Online Social Networks and Media 9 (2019) 17–29

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Where should one get news updates: Twitter or Reddit

Shalini Priya*, Ryan Sequeira, Joydeep Chandra, Sourav Kumar Dandapat

Department of Computer Science and Engineering Indian Institute of Technology Patna, Patna 801103, India

ARTICLE INFO

Article history:
Received 11 April 2018
Revised 13 November 2018
Accepted 16 November 2018

Keywords:
Social media analysis
Reddit Vs. Twitter
Journalistic requirement

ABSTRACT

The last decade witnessed an enormous growth in popularity of several social media platforms. Although these platforms are generally meant to share information and opinions, their ubiquity is being increasingly exploited for spreading news and events in real time. Hence these platforms have become a natural choice for news agencies for getting updates, comments and experts' opinion of ongoing events which is crucial to understand the societal impact and for writing reports/editorial. However, with the plethora of these platforms available, each having its own uniqueness in content presentation, spreading patterns and also in the user interests, a comparative study of the efficacy of these platforms for different journalistic purposes would be useful.

In this paper, we perform a comparative study of two leading social media platforms *Reddit* and *Twitter*. We have analyzed Reddit comments and Twitter feeds of six news categories to establish the efficacy of these platforms in terms of different journalistic requirements. Observations reveal that there exist significant differences across these platforms that can be suitably exploited depending upon the scope of the requirements; for example, while Twitter is a better choice for the evolutionary study of events, Reddit is the more natural choice for exploration during the initial phase of any event. While the availability and spread of updated information on Twitter can be key in emergency and disaster situations, critical analysis of posts in Reddit can be important for editorials.

1. Introduction

Increasing use of smartphones and high penetration of data connectivity have lent to the enormous growth in popularity of social media platforms. The propensity of users to share news updates and opinions over social media platforms makes them an important source for obtaining news and reactions in real time. According to Pew Research center survey report in 2017[1], 67% of the adults get news from social media. The news agencies consider these platforms as a rich source of information and use them for mining news and user opinions [1]. The recently released Cision 2017 Global Social Journalism revealed that nearly half of the journalists can not do their work without help of social media[2]. Journalists use the social media for meeting various journalistic requirements that include obtaining breaking news and real-time news-updates, finding opinions, enriching arguments, confirmation of facts and deriving information source [2–5]. A large variety of social media platforms are currently in place that includes, microblogs, discussion forums, question-answer sites and social news aggregators to name a few [6]. Although most of these platforms are used for sharing facts and updates, sharing users' view and for the propagation of news events, the microblog and the news aggregator sites are currently playing a predominant role in event identification, news propagation, and opinion sharing [7–9]. Twitter has traditionally been recognized as the major microblogging platform for obtaining such updates, however, issues like short content lengths, large vocabulary gaps and the inherently noisy nature of the tweets pose subtle challenges in mining the required information [10,11].

On the other hand with the increasing popularity of news aggregator sites like Reddit, journalists are exploring these platforms for meeting the journalistic requirements mentioned above. Thus, considering the diverse features, functionalities and user behavior of the microblog and news aggregation platforms, one of the current needs is to determine the suitability of a platform in mining relevant information depending upon the specificity of the requirement. For example, platform "A" might provide quicker updates on an event, whereas platform "B" might provide more critical analysis on the same events. Further, platform "A" might

Conflict of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

* Corresponding author.
E-mail addresses: shalini.pcs16@iitp.ac.in (S. Priya), ryan.mtcs16@iitp.ac.in (R. Sequeira), joydeep@iitp.ac.in (J. Chandra), sourav@iitp.ac.in (S.K. Dandapat).

[1] http://www.pewresearch.org/fact-tank/2017/10/04/key-trends-in-social-and-digital-news-media/

[2] https://www.cision.com/us/resources/white-papers/2017-global-social-journalism-study/?tc=beyondbylines

https://doi.org/10.1016/j.osnem.2018.11.001
2468-6964/

discuss political news with more rigor than platform "B" while platform "B" might be a good choice for movie reviews. As social media is mainly driven by normal users, the suitability of a platform, however, depends on its usage by the corresponding users. For example, if an event is being discussed by a significant number of people, then the event is important. For emergency situation getting frequent updates is important and hence platform with active users is more suitable for this kind of event. Thus, different characteristics of the contents posted, posting behavior of the users and spreading pattern of the posts across these platforms can prove to be suitable for meeting specific journalistic requirements like searching important events, obtaining live updates or analyzing news stories.

In this study, we explore two social media platforms, Twitter and Reddit, (one of which is a leading microblog and the other a popular news aggregation platform) with the objective of comparing the relative benefits that these two platforms provide in obtaining information catering different journalistic needs. Both these platforms pose major differences in their characteristics and key functionalities; whereas messages in Twitter propagate based on a follower–followee relation, Reddit posts are visible within interest-based communities called subreddits. Reddit users obtain the posts by subscribing to the subreddits of their interest. Moreover, there is a restriction on content length in Twitter while no such restriction is imposed in Reddit. Based on a detailed investigation of the literature, we discover that the features that play major role in meeting the journalistic requirements can be primarily grouped into three major classes — (a) social features of the users, (b) temporal features and c) content-based features of the postings [12,13]. For each of these classes, we propose certain measurable attributes that suitably represent a class and observe their variations on Twitter and Reddit for several categories of news. The news categories are selected based on the Wikipedia "Current Events" portal that classify daily news events into six major categories namely (*i*) *internal affairs*, (*ii*) *law and order*, (*iii*) *science, technology, and environment*, (*iv*) *disaster and emergency*, (*v*) *world affairs*, and (*vi*) *sports*.

We collected a total of around 500 news events that were popular in both Reddit and Twitter. Analysis of tweet feeds and Reddit comments for these events (that generated 2.3 million Reddit comments and 250.0 million tweets), based on the key identified features, reveal significant differences across these social media platforms. Observations indicate that each of these platforms poses significant advantages over other depending on the news category and the intended requirement. For example, Twitter is more suitable for understanding the evolution pattern of event chains (like the unfolding of a long-standing political crisis caused by a major event and the fallouts of the same) due to the larger lifespan of the tweets. On the other hand, Reddit is a better choice for exploration of events at their initial phases as the majority of Reddit comments arrives within a very short span and hence it becomes easier to get a holistic view.

The organization of the rest of the paper is as follows. In Section 2, we provide a brief overview of the related literature, then in Section 3, we outline the methodology of the proposed work. Description of the dataset and steps of pre-processing are outlined in Section 4. In Section 5, we highlight identified features for comparing two major social networking platforms. Section 6 highlights the experimental details, performance analysis, and results. Finally, Section 7 concludes this study with important insights.

2. Related works

In this section, we discuss the major works that explore the key role of social media with respect to the journalistic requirements and subsequently highlight important features that have been extensively used for catering different journalistic requirements.

