are considered when the dataset is not enough. Therefore, SMOTE (Synthetic Minority Over-sampling Technique) [11] has been applied to create synthetic samples from the minor class (satire and disinformation) instead of creating copies.

Features extracted before cleaning: A variety of information has been extracted from the text which includes the numerical count value of style-based features described in Table 2. Dictionaries for modal verbs, tentative, generalizing verbs, and certainty words have been

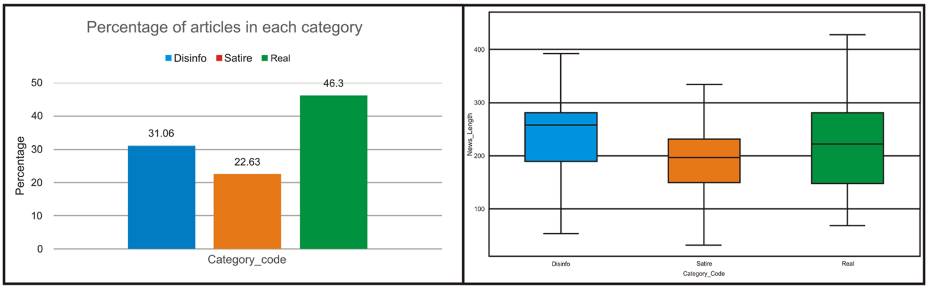


**Fig. 4** Percentage and length distribution of data to each category

created and the count has been checked for these values for every tweet. The extracted information was saved into a different file prior to preprocessing pipeline.

Linguistic Cleaning: A pipeline of cleaning steps described in framework architecture has been applied to the Fak_Cov corpus including *removal of special characters, punctuation marks, URLs, uppercase/lowercase letters, stopwords and lemmatization*.

Features extracted after preprocessing: Thereafter, sentiment analysis has been performed using TextBlob and NRC lexicons. With TextBlob, we obtained polarity and subjectivity. While scores per emotion have been obtained using the NRC lexicon.

Analyse data distribution: Furthermore, differences between disinformation, real and satire news have been calculated using some of the numerical and categorical features out of ten feature sets and have been represented with the help of boxplots. In Fig. 5, the x-axis represents the classes (real:0, disinfo:1, satire:2), while y-axis shows the value range of each feature. With the help of box plots, it has been observed that most of the features, namely char count, hashtag count, mention count, unique word count, word count, polarity, subjectivity and stop word count are highest in fake content. The general motive behind fake news is to reach out to maximum people, so in order to make it attractive and hide deception, features like hashtag, mention, stop word, extreme sentiments are used.

Outliers Removal: Moreover, the box plots interpretation showed that a lot of the observable data is highly skewed since few outliers have extremely high or low values. Therefore, logarithm transformation (automatic statistical method) has been used for the treatment of skewed variables [17]. However, in a few cases, there are valid reasons for these observations to be outliers. Hence, we experimented with different classification models using both data with outliers and normally distributed data. Table 5 shows the skewed values of features in raw data and after outlier removal. It has been observed that few features were highly positively skewed and after logarithmic transformation value became near to zero i.e. normally distributed. This happens generally because the exact zero value for explaining continuous data in statistics rarely exists in nature. The features whose skewed value were already near to zero have not been transformed.

### 3.3 Feature engineering

Before discussing the computation results, the vectorization technique, handcrafted features and embedding technique used are given below:

- TF-IDF Vectorization with N-gram features: TF-IDF, term frequency-inverse document frequency is used to calculate relative frequency and works well at various language levels. While n-gram models capture the sequence of words. In the paper, word-level n-gram has been used to represent the context of the sentence and generate features using TF-IDF for classification [55]. It is an efficient and popular technique in text categorization.
- Style-based Features: The style-based features as described in the above sections have rich literature. Lexical, syntactic and psycholinguistic features are combinedly called linguistic features. It includes character-level (e.g. Char count, mean word length), word level (e.g. Total word count, unique words), and sentence-level features (punctuation count). Rosas et al. performed exploratory analysis on the identification of linguistic differences in fake and legitimate news content [41]. Rashkin et al. used Linguistic Inquiry and Word Count (LIWC), a lexicon widely used in social science studies along with other linguistic features for analyzing language in fake news [44]. Moreover, some fake news detection models are based on sentiment score [46]. TextBlob's API

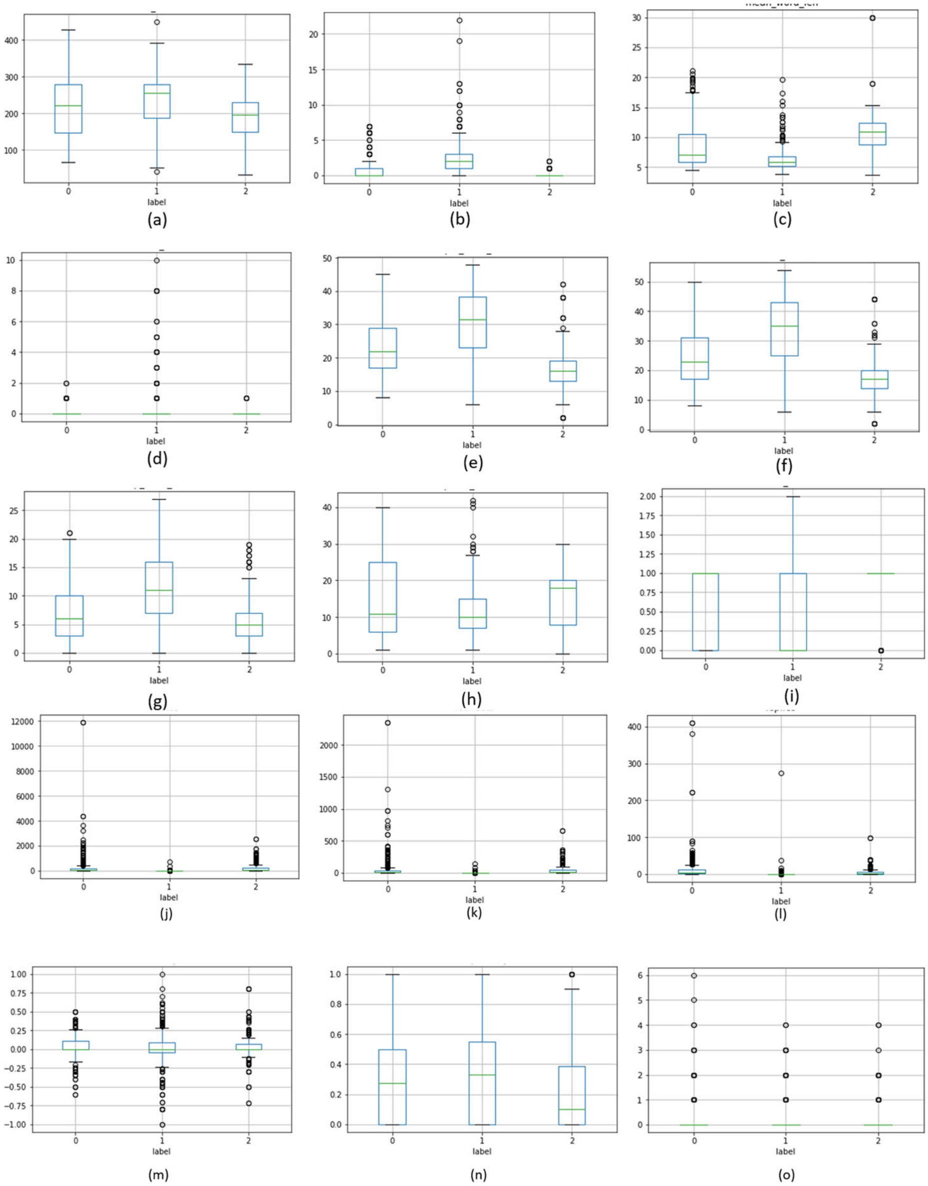


**Fig. 5** Data distribution boxplot of linguistic features *(a)* Char_count *(b)* Hashtag_count *(c)* Mean_word_len *(d)* Mention_count *(e)* Unique_word_count *(f)* Word_count *(g)* Stop_word_count *(h)* Punct_count *(i)* Url_count *(j)* Likes *(k)* Retweets *(l)* Replies *(m)* Polarity *(n)* Subjectivity *(o)* Modal_verbs

(http://textblob.readthedocs.io/en/dev/) has been used to compute the sentiment score of a tweet. Disinformation articles are emotionally charged with provocative language. Ajao et al. proposed a hypothesis that there exists a relation between fake messages or rumours and sentiments of the texts posted online [3]. The style-based features used in this pilot-study are described in Table 2.

**Table 5** Skewed values of features

| Features | Skewed value | after outlier removal | Features | Skewed value | after outlier removal |
|---|---|---|---|---|---|
| Likes | 12.43 | -0.15 | Tentative | 2.81 | 0 |
| Retweets | 10.39 | 0.48 | Generalizing | 4.84 | 1.32 |
| Replies | 10.97 | 1.07 | Certainty | 2.98 | 0.60 |
| URL_count | -0.44 | - | Noun phrases | 2.69 | 0.81 |
| Hashtag_count | 3.76 | 1.92 | Sentences | -0.34 | - |
| Mention_count | 6.64 | 6.89 | Avg_character_count | -0.09 | - |
| Avg_Punct_count | 0.62 | - | Avg_words_count | 0.56 | - |
| Stopword_Count | 0.77 | - | Numbers+quantifiers | 3.76 | 0.24 |
| Mean_word_length | 1.60 | 0.57 | Question_mark_count | 4.89 | 1.03 |
| Word_count | 0.44 | - | Positive_count | 7.31 | 2.43 |
| Unique_word_count | 0.37 | - | Negative_count | 4.53 | 3.96 |
| Char_count | -0.06 | - | Anxiety/angry/sad_Count | 5.72 | 2.43 |
| Polarity | -0.53 | - | Exclamation_count | 6.34 | 1.04 |
| Subjectivity | 0.61 | - | Unique_nouns/verbs/adverbs | 2.54 | 1.77 |
| Modal verbs | 3.11 | 1.02 | Spellcheck_count | 5.63 | 2.48 |
| Swear_word_count | 4.54 | 3.61 | | | |

- Social Features: Since content style-based features are not sufficient for detecting fake information, hence, social features are required to be incorporated. Therefore, while extracting data using Twitter API, a lot of additional features have also been extracted related to the popularity of the tweet. Literature shows various claims regarding the difference between legitimate and false news in terms of social and propagation parameters. Vosoughi et al. also claimed that false news spreads faster than real news online [56]. Therefore, social features like the number of likes, retweets, and replies have been explored in order to distinguish real, disinformation and satirical tweets.
- Word Embeddings: The text embeddings are Latent textual features that can be used directly as input to classification models for prediction. TF-IDF only considers the content, while word embeddings consider the content as well as context. Word embeddings are a kind of word representation that provide similar words with similar meaning, a similar representation [22]. Word2vec is not a single algorithm but works in the combination of two methods – CBOW (Continuous bag of words) and Skip-gram model. Both the methods are based on shallow neural networks which map a word to a target variable or word. Moreover, learned weights by both the techniques are termed as vector representations. Word embedding protects the word relationships in such a way that words with similar context will have similar vector representation or embeddings. Thus, the words occur close to each other in the vector space which in turn helps classification algorithms to work effectively. In this paper, Glove (Global vectors for word representation) has been used as a word embedding technique [18]. It is an unsupervised learning technique to generate vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus and the resulting representations showcase interesting linear substructures of the word vector space. This dataset contains English word vectors pre-trained on the combined Wikipedia 2014 + Gigaword 5th Edition corpora (6B tokens, 400K vocab). This dataset contains 50, 40 100, 200 and 300-dimensional pre-trained word vectors. We have used 100-dimensional word embeddings for English in our experiments. Literature shows rich research in text classification using GloVe embeddings because pre-trained embedding is an efficient way to quantify word co-occurrence (which also captures some aspects of word meaning). Figure 6 shows word embeddings in the Fak_Cov dataset using the TSNE dimensionality reduction visualization technique.

### 3.4 Proposed framework

Our proposed fake news detection framework has three major components: integrated feature generation; supervised classifiers and Analysis of variance test for selecting important features. The proposed generic framework for fake news detection is explained using a flow chart (Fig. 7). Following are the steps of the methodology:

- Developed a Twitter scraper to extract tweets and related information about the input query.Extracted information stored in a database and annotated by two human annotators considering intention and authenticity.
- Style-based and propagation-based features have been extracted to generate integrated feature set for classification.
- The ten extracted feature sets with a total of thirty-one features have been added iteratively to the classification models in order to improve the performance at each level. The performance has been monitored at each iteration.

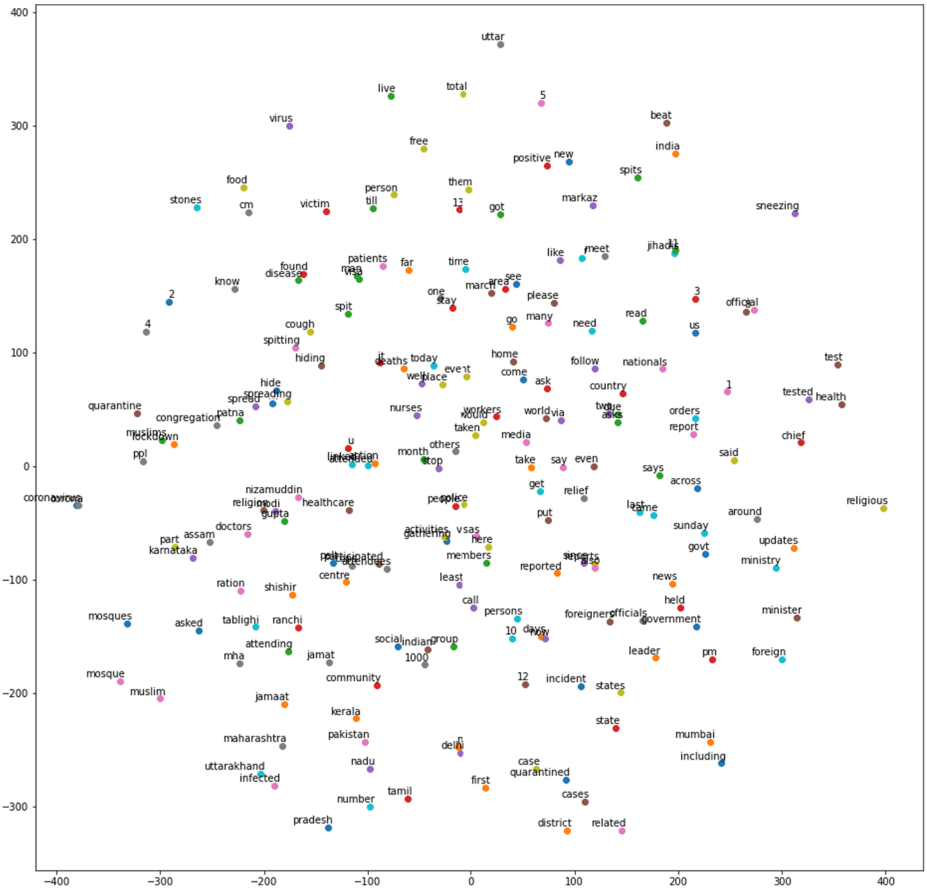


**Fig. 6** Word embeddings using TSNE

– Finally, the ANOVA statistical test has been applied to select significant features for differentiating the three categories: disinformation, satire and real.