2.1. Usage of social media in journalism

Over the past few years, several studies have been focusing on how social media is being exploited by journalists [14–18] to fulfill their requirements. In [4] Hasanain et al. explored different ways in which journalists use Twitter by asking questions. Their empirical investigation was based on the questions asked by Arab journalists on Twitter. It was observed that Twitter caters to several journalistic requirements like enriching arguments, finding facts and sources, understanding opinions and to disseminate information. Next, we discuss some of the journalistic needs as studied in literature.

2.1.1. Opinion mining

Social media websites are rich sources of data for opinion mining. Sentiment classifiers [19] and graph-based clustering technique [20] are used for mining opinions from social media posts. Even journalists post questions on social media to collect information and opinion of people on relevant topics and articles [21,22]. Slang present in the text is an indicator of the extreme views of users towards a particular topic [23,24]. SentiWordNet in conjunction with other lexical resources is used for the measuring of Internet slang and further the views towards the product or event [23]. Authors in [24] created an annotated dictionary of Internet slang to help identify the level of specific attitudes and moods in the posts.

2.1.2. Event discovery

Authors in [5] observed that social media also plays a major role in identifying issues that need to be reported. Identifying breaking news events from social media is another important aspect that has received tremendous attention. Authors in [25] showed that news usually first breaks in on Twitter. Authors in [26] also showed that news reported through Twitter significantly overlaps with that of newswire providers. For early detection of breaking news, the spread of the corresponding posts, the count of the likes, upvotes [27,28], and retweets of these posts [9,29] are considered as promising features.

2.1.3. Measuring event importance

Journalists perceive the importance of an ongoing event by following the discussions on social media. Importance of a discussion can be derived by observing the key personalities involved, the volume of discussion on a topic as well as the current trends [2]. Authors in [30] studied the level of importance that journalists attach to the types of information in their daily work.

2.1.4. Faster fact checking and verification

It is observed that fact checking and verification of general and background information is one of the major purposes for which journalists seek social media content. Further, lack of available time for reporting is the major problem encountered by the journalists. Ansari and Zuberi [31] indicated that media professionals are time-bound, and hence they primarily seek selective rather than exhaustive information. Survey conducted in [30] reported that around 61.5% of the respondents agreed that social media decreases the time used to gather information for reporting.

2.1.5. Summary extraction

Social media acts as a source of diverse information and views for important ongoing events. Extracting highlights or summary of an event from the vast information pool remains a challenge.

Nicholas et al. [32] presented a visual analytic tool, Vox Civitas, designed to help journalists and media professionals to extract news highlight from large-scale aggregations of social media content of broadcasted events. Further, in [20], the social media contents have been explored to obtain the summary of the diverse views of the users with respect to an event.

2.1.6. Identifying influential users

Influential users take more active role in persuading other users and spreading the news. They also help to identify important news topic. Hence identifying influential users is another important task involving the social media. Several user based features like the number of retweets and mentions of the posts of a user have been considered a measure of influence [33]. Other techniques like random walk and time-sensitive query approach [34,35] have also been used to measure user influence. To the best of our knowledge, there is no direct way to measure user influence on the Reddit platform, however, it is reported that users with high karma points have more high-quality content contribution [36].

2.1.7. Identifying rumour

Social media is often used to propagate dis-information and rumors, however, social media contents also provide cues that can be used to identify these news. Authors in [37,38] studied the dissemination of false rumors and legitimate news during crisis, through the Twitter network. Tweets about factual news spread widely and quickly, but tweets about dis-information underwent a greater number of modifications in content over the spreading process [38]. Such artifacts in the contents can be used to distinguish false news from real ones.

Although these studies indicate that Twitter assumes huge significance in terms of journalism based services, however, it has also been observed that a single information source usually does not cater to these diverse requirements of the journalists [39]. Selection of information source highly depends on the purpose and also on the news category (like politics, sports etc.). We next provide a brief overview of the features that have been investigated to cater to the above mentioned journalistic needs for different news categories.

2.2. Investigating the key features

We next discuss few major features that are exploited for different journalistic needs for different news categories. These set of features can be primarily categorized into three categories as *social, temporal* and *content features*.

2.2.1. Social features

Social features include attributes that represent the social activities and influence of the users in the social network engaging with respect to a content Most studied social features include the following:

Re-tweets count and Karma points: To estimate the social position or influence of users', features like retweet count (total no of retweets by other users for the tweets posted by a user), karma point (depends on different parameters like how much a user posted on Reddit and how popular those posts are) are extensively used in many studies [33,35,36]. Re-tweet count is also used to measure the popularity of an event or to find breaking news. Here re-tweet count indicates total no of re-tweets of relevant posts of an event [9,29].

Up-votes, like, favourite: These features are extensively used to asses the popularity of an ongoing event and to identify if the event should be considered as breaking events [27,28].

Activity rate of user: An event spreads out more when the users involved in it are more active. Hence, activity rate is considered as a major feature for predicting the popularity of events and also for finding the trending events [40,41].

2.2.2. Temporal features

Temporal features include those attributes that vary with time.

Inter-arrival time: Inter-arrival time between comments/feeds can be used to identify the most happening events and also determine the possibility of obtaining real-time updates of an ongoing event. Lower inter-arrival time also indicates faster spreading [42–44].

Life-span of event: Lifespan on an event characterizes certain important properties of an event. For example, it has been observed that dis-information does not remain popular for a long time [45]. Moreover, life-span is also used to detect the formation and persistence of trends [46].

2.2.3. Content features

Content-based features include different textual characteristics and aspects of posts.

Additional Information: Additional information refers to certain information related to the entities linked with an event, like information about the key personalities, history of the location and history of other similar events. In several posts such information is embedded through URL that redirects to related websites, videos and photographs [47]. These information are extracted by finding named entity in post and linking those to knowledge bases [48,49].

URL's: A huge amount of content posted in conjunction with popular events makes a challenge in finding the trustworthy sources and information verification for the journalists. Hypertext links are effective for trust transfer [50], information verification [51] and getting the background information [52].

Objectivity and Subjectivity: A post which describes mere fact about an event rather than augmenting with the user opinion can be classified as objective post while posts with users' view can be classified as subjective. Objectivity is an important feature which measures how much factual content is present in a post [53–55]. The authors in [53,54] used TextBlob Python library to find the objectivity of the tweets. Supervised classifier like Naïve Baye's was trained on twitter dataset to find the subjectivity of the Reddit comments that is crawled from the world news subreddit [55].

Slangs and Non-dictionary words: Slangs typically represent extreme view/attitude of users towards a topic [23,24]. SentiWordNet in conjunction with other lexical resources are used for assigning score of Internet slangs and further to estimate the views of users' towards a product or an event [23]. Author in [24], created an Internet slang annotated dictionary to help in identifying the level of specific attitudes and moods.

Thus these broad dimensions of the works highlight the large spectrum of journalistic requirements and the different features that are exploited for extracting information for the same. With the help of these literature discussed above, we create a taxonomy(summary) that identify common types of needs that journalists have and the most suitable features to cater those requirements and it is shown in Table 1.