# 4 Experiment results

## 4.1 Setup

The experiments were run using 5-cross-validation and in each validation round 20% for testing data and 80% for training data have been tested on several widely-accepted supervised learning classifiers [40]. The performance of the experiments has been evaluated in terms of accuracy, precision, recall and f1 score. The learning algorithms were used to train the model and then to predict the labels for the test dataset. Hyperparameter tuning is an important step in machine learning-based classification [9]. It is often carried out by hand, gradually purifying a grid over the hyperparameter space. In our analysis, different hyperparameters were tuned to gauge model performances with the goal to achieve bias-reduced

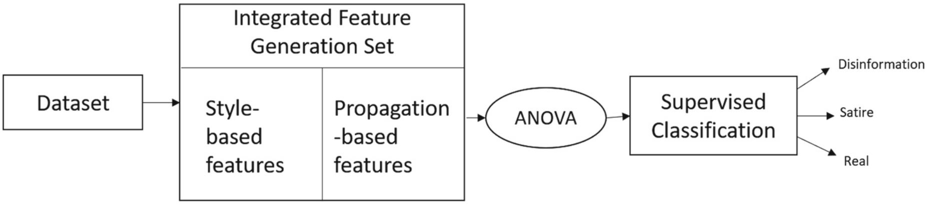

Dataset
Integrated Feature Generation Set
Style-based features
Propagation -based features
ANOVA
Supervised Classification
Disinformation
Satire
Real

**Fig. 7** Flow chart of the proposed generic framework

performance evaluations. After initial testing, the hyperparameters were set to a specific set of values for all the experiments as shown in Table 6. The performance of classifiers has been measured using accuracy, F-1 score and kappa score. We purposely used the F1 score as this score shows the harmonic mean between recall and precision. Also, using such a score facilitates the comparison with the existing studies. Whereas, kappa score is a measure to compare the observed frequency with expected accuracy and thus, is less misleading. Figure 8 shows the ensemble model architecture used in the analysis. In Ensemble, the training data is divided into various subsets and then trained on different classifiers at level 0. The output of classifiers is used as training data for level 1 classifier i.e., Logistic regression. The goal of using Ensemble is to improve the accuracy of weak classifiers by combining them. Therefore, in each experiment, models have been chosen according to the performance of classifiers in a particular scenario. In most of the experiments, it has been observed that MNB, KNN and DT are obtaining less accuracy so they are chosen as base models in Ensemble. Different experiments performed to select a feature set and model which can best capture the deceptive tweets are explained in the following subsequent sections.

## 4.2 Computational results using N-grams with TF-IDF

In the first experiment, feature extraction method TF-IDF and varied size of n-grams with n=1 (unigrams), 2 (bigrams) up to n=4 have been studied. Figure 9 shows different n-grams

**Table 6** Selected values of hyperparameters majorly effecting the overall performance

| Classifier | Hyperparameters effecting the performance of classifier in Fak_cov dataset |
|---|---|
| Support Vector Machine (SVM) | Kernel= Linear & RBF; Decision function shape=ovo; penalty term C =1; degree=2 |
| Random Forest (RF) | Number of estimators =3 |
| eXtreme Gradient Boosting (XGB) | Learning rate= 0.5; n=3; depth=2; minchild=1; gamma=5; objective =multi-softmax |
| Decision Tree (DT) | Depth= 2 |
| K-nearest neighbor (KNN) | Number of neighbours, k=5 |
| Multi-layer Perceptron (MLP) | Activation function= tanh, Relu, logistic; solver=adam; hidden layer= (25,20,10); max iterations=500 (Majorly tanh performed best) |

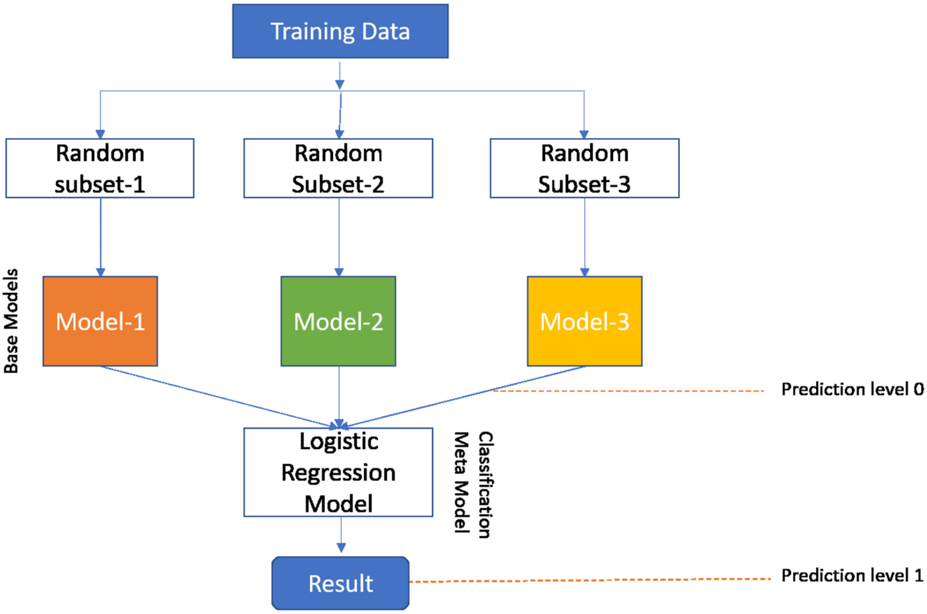


**Fig. 8** Ensemble Model Architecture

obtained for Fak_Cov corpus. For example, the unigrams display the sequence of single words with a diverse vocabulary. The performance of the aforementioned machine learning classifiers has been checked on the corpus to predict whether the tweet is real, satire or disinformation. Table 9 shows the accuracy obtained from different classifiers using uni, bi, tri and quad-grams. It has been observed that classifiers are performing well with unigrams and performance is decreasing as we increased the value of n. Moreover, Chi-square test

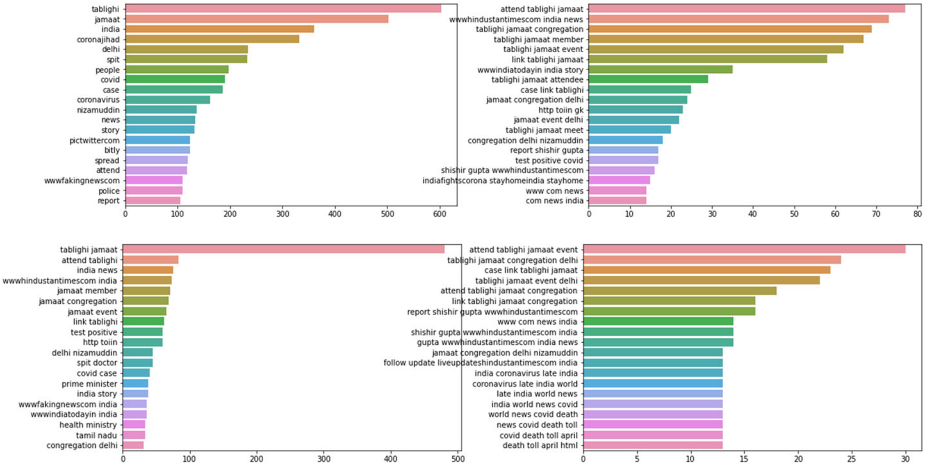


**Fig. 9** Generating n-grams: unigrams *(top-left)*, bigram *(bottom-left)*, trigrams *(top-right)*, and quad-grams*(bottom-right)*

[42] has been used in order to test what unigrams (n=1), bigrams (n=2), trigrams (n=3) and quad gram (n=4) are most correlated to each other. We observed that unigrams performed best in terms of correlation with the category. It is a thumb rule that training size should be ten times of features to avoid dimensionality problem. The experiment was run three times by considering different values of feature size (FS) as 30, 50, and 100. From the results obtained in our experiments, the classifiers performed well with values chosen to train models with high predictive power. However, there is a minute difference between values but FS=30 shows the best results. Since we have restricted to a smaller number of features due to the small dataset in all the grams experiments.