A vast of features have been investigated for meeting various key requirements however to the best of our knowledge, none of the works attempt to study the suitability of the social media platforms with respect to the features. Such a study would be useful in judging the suitability of a media platform based on the specific requirement.

3. Methodology

In this section, we outline the methodology adopted to analyze the differences in characteristics of Reddit and Twitter for different

Table 1
Taxonomy that summarize the importance of using features for the corresponding journalistic needs.

Journalistic needs	Social feature	Temporal feature	Content feature
Finding influential people	Retweets [35] & Karma points [36]	–	–
Find information source	–	–	URL's [47]
Enrich the argument & background information	–	–	Additional information Gain [48,49]
Finding Opinion	–	–	Graph based approach [20]
Information verification & background information		Life-span of event [45]	URL [50–52]
Breaking News	Retweet-count [9,29] & Up-votes [27,28]	Spreading pattern [29,56]	–
Extreme Opinion	–		Slangs [24]
Real-time information	Activity rate [40,41]	Inter-response time [42–44]	–
Find Fact	–	Spreading pattern [53–55]	Objectivity [37,38]

news categories. For conciseness, messages posted on Reddit will be termed as *comments* and tweets (feeds and comments) will be represented as *feeds*. We next highlight the major functional differences between Reddit and Twitter.

3.1. Functional differences between Reddit and Twitter

Both Reddit and Twitter pose major differences in their characteristics and key functionalities. One of the major differences between Reddit and Twitter lies in the social connectivity of the users across these two platforms. In Reddit users create/follow a number of subreddits based on their interest. A subreddit is an interest based group or forum where members post and discusses on a specific topic of interest. Users are able to create their own content by: (1) submitting a link or URL; (2) writing a self-post; (3) giving opinion on other user's posts or comments. For example "/r/worldnews" is a subreddit with around 18M users and common interest of the users of this subreddit is world event. Similarly "/r/sports" is a subreddit where users post content related to sports events. On the other hand, in Twitter, users follow other users based on either personal(professional) relation or may be based on some common topic of interest between users.

Another important difference lies in the restriction imposed by Twitter in post length. Twitter imposes a limit of 280 characters (earlier it was 140 characters) on the posted tweets[3]. Unlike Twitter, Reddit comments can be much larger as it has no such length constraints. This encourages users to write critical analysis with several additional information related to a post. Thus these two platforms show significant differences, not only in their content and posting patterns but also in the behavioral aspects of the users.

3.2. Outline of the proposed approach

The basic objective of this study is to understand the potential of two major OSN platforms, Twitter and Reddit, in providing useful information to cater different journalistic requirements. This section provides a brief outline of the approach used for comparing both platforms for the aforesaid objective.

In this study, we have considered the news categories based on the "Current Events" portal of Wikipedia. Major categories as per this portal are (*i*) Internal Affairs (IA), (*ii*) Law and Order (LA), (*iii*) Science, Technology, and Environment (STE), (*iv*) Disaster and Emergency (DE), (*v*) World Affairs (WA) and (*vi*) Sports(S). For each of the news categories, we consider only those events that are widely discussed on both the platforms(discussed in Section 4).

Subsequently, we extract several features from the comments and feeds corresponding to these news articles to characterize their differences across both the platforms.

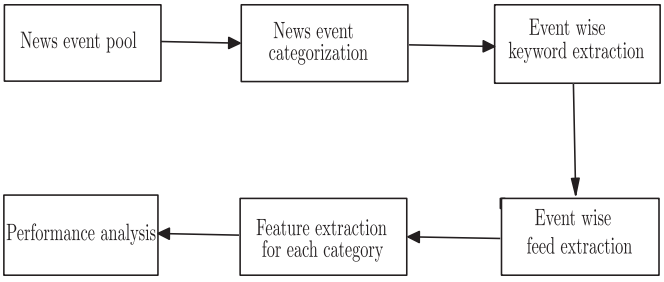


Fig. 1. Block diagram of proposed approach.

A summary of the steps followed is highlighted in Fig. 1. We next detail the steps followed for news categorization. We followed a multi-class classification based approach to determine the appropriate category of a given news article (as discussed in Section 3.3).

3.3. News categorization

We collected 500 news article which were widely discussed in Reddit as well as Twitter (detail is provided in Section 4).

Subsequently, the news articles were classified into one of the predefined categories. Using our annotated dataset we trained several classifiers like naïve bayes, SVM, logistic regression and decision tree. Each of these articles was annotated by 2 independent annotators and the inter-rater agreement using the Cohen–Kappa measure [57,58] was found to be **76.4%**. The evaluation of inter-rater agreement is based on the comparison of actual pairwise agreement among the annotators with the possibility of their agreement by chance. High Cohen–kappa value indicates good quality annotation.

Similar to the method adopted in [59], we train all the models using unigram and bigram features. We consider **80**% of total articles as training set and the rest as the test set. The best accuracy (around 65.8%) is achieved by SVM and hence we use this model for classification. The comparison table of the performance for different classifiers is given in Table 2.

We next describe briefly the scopes of each of the news categories that we have considered.

1. *Disaster and Emergency (D&E) :* This category is considered for articles related to natural disasters and emergency situations that have resulted in severe damages or loss of life and property. Typical examples include news related to hurricanes and storms, earthquakes, terror attacks etc. A key feature of this news category is the sudden burst of social media posts that provide updated information about the damages.
2. *Science, Technology, and Environment (ST&E):* A large set of news in the hot news section was related to advancements in sci-

[3] We collected the data during August and September of 2017, when the limit of Twitter post was 140.

Table 2
Performance comparison of different classifiers.

Classifier	Precision	Recall	F-measure	Accuracy
Logistic regression	0.51	0.55	0.52	0.52
Decision tree	0.41	0.67	0.50	0.43
Naive Bayes	0.45	0.54	0.49	0.51
Support Vector Machine (SVM)	**0.63**	**0.58**	**0.60**	**0.658**

ence, technology and environmental issues like climate change. The discussions in social media on these topics also center around the newly developed technologies and spreading awareness about environmental issues. Some typical examples in this category include Falcon Heavy Rocket Test, Ozone depletion etc.

3. *World Affairs (WA):* This category includes news events which find attention from several nations. Typical examples include news from North Korea crisis, South China Sea issue and Brexit.
4. *Internal Affairs (IA):* We consider news events that are specific to a particular region or country as matters of Internal Affairs. We annotate an event as local if no more than one, i.e., the country of occurrence is mentioned in the article. News articles in which more than one country name features are not considered as event of this type. For example: "Demonetization in India" would be covered as an internal affair event.
5. *Law and Order (L&O):* This kind of articles mainly provide information about some new rules of a country or breaking a law particular to that country. For example, the news article with the headline, "Pope admits Vatican has 2000-case backlog of sexual abuse cases" would be typically considered in this category.
6. *Sports(S):* Sports category include all news related to national or international sports events that have generated high interest among the followers.
   For example, "Philadelphia Eagles Beat New England Patriots 41–33 in the Superbowl 52".

We next highlight the details of the dataset used for our study and the preprocessing techniques used.