### 4.3 Computational results using Iterative Feature Selection

Apart from text data (which is now TF-IDF sparse matrix), the other nine feature sets (containing 31 features in total) under stylistic and social categories as described in Table 2 have been explored in order to further improve the overall performance of classification models. Since, the handcrafted features are divided into nine categories from F(2-10): quantity, complexity, uncertainty, sentiment, subjectivity, diversity, informality, additional, and popularity. Thus, the feature sets F(2-10) have been run sequentially by adding the next feature set in each iteration. Finally, we checked the performance of models over all the numeric features iteratively. Note, SVM1 represents the SVM model with RBF kernel, while SVM2 represents Linear kernel. To this end, the TF-IDF frequency matrix (F1) is also combined with F(2-10). Data frame mapper functionality in python has been used to combine transformed text with numeric features. Furthermore, to check the importance of features, ANOVA (Analysis of Variance) test has been done on each feature while defining the problem statement [24]. Other tests like Chi-square test is not suitable in this scenario since it works better with two categorical variables. Moreover, a t-test is also not suitable as there are more than two categories (labels). Essentially, the collected data is about one categorical independent variable i.e., ‘label’ (disinformation, real and satire) and a set of quantitative dependent variables i.e., features as mentioned in Table 7. ANOVA is used to find the correlation between the categorical variable and other variables [54]. In the experiments, ANOVA has been applied separately on each feature assuming that there is no interaction between classes. Finally, the p-value is calculated for each dependent variable (one feature at a time). For a particular feature, if the test gives a p-value less than 0.05 then there exists a strong correlation between the two variables [15, 32]. Hence, the null hypothesis, Ho (as stated in the section problem statement) is rejected and the alternative hypothesis, Ha is accepted. This implies that label type does have a significant effect on the respective feature. The p-values obtained are shown in Table 8. For example, the p-value obtained for feature ’#characters’ is 0.000004 which is very less than 0.05; thus, the feature is important for classification. The selected features are retweets, replies, word_count, unique_word_count, character_count, positive_count, negative_count, emotion_count, subjectivity and mean_word_length since the obtained p-value is less than or equal to 0.05 (lower the better) for these features in the Fak_Cov dataset.

**Table 7** Variables in ANOVA test

| Variable | Name |
|---|---|
| Dependent variables | 31 Features considering one at a time |
| Independent variable | labels = Disinformation, real and satire |

**Table 8** Feature Importance using ANOVA test

| Feature Set | Features | ANOVA (p-value) | |
|---|---|---|---|
| F2 | # Characters | 0.000004 | ✓ |
| | # Words | 0.000001 | ✓ |
| | # Noun Phrases | 0.34 | |
| | # Sentences | 0.89 | |
| F3 | Average # characters per word | 0.3 | |
| | Average # words per sentence | 0.09 | |
| | Average # punctuations per sentence | 0.845 | |
| F4 | # Modal verbs | 0.79 | |
| | # Certainty terms | 0.58 | |
| | # Generalizing terms | 0.91 | |
| | # Tentative terms | 0.12 | |
| | # Numbers and quantifiers | 0.56 | |
| | # Question marks | 0.89 | |
| F5 | # Positive words | 0.05 | ✓ |
| | # Negative words | 0.04 | ✓ |
| | # Anxiety/angry/sadness words (emotion) | 0.05 | ✓ |
| | # Exclamation marks | 0.9 | |
| | Content sentiment polarity | 0.2 | |
| F6 | # Subjective verbs | 0.03 | ✓ |
| F7 | # Unique words | 2.45 * e-07 | ✓ |
| | # Unique nouns, verbs, adjectives, adverbs | 0.8 | |
| F8 | # Typos/spellchecks | 0.73 | |
| | # Swear words/netspeak/assent/fillers | 0.68 | |
| F9 | # Hashtags | 0.221 | |
| | # Mentions | 0.662 | |
| | # Stopwords | 0.51 | |
| | # URL | 0.32 | |
| | Mean word length | 0.000052 | ✓ |
| F10 | # Likes | 0.11 | |
| | # Retweets | 0.054 | ✓ |
| | # Replies | 0.000019 | ✓ |

Tables 9, 10, 11 and 12 show the obtained results with respective feature sets. Results show that

- Non-latent features outperformed latent ones: Latent textual features (text embeddings) have been experimented which gives a vector for each tweet content. It has been observed that non-latent features (n-grams with TF-IDF) have achieved maximum accuracy of 98.24% which is far better than the accuracy achieved from latent ones i.e., 86.76%.
- Combined features of multiple perspectives outperformed single-level features: Combining features across different sets enhances the performance compared to when separately using each feature set. Results indicate that introducing F9 i.e., additional

**Table 9** Using TF-IDF

| Classification Model | Accuracy | | | |
|---|---|---|---|---|
| | N=1 Unigram | N=2 Bigram | N=3 Trigram | N=4 Quadgram |
| MNB | 94.62 | 72.73 | 63.26 | 62.5 |
| LR | 97.11 | 92.8 | 66.67 | 62.5 |
| SVM1, SVM2 | **98.24**, 98.11 | 92.05, 92.8 | 90.15, 91.29 | 62.88, 88.26 |
| RF | 97.62 | 93.56 | 62.88 | 50.76 |
| XGB | 97.62 | 74.62 | 44.7 | 44.7 |
| DT | 75.76 | 67.05 | 44.7 | 44.7 |
| KNN | 87.88 | 84.85 | 84.47 | 25.0 |
| Ensemble | 98.04 | 96.21 | 78.79 | 61.36 |
| MLP | 98.08 | 81.82 | 62.5 | 51.52 |

feature set (such as the number of hashtags, mentions, stopwords, URLs) have increased accuracy and F-1 score of models strongly from 0.74 to 0.90 with Ensemble classifier. Finally, the performance of models has improved when features from a different perspective are added (i.e., user engagement features).

- Relative frequency vector matrix of text performed better than other feature groups: It has been observed that the performance obtained using frequency feature is better than other feature sets in terms of accuracy and F-1 score. Thus, the relative frequency vector features played a better role in differentiating disinformation from satire and real ones.
- Random forest performed acceptably well with a smaller number of features: Random Forest (RF) has shown acceptably well performance throughout the experiments even with a small number of features.
- Whereas, as the number of features is increasing, the proposed ensemble classifier outperformed other classifiers. Moreover, an ANOVA test was conducted to extract important features which further improved the performance.

The classification models RF, Ensemble, and MLP are clearly best suited for the predictive models. Notably, sentiment feature set (F5) has negative values; hence, MNB could not perform in further experiments. Furthermore, additional experiments have been conducted using CNN, BERT and RNN-LSTM on the corpus and accuracies obtained were 40.86%, 56.25% and 46.77% respectively. Evidently, machine learning classifiers outperformed deep learning models on the Fak_Cov corpus. Overall, it has been observed that Ensemble and MLP found more suitable for this problem. Hence, the models are useful in automating the process of flagging data that has a high probability of being fake.

## 5 Benchmark observations and visualization

- Ahmed et al. used machine learning with n-grams and achieved the highest accuracy of 92% using unigrams with Linear SVM classifier [1]. Whereas, our model achieved an accuracy of 98.24% on multiclass Fak_Cov corpus with unigrams TF-IDF using Linear SVM classifier.
- Style-based features have been mainly trained on supervised machine learning classifiers. Rosas et al. relied on SVM to achieve an accuracy of 74% using features like

**Table 10** Using iterative feature engineering

| | F2 | | F2-3 | | F2-4 | | F2-5 | | F2-6 | | F2-7 | | F2-8 | | F2-9 | | F2-10 | | Selected | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| MNB | 0.51 | 0.50 | 0.61 | 0.60 | 0.50 | 0.46 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| LR | 0.64 | 0.62 | 0.49 | 0.45 | 0.60 | 0.59 | 0.58 | 0.57 | 0.57 | 0.56 | 0.57 | 0.56 | 0.58 | 0.57 | 0.73 | 0.73 | 0.76 | 0.75 | 0.82 | 0.81 |
| SVM1 | 0.61 | 0.54 | 0.68 | 0.68 | 0.67 | 0.67 | 0.66 | 0.67 | 0.66 | 0.66 | 0.66 | 0.66 | 0.63 | 0.63 | 0.62 | 0.55 | 0.71 | 0.65 | 0.72 | 0.65 |
| SVM2 | 0.70 | 0.65 | 0.64 | 0.62 | 0.62 | 0.60 | 0.58 | 0.57 | 0.58 | 0.57 | 0.59 | 0.52 | 0.58 | 0.59 | 0.77 | 0.77 | 0.80 | 0.79 | 0.84 | 0.85 |
| RF | **0.76** | **0.76** | **0.73** | **0.73** | 0.63 | 0.63 | **0.74** | **0.74** | 0.69 | 0.69 | 0.69 | 0.69 | 0.69 | 0.69 | 0.84 | 0.84 | 0.86 | 0.87 | **0.94** | **0.94** |
| XGB | 0.60 | 0.58 | 0.72 | 0.72 | 0.71 | 0.71 | 0.67 | 0.65 | 0.63 | 0.60 | 0.52 | 0.43 | 0.53 | 0.53 | 0.72 | 0.72 | 0.77 | 0.73 | 0.80 | 0.79 |
| DT | 0.59 | 0.52 | 0.59 | 0.52 | 0.59 | 0.52 | 0.59 | 0.30 | 0.59 | 0.52 | 0.59 | 0.52 | 0.61 | 0.60 | 0.75 | 0.74 | 0.71 | 0.72 | 0.70 | 0.71 |
| KNN | 0.72 | 0.71 | 0.67 | 0.62 | 0.67 | 0.67 | 0.68 | 0.68 | 0.67 | 0.67 | 0.71 | 0.71 | 0.73 | 0.72 | 0.72 | 0.71 | 0.78 | 0.79 | 0.81 | 0.81 |
| Ensemble | 0.73 | 0.73 | 0.66 | 0.66 | 0.68 | 0.68 | 0.73 | 0.73 | **0.73** | **0.73** | **0.76** | **0.76** | **0.74** | **0.74** | **0.90** | **0.90** | **0.91** | **0.91** | **0.94** | **0.94** |
| MLP | 0.67 | 0.64 | 0.70 | 0.70 | **0.72** | **0.72** | 0.71 | 0.71 | 0.69 | 0.69 | 0.70 | 0.70 | 0.72 | 0.71 | 0.81 | 0.81 | 0.88 | 0.88 | 0.92 | 0.92 |