4. Dataset statistics and preprocessing

In this section, we highlight the dataset and the preprocessing techniques applied to the same.

4.1. Dataset

We have extracted the news articles from Reddit with the help of publicly accessible PRAW API (https://praw.readthedocs.io/en/latest/). We crawled the posts from the *World News* subreddit, *Hot News* section (https://www.reddit.com/r/politics/hot/), that were posted over a continuous period of 2 months from 1st August, 2017 to 30th September, 2017. World news subreddit contains all news category considered in this study except sport. For getting sport articles we crawled the *Hot News* section of *Sports* subreddit. The collected dataset are not location specific, the articles are crawled from hot news irrespective of its location. All the news articles, tweets and comments are in English language. Each news post was examined manually to determine the corresponding event. Multiple news posts that mapped to the same event were subsequently discarded from the data set.

Initially, we extracted news articles from the posts that were hot listed every day, thereby collecting around 800 news events in total. Out of these, we found that around 500 articles, representing different news events, were widely discussed on Twitter as well. An event is considered to be widely discussed if there are at least 5000 tweets/retweets corresponding to that event and if it is reported in "Hot News" section of subreddit.[4] However the choice of this threshold 5000 is not very principled. This threshold have been set based on the statistics of trending events that were observed for few days. We classified these articles into 6 categories using the procedure detailed in Section 3.3. Approximately 55, 73, 82, 67, 113 and 110 events were present in each category IA, L&O, ST&E, D&E, WA, and S respectively. We collected the comments and tweets relevant to the news articles.

Relevant comment extraction: The comments related to the news article were also collected using the PRAW API, thereby collecting a total of around 2.3 million comments. Other available data like the up-votes, the arrival time of the comments, commenter name etc., were also crawled for each of the comments. The profile of the commenter was retrieved using the same API. The number of unique Reddit users obtained was approximately 25K.

Relevant traction: Tweets corresponding to the news articles were searched with the help of tweepy API based on the keywords that were extracted from the headlines. For each news article, we extracted the bi-grams from the news headlines to construct a keyword set which is used to extract the relevant tweets — similar to the method followed in several existing works [60].

A tweet is considered relevant if it contains any bi-gram from the keyword set. For every relevant tweet, the corresponding comments were also collected, thereby obtaining around 250 million tweet feeds (tweets and comments) for all the events. We also collected information like *retweet count, arrival time* of the tweet, *tweet creator name, follower count* and *followee count* etc. The number of unique Twitter users obtained was 52 K.

To verify whether the crawled tweets were really relevant to the articles, we calculate the cosine similarity between the title of the news article and corresponding crawled tweets for that article [42]. The average cosine similarity of all the articles with tweets is 0.70 which indicates that with high probability, the extracted tweets are relevant. We summarize some of the important details about the dataset in Table 3.

We next describe the preprocessing steps followed to clean and filter the collected data.

4.2. Preprocessing

Each collected tweet has several attributes. We select only the following attributes for further processing: tweet text, tweet id, tweet creation time, retweet and favorite count, user screen name and the number of followers of the users. We crawled the history of each Twitter user based on their user screen name and collected a maximum of their last 3200 tweets.[5] The information for these tweets contain the tweet ids, the timestamps at which they were posted and their retweet and favorite count.

Similarly for each Reddit user, we collected their comments and posts (last 1000 can be crawled using the API), the time of posting or commenting and the subreddit on which the post or comment has been done. Further, for all the keywords collected from both

[4] We have selected popular events so that a large number of Reddit comments and tweet feeds can be obtained for analysis.

[5] This is the maximum number that can be obtained for a user using the Tweepy API.

Table 3
Summary of data used in our experimental evaluation. This table gives the number of articles, approx value of total comments, total tweets, number of Reddit and the number of Twitter user in each category. The value shown in the parenthesis is the mean number of comments/tweets per event.

News class	#News events	#comments (Mean)	#tweets (Mean)	% of reddit user	% of twitter users
Internal affairs	55	36 K(546)	43 M(63 K)	48.2	34.6
World affairs	113	31 K(450)	48 M(68 K)	78	59.6
Sc., Tech. & Env.	82	32 K(529)	20 M(33 K)	64	38.5
Sports	110	88 K(764)	63 M(54 K)	81	42.3
Disaster and emergency	67	31 K(799)	20 M(52 K)	76	53.8
Law and order	73	14 K(376)	33 M(86 K)	68	40.4

the Reddit comments and tweet feeds, the contractions are initially expanded using a contracted verb form dictionary. For example, *I'm* will be expanded to *I am*. We, subsequently, remove every punctuation present in the sentences and remove the stop words using the nltk library. Further, we remove all duplicate tweets and comments. We also do not consider user mentions and re-tweet tags as a keyword of the tweet. We also perform stemming of the keywords using the Porter Stemmer of the nltk toolkit. Finally, to discover the diversity in opinions obtained, we use a preprocessing technique detailed in [20] to form the communities of the comments and tweets. We next discuss the major group of features along with the specific attributes that are subsequently used for comparing the social media contents across both the platforms.

5. Feature sets

As stated earlier, we consider three groups of features for our analysis: social, temporal and content based. These features are used extensively for extracting information for different journalistic requirements.

5.1. Social features

Social features of the social media contents include attributes that represent certain activities of the users with respect to the content. Some of these activities include retweeting, up or down voting, commenting etc. The social features play three important roles, (*a*) they provide a measure of the importance of the news that can help in predicting its popularity [61], (*b*) they help in obtaining an idea about the major topics of the discussion related to the news [62] and (*c*) they also help in finding influential users. The major attributes that are included in this category are as follows:

5.1.1. Up-votes and re-tweet count

A higher value of these parameters can indicate that the comments or tweets related to a news event are being followed by a large number of users and hence indicates its importance in the respective platform. In Reddit, we consider the mean up votes received (per category) by the comments and for Twitter, we consider the mean number of retweets (per category) of a feed. Thus if U_i is the number of up votes received by the ith comment of any event belonging to category 'c' and n_c is the total number of comments of category 'c', then the average number of the up votes in that category 'c' is given by

$$U_c^{(avg)} = \frac{1}{n_c}\sum_{i\in c}(U_i) \tag{1}$$

By using the same formula the average number of re-tweets $R_f^{(avg)}$ of any class of news events can be calculated.

5.1.2. Mean number of influential users

Influential users commenting on a news feeds/comments can provide an indication of the importance of the news. For Twitter, tweets from influential users can also determine the popularity of a news as these tweets are likely to spread rapidly through the follower network. However, Reddit does not provide such explicit spreading mechanism and hence the presence of influential users in the comments can only be used to determine the importance of the news. Since social media platforms possess unique social features, calculating the influence of a user requires considering features that are specific to the platform. For example, Reddit provides *Link Karma* and *Comment Karma* that are reward mechanisms for posting posts and comments respectively. Users upvote or downvote a post or comment made by a user thereby leading to a subsequent increase or decrease in the karma points. The users with high karma points are the users who have contributed high-quality content and make insightful, interesting comments [36]. For all the Reddit users who have commented in at least one collected news event, we consider the top 10% with the highest Karma points as influential. If k_a denotes the fraction of influential users who have commented on any event a belonging to news category e, then the mean number of influential Reddit users for the news category e is

$$K_e^{(c)} = \frac{1}{n_e}\sum_{a\in e}(k_a) \tag{2}$$

while n_e is number of event that belongs to category e.