**Table 11** Using TF-IDF with Selected features (Data Mapper)

| Classification Model | Accuracy | F-1 measure | Kappa |
|---|---|---|---|
| MNB | - | - | - |
| LR | 0.81 | 0.81 | 0.70 |
| SVM1, SVM2 | 0.72, 0.85 | 0.65, 0.86 | 0.54, 0.81 |
| RF | 0.96 | 0.96 | 0.94 |
| XGB | 0.91 | 0.91 | 0.86 |
| DT | 0.76 | 0.68 | 0.60 |
| KNN | 0.81 | 0.81 | 0.70 |
| Ensemble | **0.99** | **0.99** | **0.98** |
| MLP | 0.98 | 0.98 | 0.97 |

n-grams, LIWC, readability, punctuations combinedly [41]. We followed their methodology on our dataset but these features have not performed well. Nonetheless, our model achieved 90% accuracy with the Ensemble model considering some more style-based features such as quality, complexity, uncertainty, sentiment, subjectivity, etc.

- Shu et al. proposed a tri-relationship fake news detection framework and achieved an accuracy of 89.3% on PolitiFact and Buzzfeed dataset with publisher bias, news stance and user engagement features [50]. Their methodology has shown similar results on our dataset. Whereas, the proposed model used an integrated approach of combined ten feature sets of a total of 31 features achieved the highest accuracy of 99% with Ensemble.
- Moreover, Reis et al. explored the importance of features for automatic detection of false news using chi-square test and achieved the best accuracy with XGB of 86% [45]. The authors explored style based on user engagement features also. we employed their methodology on our dataset but better results were achieved with our proposed model i.e., the accuracy of 94% using Ensemble and RF. When selected style-based and social features have been used along with n-grams with TF-IDF, accuracies of 99% and 98% have been obtained using ensemble and MLP respectively.
- Zhou et al. explored various non-latent style-based features combinedly and obtained 84.5% accuracy [61]. Also, the authors explored Latent features, namely word2vec and Doc2vec with obtained accuracies of 68.8% and 69.8% respectively. Whereas, the

**Table 12** Using word embeddings

| Classification Model | Accuracy | F-1 measure | Kappa |
|---|---|---|---|
| MNB | 52.94 | 0.52 | 0.30 |
| LR | 55.04 | 0.54 | 0.32 |
| SVM1, SVM2 | 68.49 | 0.68 | 0.49 |
| RF | 82.56 | 0.82 | 0.72 |
| XGB | 82.77 | 0.82 | 0.72 |
| DT | 77.73 | 0.77 | 0.64 |
| KNN | 61.34 | 0.59 | 0.36 |
| Ensemble, Ensemble1 | **86.67, 86.76** | **0.80**, **0.86** | **0.69**, **0.78** |
| MLP (tanh), Relu, logistic | 67.86, 68.07, 69.54 | 0.67, 0.67, 0.69 | 0.78, 0.78,0.78 |

proposed model in this paper achieved 86% accuracy with Glove word embeddings over Ensemble classifier. Although, our model also shows better results with non-latent features as compared to latent ones.

- Khan et al. obtained 95% accuracy with Glove embedding using Bi-LSTM on a combined corpus of LIAR and Fake or Real datasets [31]. Nonetheless, our dataset achieved very less accuracy with Glove using LSTM. The reason is the unexplored domain and limited dataset. However, the accuracy of 86.67% has been obtained using the proposed Ensemble model and 88.76% using the Ensemble model by changing the classifiers to further boost the accuracy.

In this paper, we performed dimensionality reduction techniques, namely PCA (principal component analysis) and t-SNE (t-distributed Stochastic Neighbor Embedding) to plot the observations in two dimensions [14]. Figure 10 shows the visualization of fake, real and satire vectors in two-dimensional space using PCA and t-SNE techniques. Apparently, Satire and disinformation are kinds of deceptive content that differ in intention. Therefore, it is clear in the graphs that Satire and disinformation are not perfectly classified, while Real news is perfectly segregated from the other two categories. Overall, all features combinedly improve the accuracy of the framework and models, namely XGB, Ensemble and MLP outperformed. Moreover, a slight increase in the accuracy of classifiers is obtained after feature selection using the ANOVA test. The developed corpus that has been analyzed in this paper is multi-class (disinfo, satire and real) and focused on the current sensitive issue during the critical times of Covid-19. Therefore, a number of experiments have been implemented in order to find a suitable one in such a scenario. Essentially, the F-1 score has also been computed since it shows the harmonic mean between recall and precision. Also, Cohen's Kappa score used as a multi-class metric to measure the degree of agreement between the actual values and the predicted values. In our experiments, classifiers have obtained a good kappa score (nearly equals 1 which shows strong agreement).

Furthermore, the Ensemble model obtains the highest accuracy with selected features (Table 10). In order to gain insights into the way the model is working, we did the model interpretation of the Ensemble classifier. It has been observed from the confusion matrix (Table 13) of the Ensemble model that total 10 data records were misclassified and out of which 8 satirical news have been misclassified as disinformation (as shown in Fig. 11). As

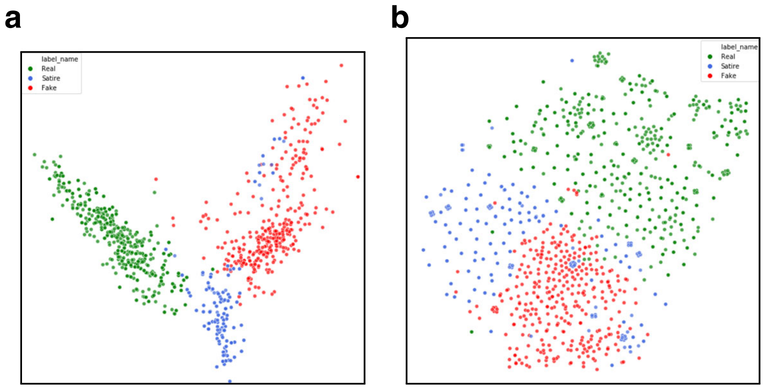


**Fig. 10** Visualization of tweet articles from FaK_CoV Corpus (a) using PCA and (b) using t-SNE

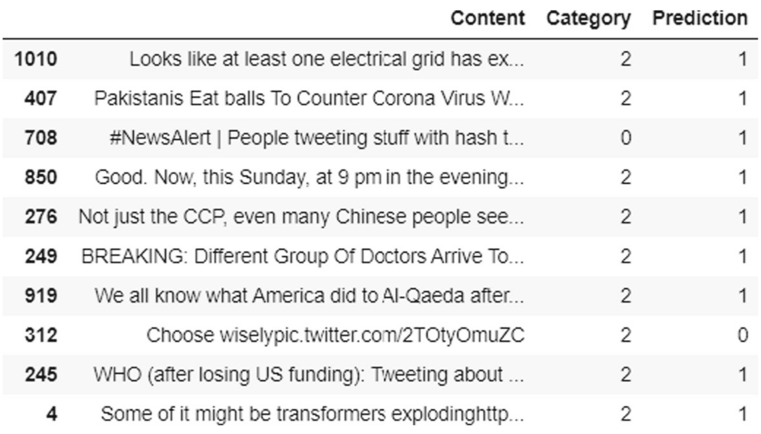
**Table 13** Confusion Matrix

| Pred/True | Real | Disinfo | Satire | Overall | % Accuracy |
|---|---|---|---|---|---|
| Real | 119 | 1 | 0 | 120 | 99.16 |
| Disinfo | 0 | 85 | 0 | 85 | 100 |
| Satire | 1 | 8 | 50 | 59 | 84.75 |
| Overall | 120 | 94 | 50 | 264 | 94.62 |

per the literature, satirical content is considerably similar to fake content. Hence, all satirical news have been manually analyzed to know the reason why these were misclassified. Overall, each misclassified satire article consists of text with some image but in this work, only textual features have been explored. For example, the article “Pakistanis eat balls to counter Corona Virus which originated from Bats” looks fake but the image as shown in Fig. 12 depicts the fun part behind the news. These types of errors always occur since there can be articles that truly belong to two or more categories at the same time.