For Twitter, existing literature use several measures of user influence that includes the follower count of a user [9], the average retweet count of her tweets [9,35] as well as several PageRank based measures [9,63]. Here we consider two measures of user influence, the number of followers of a user as well as the average retweet count of her tweets. Similar to Reddit users we consider the top 10% users with highest influence score who have commented on any collected event as influential Twitter users. If f_a denotes the fraction of influential Twitter users who tweeted in any event a that belongs to category e, then the mean number of influential Twitter users for the news category e is given as

$$K_e^{(f)} = \frac{1}{n_e}\sum_{a\in e}(f_a) \tag{3}$$

As stated earlier, the presence of influential users, determined based on PageRank based measures, in the news tweets for any news topic can provide an indication about its popularity that it is expected to receive. We cannot make such claims on Reddit feeds and hence avoid comparing these platforms based on such influence models.

5.1.3. Mean activity rate

To examine the mean activity rate, for each news category, we have extracted five random samples of 2000 users randomly (without replacement) from Reddit and Twitter to make the evaluation more robust and generalized. A selected user of Reddit in a particular news category must have posted a Reddit comment for that

news category. Same is true for the Twitter user as well. The users selected have posted either a Reddit comment or a tweet in the respective news category. We subsequently obtain all the Reddit comments posted by these Reddit users and the last 3200 tweets of the Twitter users for computing mean activity rate[6]. This parameter measures the activity rate of each user (irrespective of any specific news event) by observing their comment history. If n_u denotes the total number of comments or tweets posted by a user u, and $p_i^{(u)}$ and $p_{i+1}^{(u)}$ denote the time of posting the ith and $(i+1)$th comments/tweets, respectively, then the mean activity rate (MAR) of user u is given as

$$\alpha_u = \frac{1}{\frac{1}{n_u-1}\sum_{i=1}^{n_u-1}\left(p_{i+1}^{(u)} - p_i^{(u)}\right)} = \frac{1}{\frac{1}{n_u-1}\left(p_n^{(u)} - p_1^{(u)}\right)} \tag{4}$$

Thus the average MAR of all the users who commented on event a is given as

$$\mu_a^{(c)} = \frac{1}{n_{au}^c}\sum_u \alpha_u \tag{5}$$

while n_{au}^c denotes the number of users who commented at least once for event a. Similarly we denote the corresponding MAR of users who tweeted on event a as $\mu_a^{(f)}$.

5.2. Temporal features

Temporal features of the comments/feeds include those aspects that vary with time, like the mean inter-arrival time of the comments/feeds, their lifespan and so on. Temporal features considered for this study are as follows:

5.2.1. Spreading

This feature attempts to capture the dynamics of spreading by means of arrival rate and burstiness of comments/feeds after the occurrence of the event. While arrival rate (fraction of comments/tweets each hour) provides an idea of the popularity of the news event in that social media, burstiness provides us the idea of the important phase of the event. So to understand the arrival rate we find out the quartile statistics of the fraction of data arrived per hour and to understand the burstiness, we have computed how much time it takes to receive 25%, 50%, 75% and 100% of overall comments/feeds.

5.2.2. Lifespan of event

To derive the life span of an event, we consider the hourly frequency of the comments/tweets after the article was posted. We introduce a term called *last hour* in which the frequency dropped beyond 20% of the peak frequency for the first time after reaching the peak (maximum frequency). This method is a slight modification of the approach proposed in [45] to measure the life span of an event in Twitter. We define the lifespan (LS) of an event as the difference in timestamps between the first and last hour of the event. This measure provides an idea of the duration for which an event was able to attract the attention of users.

5.2.3. Inter-arrival time of comments/feeds

Inter-arrival time (IT) of comments or tweet feeds is defined as the time difference between the posting of subsequent comment/feeds relevant to the news event. Thus, if $\Delta_a^{(c)}$ denotes the mean inter-arrival time of the comments for event a, and $t_{a,i}^{(c)}$ represents the arrival time of the i^{th} comment of event a, then

$$\Delta_a^{(c)} = \frac{\sum_{i=1}^{n_a^{(c)}-1}(t_{a,i+1}^{(c)} - t_{a,i}^{(c)})}{n_a^{(c)}-1} = \frac{t_{a,n}^{(c)} - t_{a,1}^{(c)}}{n_a^{(c)}-1} \tag{6}$$

One may note that there is a subtle relation between activity-rate and inter-arrival time. If activity rate is higher then it is most likely that inter-arrival time will be lower. However, activity rate is a social feature which indicates how active a particular user is. On the other hand inter-arrival time is mainly defined on an event to capture how much attention that particular event is getting from users.

5.3. Content-based features

Content-based features deal with the typical textual characteristics of the comments and feeds across the platforms. These include the richness of information, the presence of extreme views (presence of slangs), non-dictionary words and the number of diverse opinions highlighted through the comments and feeds. It is to be noted that while calculating the formula for the content-based features of each comment/feeds, we normalize the comment/feed score by the number of words in the comment/feed respectively, to minimize the possibility of any length biases in Reddit comments. The measures used for each of these features are as follows:

5.3.1. Average number of slangs and non-dictionary words

Slangs in the comments and feeds can reveal expression of extreme views or biases of the readers. Slang identification is helpful in the detection of harmful, aggressive, offensive, and help-seeking messages [24]. An aggressive, offensive or hateful message expresses the extreme attitude/view towards a particular event. And hence, extreme opinion holder are likely to hold the biased view over time [64]. We identify the common slangs using two available online dictionaries, www.noslang.com and www.internetslang.com. Similarly, non-dictionary words are noise that may be misspelled words, or word short forms typical to that platform. Presence of non-dictionary words increases the noise content in the text and thus makes information mining difficult. We calculate the mean number of non-dictionary terms per word present in each comment or tweet. If $l_{ai}^{(c)}$ and $b_{ai}^{(c)}$ denotes the total number of words and number of non-dictionary words in the ith comment of article a, whereas $l_{ai}^{(f)}$ and $b_{ai}^{(f)}$ denote the same for the feeds, then the mean number of non-dictionary terms (MD) per word in the comments and feeds of article a would, respectably, be given as

$$A_a^{(c)} = \frac{1}{n_a^{(c)}}\sum_i \frac{b_{ai}^{(c)}}{l_{ai}^{(c)}} \quad \text{and } A_a^{(f)} = \frac{1}{n_a^{(f)}}\sum_i \frac{b_{ai}^{(f)}}{l_{ai}^{(f)}} \tag{7}$$

For an article a, $n_a^{(c)}$ and $n_a^{(f)}$ represents the number of comments and feed. Thus both these values are normalized to range between 0 and 1. We use a similar calculation to derive the mean slangs (MS) per word and denote the same as $S_a^{(c)}$ and $S_a^{(f)}$, respectively.

5.3.2. Information richness

We consider a comment or feed as information-rich if it provides certain additional information beyond the news article. As a measure, we introduce mean additional information gained per comment or feed that considers the number of additional named entities or noun phrases present in the comments and feeds (with respect to the news article). The named entities are extracted using the NLTK named entity recognizer[7], and noun phrases are extracted using nltk.tree python module after removing the non-dictionary words. Thus for news article a, if P_a represent the set of entities (noun and noun phrases) present in the news article, where as $Q_a^{(c)}$ and $Q_a^{(f)}$ represent the entities in comment c and

[6] This is the maximum number that can be obtained for a user using the Tweepy API.

[7] http://www.nltk.org/book/ch07.html

tweet f, respectively, then the information richness of c and f, are given by

$$I_a^{(c)} = \frac{|Q_a^{(c)} - P_a|}{|P_a \cup Q_a^{(c)}|} \text{ and } I_a^{(f)} = \frac{|Q_a^{(f)} - P_a|}{|P_a \cup Q_a^{(f)}|} \tag{8}$$

5.3.3. Additional information gain

Although information richness introduced above provides insight about the richness of the content in a comment or tweet, however a more stringent measure of content richness would be amount of additional information introduced by the comment or tweet. If any information provided in a recent comment or feed has already been introduced through a previous comment, then no additional information is gained by the recent comment. Thus, it would be more relevant to determine the cumulative information gain or the convergence of the additional information. Let us assume that we want to compute additional information gained due to the comment, c or feed, f. Further, let E represents the set of entities (noun and noun phrases) introduced through the news article as well as all the comments and feeds that are posted before the posting of c or f. Then the additional information gained through the comment (c) and feed (f) is given by

$$G_a^{(c)} = \frac{|E - Q_a^{(c)}|}{|E \cup Q_a^{(c)}|} \text{ and } G_a^{(f)} = \frac{|E - Q_a^{(f)}|}{|E \cup Q_a^{(f)}|} \tag{9}$$

For an article a, if $n_a^{(c)}$ and $n_a^{(f)}$ represents the number of comments and feeds, respectively, then the mean additional information (*MAI*) per comment and feed is given by

$$\hat{G}_a^{(c)} = \frac{G_a^{(c)}}{n_a^{(c)}} \text{ and } \hat{G}_a^{(f)} = \frac{G_a^{(f)}}{n_a^{(f)}} \tag{10}$$

5.3.4. Opinion diversity

Opinions regarding an event represent the individual viewpoints expressed by the readers about that event. This feature is important as in many cases it can help in understanding the critical aspects of a news that are of concern to the readers [20]. Thus for comparing the efficiency of the platforms with respect to this functionality, one of the major tasks is to identify the number of relevant yet diverse opinions expressed by the readers, with respect to a news event. For obtaining the same we use a graph community-detection based approach proposed in [20]. This approach, when applied to tweets, generates better results compared to traditional opinion summarization techniques like Latent Semantic Analysis (LSA) [65] and LexRank [66]. In this approach, tweets relevant to an event are linked with each other based on their content and contextual similarity to form an event-specific tweet-graph. The graph is then divided into communities based on the similarity of the opinions that the tweets represent. Subsequently, the communities that are most relevant to the news event are returned, where each community represents a unique view/opinion. We measure the opinion diversity of an article by considering the number of communities that are returned in the previous step. For an event, a, D_a^c, D_a^f denote the number of diverse communities (measurement of opinion diversity) in Reddit and Twitter, respectively.

5.3.5. Mean URL's

A larger value of this parameters can indicate that the posted comments or tweets contain the additional source that can help in verifying the contents and getting more information about that event. If R_i is the number of URL's mentioned in the i^{th} comment of any event belonging to news category 'y' that has a total of n_y comments, then the average number of the URL's in the comments of that category is given by

$$R_c^{(avg)} = \frac{1}{n_y} \sum_{i \in y} (R_i) \tag{11}$$

Using a similar formula the average number of URL's, $R_f^{(avg)}$, in the tweet feeds of any category can be calculated.

We next highlight the major observations that we capture.

6. Experimental results and insights

For each category of the news event, we investigate all three important feature sets mentioned in Section 5 and infer some interesting findings from the experimental results.

6.1. Comparison of social features

In this section, we outline the major differences of the social features observed in both the platforms and then highlight their implication. The social features are important as the users in a platform play a major role in creating a buzz around the news, thus determining its popularity and the attention it would receive as highlighted in Fig. 2.

6.1.1. Mean up votes/retweets

Fig. 2a shows the mean number of up votes and retweets received by the Reddit comments and tweet feeds, respectively, for the news in each category. This measure provides a measure of the attention a news receives in both the platforms. As can be seen, this pattern varies across the platforms depending upon the news types. While the average number of Reddit up votes is much higher for news related to internal affairs and law and order, the number of retweets is higher for science, technology, and environmental news as well as news related to disaster and emergency. The corresponding values are nearly similar in both platforms for the world affairs category. This observation highlights the fact that news related to disaster and emergency or scientific discoveries and technological advancements receive much more attention on Twitter than in Reddit.

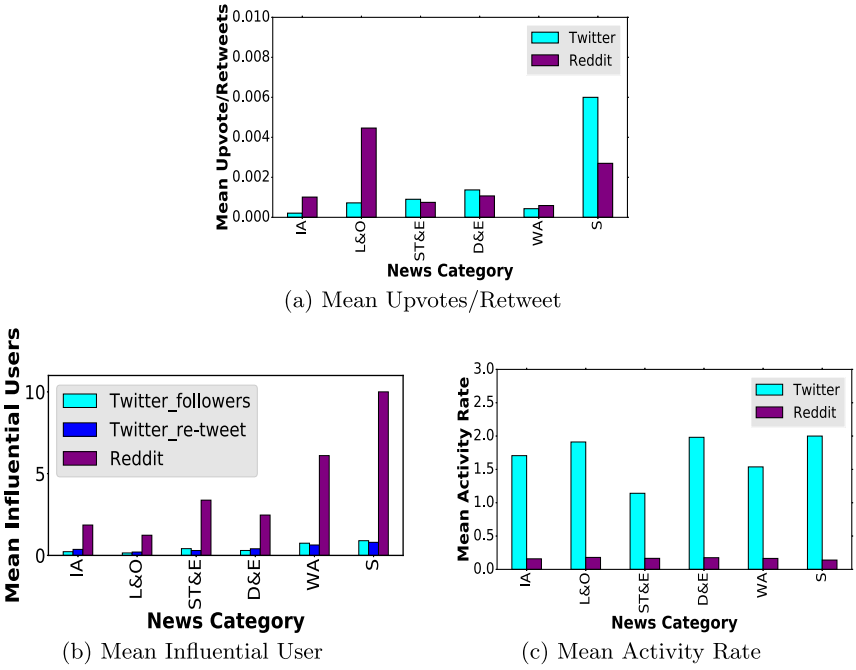
6.1.2. Mean influential users

Fig. 2b shows the fraction of influential users involved in both platforms for each news category. Result shows that fraction of influential users in Reddit (measured using karma points) is higher compared to Twitter in both the scenarios (1) when it is measured using followers count (2) when it is measured based on average retweet count. This is because our study includes events collected from "World news" subreddit and hence the Reddit user base remains same across all news categories (users subscribed to "world news" subreddit). However, for Twitter, although the events were popular, the user base is quite diverse and not many influential people were involved in every category, thus lowering the value of $K_e^{(f)}$. This observation highlights the importance of subreddits in obtaining expert opinions. Expert opinion or opinions of influential readers are more available if a news is aligned to any existing topical subreddit.