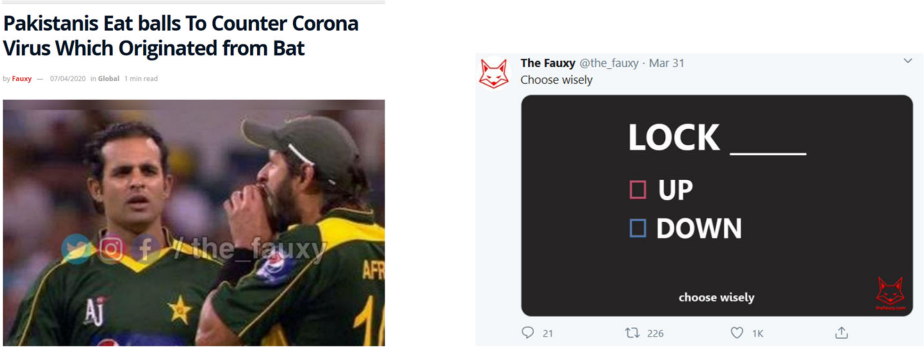
## 5.1 Comprehensive Model Test

The Model has been trained for the Fak_Cov corpus which is collected from the time event happened i.e. 29th March 2020 to 14th April 2020 using different related keywords and finally labelled as Real, fake and Satire. In addition to the training dataset, there may appear other new articles on the web. The classification model provides the conditional probability of an article being fake. Therefore, we have assigned a score from range 1 to 5 based on the obtained probability. A score ‘1’ depicts a very low possibility of being fake, while score ‘5’ means a high possibility of the article being fake. Nonetheless, there may be articles which are not related to such sensitive issue like communal riots in a country thus always be misclassified. For instance, in an article that clearly related to such a sensitive national issue, the conditional probability of belonging to either fake or real class is high accordingly. Essentially, three out of a few recent articles randomly scraped from Twitter are shown in Table 14



| | Content | Category | Prediction |
|---|---|---|---|
| 1010 | Looks like at least one electrical grid has ex... | 2 | 1 |
| 407 | Pakistanis Eat balls To Counter Corona Virus W... | 2 | 1 |
| 708 | #NewsAlert \| People tweeting stuff with hash t... | 0 | 1 |
| 850 | Good. Now, this Sunday, at 9 pm in the evening... | 2 | 1 |
| 276 | Not just the CCP, even many Chinese people see... | 2 | 1 |
| 249 | BREAKING: Different Group Of Doctors Arrive To... | 2 | 1 |
| 919 | We all know what America did to Al-Qaeda after... | 2 | 1 |
| 312 | Choose wiselypic.twitter.com/2TOtyOmuZC | 2 | 0 |
| 245 | WHO (after losing US funding): Tweeting about ... | 2 | 1 |
| 4 | Some of it might be transformers explodinghttp... | 2 | 1 |

**Fig. 11** Ten misclassified articles by Ensemble classifier



Pakistanis Eat balls To Counter Corona Virus Which Originated from Bat

by Fauxy — 07/04/2020 in Global 1 min read

The Fauxy @the_fauxy · Mar 31

Choose wisely

LOCK ____

□ UP

□ DOWN

choose wisely

**Fig. 12** Misclassified Satirical News extracted manually from Twitter

and observed to obtain the behaviour of conditional probability vectors. Firstly, articles clearly related to this event have been tested using an Ensemble classifier and then articles that are not related to this particular event. The model has predicted the first article from related articles (as shown in Table 14) as disinformation with high conditional probability, thus, a high score of 5 has been assigned. Whereas, the article predicted as disinformation with less conditional probability is assigned a score of 3. Clearly, this multilabel approach of scoring articles satisfies our contribution C5. The model doesn't perform well with data that is not related to such an event. This evidently indicates that the research can be used as a fundamental study for the future in this domain.

**Table 14** Random tweets tested on Ensemble using selected features

| | Text | Actual category | Predicted category | Conditional Probability | Score |
|---|---|---|---|---|---|
| Related articles | - No my friend, muslims are doing something called coronajihad, they are spitting on fruit and vegetables. | Disinfo | Disinfo | 99.70 | 5 |
| | - Shameful: How Tablighi Jamaat workers manhandled a lady health worker in Delhi LNJP hospital. | Real | Real | 97.85 | 1 |
| | - Vijay Mallya Plans to fill all his aircrafts with crude oil, will earn enough money to clear all his loans | Satire | Disinfo | 52.79 | 3 |
| Unrelated article | - I have instructed the United states Navy to shoot down and destroy any and all Iranian gunboats if they harass our ships at sea. | Real | Real | 58.74 | 3 |

# 6 Discussion

## 6.1 Case study: Tablighi Jamaat narrative

Tablighi Jamaat is a global religious organization to encourage a specific community and fellow members to practice their religion. Tablighi Jamaat, the religious congregation was held in mid-march at Delhi's Nizamuddin mosque significantly contributed to increase the coronavirus cases in India. On 24th March 2020, a nationwide lockdown was observed at the instance of prime minister Narendra Modi and it was strictly prohibited to enter the premises of the mosque during Lockdown. However, on 30th March, the Tablighi Jamaat congregation at Delhi's Nizamuddin mosque was found and more than 100 positive cases of coronavirus have arisen. But it was not enough. When the whole world was suffering from Covid-19, India was suffering from two different communalism narratives by social media platforms [19]. The first narrative presented the event as an irresponsible act by a group of people. Whereas, the second narrative stated it as propaganda of a specific religious community with the intent to spread coronavirus disease in India. Social Media accounts vending hate and fake news is behind the whole shift in the narrative. It has been observed that western media also communalized the event to carry out cyber warfare. The hashtags in Table 15 were being used to peddle the communal narrative. Moreover, Fig. 13 shows the popularity of hashtags over time used during the Tablighi Jamaat event. The hashtags used by social media users were clearly indicating that it was not a mistake rather a hidden agenda well-planned event happened purposely to increase coronavirus cases in India. Moreover, few tweets are shown in Fig. 14 which propagated disinformation and accounts like Rosy @rosyk01 has been suspended by Twitter to violate Twitter Rules. Twitter has removed offensive content linking Covid-19 to communalism under offensive hashtags #coronaJihad, #islamophobia after a PIL filed in India. The actions taken by Twitter to control misinformation are virtuous because it is widely used as the source of information.

## 6.2 Limitation

In general, the performance of models is not dataset invariant and it is quite hard to find an exclusive model for all datasets. Hence, one of the main challenges of this study is the limited data size. We do not claim that our dataset is demonstrative of the whole Twitter population. However, the strength of this analysis deceits in the proposed framework that has been adopted from fundamental theories to gauge a better understanding of human behaviour patterns. Furthermore, deep learning models such as CNN, RNN-LSTM and BERT have also been tested but not performed well due to limited dataset. In future, the current framework will be tested on a large data corpus related to the national crisis that happened due to disinformation to find the most acceptable model.