6.1.3. Mean activity rate

We next observe the mean activity rate of the users in each of the platforms for the different news categories. As shown in Fig. 2c, the mean activity rate of the users for all categories are much higher in Twitter as compared to Reddit. This is due to the fact that users on Reddit restrict their activities to their topic of interest.

Thus these observations highlight the fact that the social behavior of the users varies across both the platforms depending upon



(a) Mean Upvotes/Retweet

(b) Mean Influential User

(c) Mean Activity Rate

Fig. 2. Comparisons of social features.

the news category. Choosing the right platform, depending on the news category, is necessary when information is to be extracted from contents generated through active user engagement. Further, whereas news related to internal affairs and law and order are actively discussed in Reddit (and hence would be a more suitable platform for information extraction), updates related to disasters are more effectively communicated through Twitter. Experiments related to social features confirm that we should choose the right platform for greater coverage depending on the news category and journalistic requirements (for expert comment subscribe to suitable sub-reddit).

6.2. Comparison of temporal features

We compare the temporal features of both the platforms based on their spreading rate, the lifespan of the news and inter-arrival time of the comments and feeds in the respective platforms as defined in Section 3. Fig. 3 compares the temporal features for both Reddit and Twitter.

6.2.1. Mean comment/feed arrival rate

Fig. 3a shows the quartile graph of fraction of comments arrived per hour for each news category[8]. Variation in the fraction of comments received per hour is much higher on Reddit. More specifically, the variation is much more prominent in top 50 percentile data which indicates a burstiness in Reddit. On the contrary, Twitter exhibits more or less uniform arrival rate of comments. However, the median fraction of feeds received per hour in Twitter is always higher as compared to the comments in Reddit.

6.2.2. Burstiness of comments/feeds

The burstiness of the comments and feeds can be analyzed more precisely from Fig. 3b. The figure shows that irrespective of news category, 50% of total comments with respect to an event is received within the first 10% duration of the entire lifespan of the event in Reddit. On the other hand, in case of Twitter, 50% of feeds arrived in first 40% duration of the event lifespan. This implies that we can expect a heavy burstiness in Reddit comments in the first phase of the event. Thus Reddit can provide a quick update about an event in a short span of time after its occurrence as compared to Twitter. However, it has also been observed from our empirical data that popular events are first reported on twitter rather than on Reddit.

6.2.3. Lifespan of an event

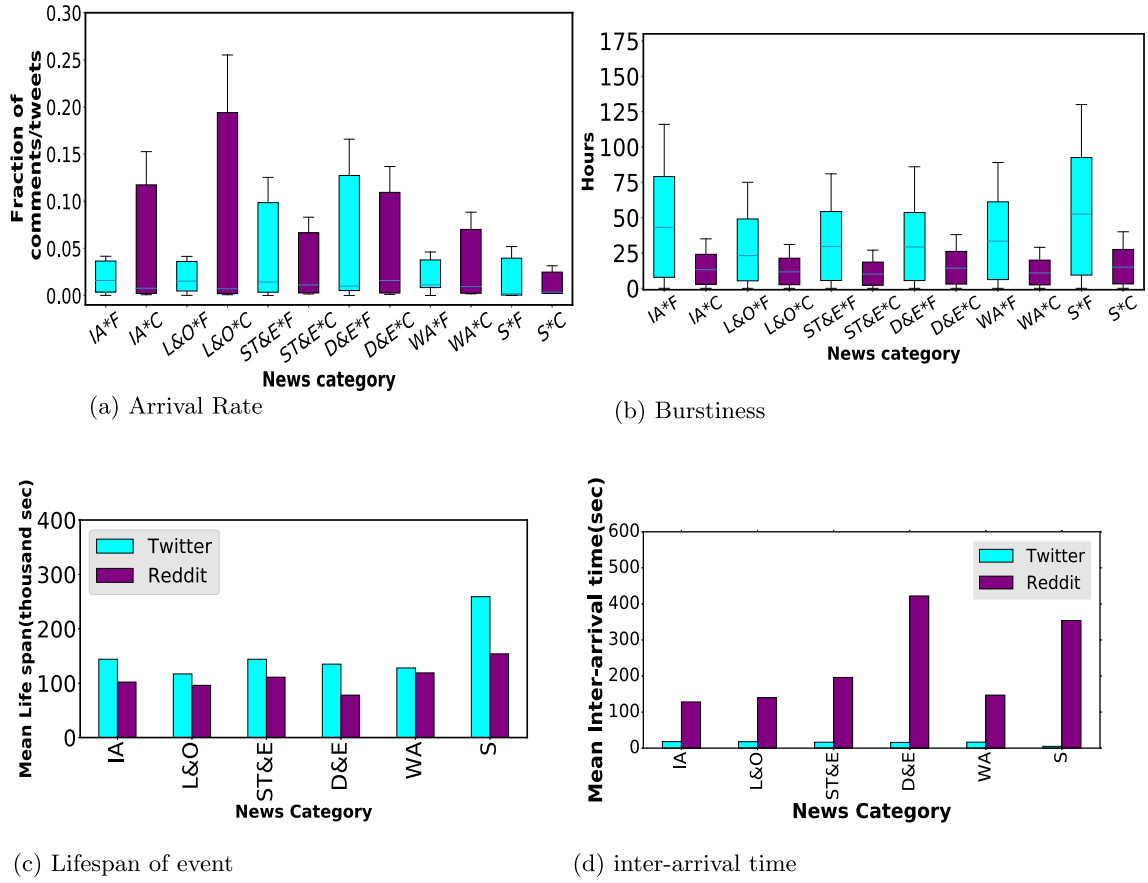
The life span of an event in Twitter is always longer as compared to Reddit (shown in Fig. 3c) irrespective of news category. This might be the effect of longer opinion convergence time of the large and more diverse user base of Twitter who are participating in a discussion. To estimate how opinion emerges in two platforms, we have done a small experiment. We compute the number of opinions (using the technique discussed in Section 5.3.4) that subsequently can also be used to represent the fraction of opinions emerging through both the platform every hour for all the news categories. Fig. 4 shows, the fraction of opinions added cumulatively with time for different types of events in both the platforms. Figure shows, within very short duration Reddit opinion converges or in other terms no more additional opinions are added. While in Twitter it takes longer time for convergence of opinion.