**Table 15** Popular hashtags inciting hate on Twitter

| #Islamophobia | 825.5K | 777.7K | 2.0B |
|---|---|---|---|
| #CoronaJihad | 538.0K | 504.1K | 914.4M |
| #NizamuddinIdiots | 156.7K | 151.2K | 276.7M |
| #TablighiVirus | 27.4K | 26.4K | 56.3M |

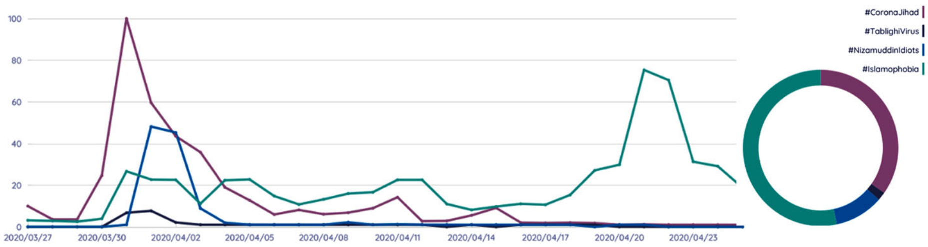
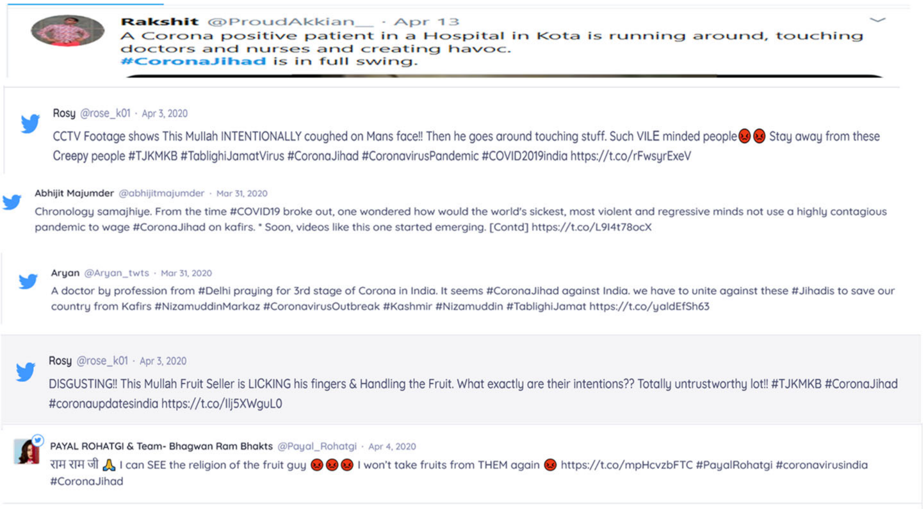


**Fig. 13** Popular hashtags and interest over time

## 7 Conclusion

The paper aims to provide a comprehensive model for the check-worthy topic that led to cyberwar in the form of communal riots. In this paper, we have developed a multiclass corpus named Fak_Cov by crawling data from Twitter when the event happened. The expert-based annotation has been done considering both authenticity and intention to label the dataset into three classes: real, disinformation and satire. Essentially, the features extracted have been inspired by well-established social science theories. This encourages interdisciplinary research on fake news detection. An integrated approach using thirty-one features of multiple perspectives such as style-based, propagation-based (particularly user engagement) has been projected. Experimental results based on the developed real-world corpus indicate that (1) Style-based features have shown acceptably well performance to distinguish disinformation with satire and real news. The best accuracy of 90% with the Ensemble classifier has been achieved with all iteratively combined style-based features. (2) The integrated features from multiple perspectives have slightly improved the performance in terms of accuracy and F-1 score of all the classifiers. (3) ANOVA test has been conducted for feature selection. Selected features further improved the performance and the highest accuracy of 94% has been obtained from Ensemble and MLP models. Thus, the selected features


Rakshit @ProudAkkian__ · Apr 13
A Corona positive patient in a Hospital in Kota is running around, touching doctors and nurses and creating havoc.
#CoronaJihad is in full swing.

Rosy @rose_k01 · Apr 3, 2020
CCTV Footage shows This Mullah INTENTIONALLY coughed on Mans face!! Then he goes around touching stuff. Such VILE minded people😡😡 Stay away from these Creepy people #TJKMKB #TablighiJamatVirus #CoronaJihad #CoronavirusPandemic #COVID2019india https://t.co/rFwsyrExeV

Abhijit Majumder @abhijitmajumder · Mar 31, 2020
Chronology samajhiye. From the time #COVID19 broke out, one wondered how would the world's sickest, most violent and regressive minds not use a highly contagious pandemic to wage #CoronaJihad on kafirs. * Soon, videos like this one started emerging. [Contd] https://t.co/L9I4t78ocX

Aryan @Aryan_twts · Mar 31, 2020
A doctor by profession from #Delhi praying for 3rd stage of Corona in India. It seems #CoronaJihad against India. we have to unite against these #Jihadis to save our country from Kafirs #NizamuddinMarkaz #CoronavirusOutbreak #Kashmir #Nizamuddin #TablighiJamat https://t.co/yaldEfSh63

Rosy @rose_k01 · Apr 3, 2020
DISGUSTING!! This Mullah Fruit Seller is LICKING his fingers & Handling the Fruit. What exactly are their intentions?? Totally untrustworthy lot!! #TJKMKB #CoronaJihad #coronaupdatesindia https://t.co/Ilj5XWguL0

PAYAL ROHATGI & Team- Bhagwan Ram Bhakts @Payal_Rohatgi · Apr 4, 2020
राम राम जी 🙏 I can SEE the religion of the fruit guy 😡😡😡 I won't take fruits from THEM again 😡 https://t.co/mpHcvzbFTC #PayalRohatgi #coronavirusindia #CoronaJihad

**Fig. 14** Screenshots of Tweets spreading fake information to shift to anti-Muslim narrative

play a strong role to detect disinformation, satire and real news in the corpus. Overall, the model achieved the best performance of 99% and 98% accuracies with relative frequency vector feature over Ensemble and MLP models respectively. To systematically divulge further patterns in disinformation compared to real and satire content, one has to employee (1) more fundamental theories to extract more effective features such as spreader's information about followers, account details etc.; (2) experimental analysis on larger datasets related to this domain; (3) consider other forms of data apart from text such as audio, video etc. Not many studies have considered the multimedia content for the development of a fake news detection model. Deep learning models have not performed well with the current dataset; therefore, will be tested again on the large corpus. All these aforementioned points will be part of our future work. Correspondingly, the domain considering national crisis due to disinformation has not been covered yet; thus, it will be a contribution to build a comprehensive cross-domain fake news detection model. Finally, the comprehensive fake news detection model on a large corpus can assist Twitter to flag misleading tweets according to the probability score obtained.

## References

1. Ahmed H, Traore I, Saad S (2017) Detection of online fake news using n-gram analysis and machine learning techniques. In: International conference on intelligent, secure, and dependable systems in distributed and cloud environments. Springer, pp 127–138
2. Ahuja R, Bansal S, Prakash S, Venkataraman K, Banga A (2018) Comparative study of different sarcasm detection algorithms based on behavioral approach. Procedia Comput Sci 143:411–418
3. Ajao O, Bhowmik D, Zargari S (2019) Sentiment aware fake news detection on online social networks. In: ICASSP 2019-2019 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 2507–2511. IEEE
4. Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss). J Appl Ecol 43(6):1223–1232
5. Amado BG, Arce R, Fariña F. (2015) Undeutsch hypothesis and criteria based content analysis: a meta-analytic review. Eur J Psychol Appl Leg Context 7(1):3–12
6. Anburajan K, Andavar S, Elango P (2020) An empirical evaluation of name semantic network for face annotation. Recent Advances in Computer Science and Communications (Formerly:, Recent Patents on Computer Science) 13(4):557–571
7. Ayedee N, Manocha D et al (2020) Role of media (television) in creating a positive atmosphere in covid-19 during lockdown in india. Role of Media (Television) in Creating a Positive Atmosphere in COVID-19 during Lockdown in India (May 19:2020
8. Bandura A (2001) Social cognitive theory of mass communication. Media psychology 3(3):265–299
9. Bardenet R, Brendel M, Kégl B., Sebag M (2013) Collaborative hyperparameter tuning. In: International conference on machine learning, pp. 199–207
10. Boididou C, Middleton SE, Jin Z, Papadopoulos S, Dang-Nguyen DT, Boato G, Kompatsiaris Y (2018) Verifying information with multimedia content on twitter. Multimed Tools Appl 77(12):15545–15571
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
12. Conroy NK, Rubin VL, Chen Y (2015) Automatic deception detection: Methods for finding fake news. Proc Assoc Inf Sci Technol 52(1):1–4
13. Deutsch M, Gerard HB (1955) A study of normative and informational social influences upon individual judgment. J Abnorm Soc Psychol 51(3):629
14. Devassy BM, George S (2020) Dimensionality reduction and visualisation of hyperspectral ink data using t-sne. Forensic Science International p 110194
15. Elssied NOF, Ibrahim O, Osman AH (2014) A novel feature selection based on one-way anova f-test for e-mail spam classification. Res J Appl Sci Eng Technol 7(3):625–638
16. Fersini E, Armanini J, D'Intorni M (2020) Profiling fake news spreaders: stylometry, personality, emotions and embeddings. In: CLEF Labs and workshops, notebook papers

17. Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. Comput Stat Data Anal 52(3):1694–1711
18. Gautam A, Jerripothula KR (2020) Sgg: Spinbot, grammarly and glove based fake news detection. In: 2020 IEEE Sixth international conference on multimedia big data (bigMM), pp 174–182. IEEE
19. Gettleman J, Schultz K, Raj S (2020) In india, coronavirus fans religious hatred. New York Times, April
20. Gillani N, Yuan A, Saveski M, Vosoughi S, Roy D (2018) Me, my echo chamber, and i: introspection on social media polarization. In: Proceedings of the 2018 World Wide Web Conference, pp 823–831
21. Glenski M, Weninger T, Volkova S (2018) Identifying and understanding user reactions to deceptive and trusted social news sources. arXiv:1805.12032
22. Goldberg Y, Levy O. (2014) word2vec explained:, deriving mikolov et al.'s negative-sampling word-embedding method. arXiv:1402.3722
23. Gravanis G, Vakali A, Diamantaras K, Karadais P (2019) Behind the cues: a benchmarking study for fake news detection. Expert Syst Appl 128:201–213
24. Hajek P, Barushka A, Munk M (2020) Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. Neural Comput Applic 32(23):17259–17274
25. Horne BD, Adali S (2017) This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news arXiv:1703.09398
26. Iwendi C, Ponnan S, Munirathinam R, Srinivasan K, Chang CY (2019) An efficient and unique tf/idf algorithmic model-based data analysis for handling applications with big data streaming. Electronics 8(11):1331
27. Iwendi C, Zhang Z, Du X (2018) Aco based key management routing mechanism for wsn security and data collection. In: 2018 IEEE International conference on industrial technology (ICIT), pp 1935–1939. IEEE
28. Johnson MK, Raye CL (1981) Reality monitoring. Psychol Rev 88(1):67
29. Kajan E, Faci N, Maamar Z, Sellami M, Ugljanin E, Kheddouci H, Stojanović DH, Benslimane D (2020) Real-time tracking and mining of users' actions over social media. Comput Sci Inf Syst 17(2):403–426
30. Kasthuri A, Suruliandi A, Raja S (2019) Gabor-oriented local order feature-based deep learning for face annotation. International journal of wavelets, multiresolution and information processing 17(05):1950032
31. Khan JY, Khondaker M, Islam T, Iqbal A, Afroz S (2019) A benchmark study on machine learning methods for fake news detection. arXiv:1905.04749
32. Kumar M, Rath NK, Swain A, Rath SK (2015) Feature selection and classification of microarray data using mapreduce based anova and k-nearest neighbor. Procedia Computer Science 54:301–310
33. Kumar S, Shah N (2018) False information on web and social media:, A survey. arXiv:1804.08559
34. Kuran T, Sunstein CR (1998) Availability cascades and risk regulation. Stan L Rev 51:683
35. Kušen E, Strembeck M (2018) Politics, sentiments, and misinformation: an analysis of the twitter discussion on the 2016 austrian presidential elections. Online Soc Netw Media 5:37–50
36. McCornack SA, Morrison K, Paik JE, Wisner AM, Zhu X (2014) Information manipulation theory 2: a propositional theory of deceptive discourse production. J Lang Soc Psychol 33(4):348–377
37. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
38. Mohseni S, Ragan E, Hu X (2019) Open issues in combating fake news:, Interpretability as an opportunity. arXiv:1904.03016
39. Moreno-Sandoval LG, Puertas E, Pomares-Quimbaya A, Alvarado-Valencia JA (2020) Assembly of polarity, emotion and user statistics for detection of fake profiles. CLEF Labs and Workshops, Notebook Papers
40. Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fmri: a tutorial overview. Neuroimage 45(1):S199–S209
41. Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R (2017) Automatic detection of fake news. arXiv:1708.07104
42. Posadas-Durán JP, Markov I, Gómez-Adorno H, Sidorov G, Batyrshin I, Gelbukh A, Pichardo-Lagunas O (2015) Syntactic n-grams as features for the author profiling task Working Notes Papers of the CLEF
43. Potthast M, Kiesel J, Reinartz K, Bevendorff J, Stein B (2017) A stylometric inquiry into hyperpartisan and fake news. arXiv:1702.05638
44. Rashkin H, Choi E, Jang JY, Volkova S, Choi Y (2017) Truth of varying shades: Analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 2931–2937
45. Reis JC, Correia A, Murai F, Veloso A, Benevenuto F (2019) Supervised learning for fake news detection. IEEE Intell Syst 34(2):76–81

46. Rout JK, Singh S, Jena SK, Bakshi S (2017) Deceptive review detection using labeled and unlabeled data. Multimed Tools Appl 76(3):3187–3211
47. Savyan P, Bhanu SMS (2020) Ubcadet: detection of compromised accounts in twitter based on user behavioural profiling. Multimedia Tools and Applications
48. Schreiner C, Torkkola K, Gardner M, Zhang K (2006) Using machine learning techniques to reduce data annotation time. In: Proceedings of the human factors and ergonomics society annual meeting, vol 50, pp 2438–2442. SAGE Publications Sage CA: Los Angeles, CA
49. Shao C, Ciampaglia GL, Varol O, Yang KC, Flammini A, Menczer F (2018) The spread of low-credibility content by social bots. Nature communications 9(1):1–9
50. Shu K, Wang S, Liu H (2017) Exploiting tri-relationship for fake news detection. arXiv:1712.07709, vol 8
51. Siering M, Koch JA, Deokar AV (2016) Detecting fraudulent behavior on crowdfunding platforms: the role of linguistic and content-based cues in static and dynamic contexts. J Manag Inf Syst 33(2):421–455
52. Stefanidis A, Cotnoir A, Croitoru A, Crooks A, Rice M, Radzikowski J (2013) Demarcating new boundaries: mapping virtual polycentric communities through social media content. Cartogr Geogr Inf Sci 40(2):116–129
53. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J (2012) Brat: a web-based tool for nlp-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp 102–107
54. Tabachnick BG, Fidell LS (2007) Experimental designs using ANOVA. Thomson/brooks/cole Belmont CA
55. Trstenjak B, Mikac S, Donko D (2014) Knn with tf-idf based framework for text categorization. Procedia Engineering 69:1356–1364
56. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151
57. Zannettou S, Sirivianos M, Blackburn J, Kourtellis N (2019) The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. Journal of Data and Information Quality (JDIQ) 11(3):1–37
58. Zhang D, Zhou L, Kehoe JL, Kilic IY (2016) What online reviewer behaviors really matter? effects of verbal and nonverbal behaviors on detection of fake online reviews. J Manag Inf Syst 33(2):456–481
59. Zhao Z, Zhao J, Sano Y, Levy O, Takayasu H, Takayasu M, Li D, Wu J, Havlin S (2020) Fake news propagates differently from real news even at early stages of spreading. EPJ Data Science 9(1):7
60. Zhou X, Zafarani R (2018) Fake news:, A survey of research, detection methods, and opportunities. arXiv:1812.00315
61. Zhou X, Zafarani R (2019) Fake news detection: an interdisciplinary research. In: Companion Proceedings of The 2019 World Wide Web Conference, pp 1292–1292
62. Zhou X, Zafarani R (2020) A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR) 53(5):1–40
63. Zuckerman M, DePaulo BM, Rosenthal R (1981) Verbal and nonverbal communication of deception. In: Advances in experimental social psychology, vol 14, pp 1–59. Elsevier