6.2.4. Mean inter-arrival time

The inter-arrival time of the tweets is much lower as compared to the Reddit feeds (shown in Fig. 3d) that points towards a presence of more active and larger user base in Twitter as compared to Reddit. This feature can help the journalists in real-time reporting and obtaining fast news updates.

These observations highlight the fact that Twitter can provide much quicker updates as compared to Reddit. This characteristic makes Twitter a natural choice for mining real-time updated information, i.e. for news categories where quick updates are required, like a disaster or emergency-related news and in situations of rapid political developments. However, Reddit comments arrive faster at the early stages of any event and hence it would be a better choice for the journalist to follow Reddit during the initial phases of an

[8] Extension *F and *C represents the Twitter feeds and Reddit comments, respectively.



(a) Arrival Rate

(b) Burstiness

(c) Lifespan of event

(d) inter-arrival time

Fig. 3. Comparisons of temporal features.

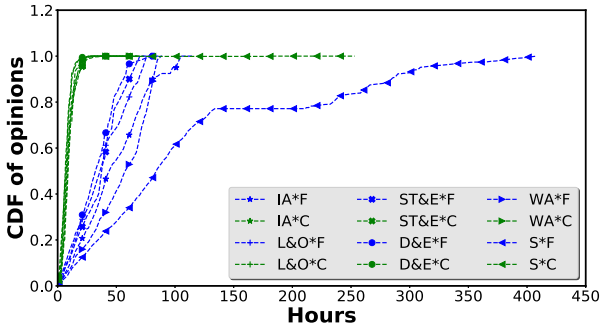


Fig. 4. Cumulative opinion across the platform.

event. The longer lifespan of the tweets makes it useful in studying the evolution and event chains of persisting news events, i.e. those events that are continuing for a long time, like a long-standing political crisis.

6.3. Comparison of content based features

As outlined in Section 3, we have considered three different parameters as content-based features: mean additional information, opinion diversity and average slangs and non-dictionary terms per word. Fig. 5 highlights the observations for the same.

6.3.1. Mean additional information

Fig. 5a compares the mean additional information in Reddit and Twitter for each of the news category. As can be observed, the mean additional information in Reddit comments is much higher for all the news categories except the sports events, indicating that Reddit can provide more additional relevant information as compared to Twitter.

6.3.2. Opinion diversity

Comparison of the number of diverse opinions, shown in Fig. 5b, also indicates a similar trend where the mean number of diverse opinions being higher on Reddit as compared to Twitter. The average number of diverse opinion is highest in law and order category for both comments and feeds. This is more natural considering the controversies associated with this news.

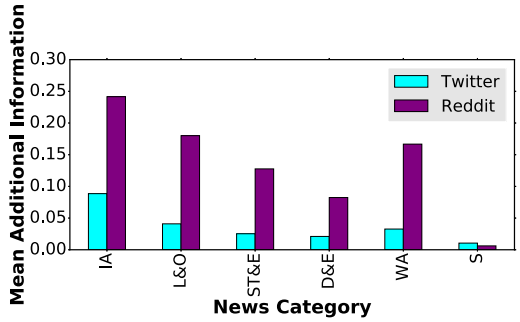
6.3.3. Mean number of non-dictionary/slang words

Comparison of the average number of slangs and non-dictionary terms per word, in both comments and feeds (shown in Fig. 5c), reveals that the fraction of slangs per word is higher in Reddit whereas the fraction of non-dictionary terms per word is much higher in Twitter. This indicates that Reddit comments contains more extreme or biased views as compared to tweet feeds [24].

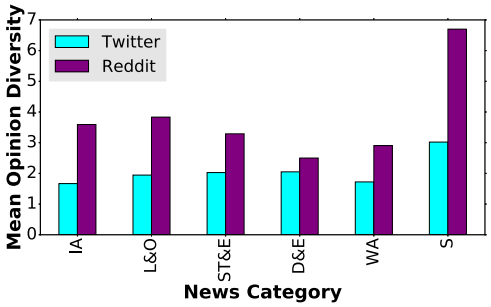
Thus obtaining the polarity or stance of a user can be much easier using Reddit comments. Further, the presence of the higher fraction of non-dictionary words in Twitter indicates tweets are noisier compared to Reddit comments and hence extracting information from tweets will be inherently challenging as compared to Reddit comments [67].

6.3.4. Mean number of URL's

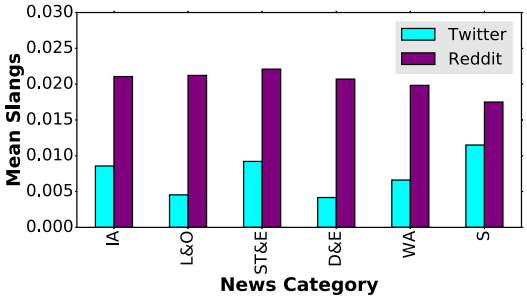
Fig. 5e compares the average URL's present in the comments/tweets. Observation reveals that fraction of URL's in the tweet feeds is more than in the Reddit comments except for science, technology, and environment news. As has been observed in



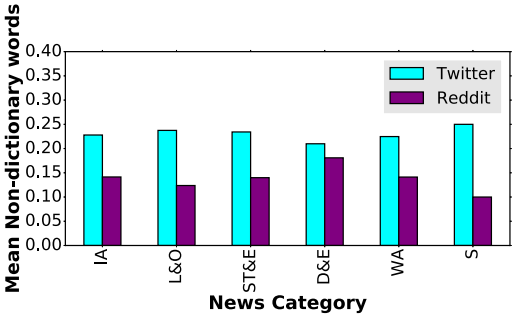
(a) Mean Additional Information



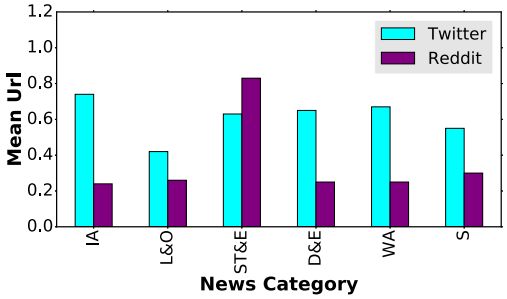
(b) Average Opinion diversity



(c) Mean number of slangs terms



(d) Mean number of non-dictionary terms



(e) Mean number of urls's

Fig. 5. Comparisons of content based features.

literature [50,51], the presence of URL's in a content enhances the credibility of the content. This implies that verification of content is quite easier with the tweets than comments.

6.4. Convergence of information gain

To compute the convergence of information gained through the comments (feeds) related to an article, we distributed them in different bins as per their arrival time. Then Eq. (9) is used to compute the average additional information gain in each hour.

Fig. 6 shows the cumulative information gain from comments and feeds of each news category. It depicts that additional information in case of Reddit converges rapidly that means it takes less time to collect all the information that are enriching the arguments. But in case of Twitter obtaining information that enriches an argument is a continuous process. News aggregators will have to wait until the end of the article lifespan to obtain all the information gained from the tweets.

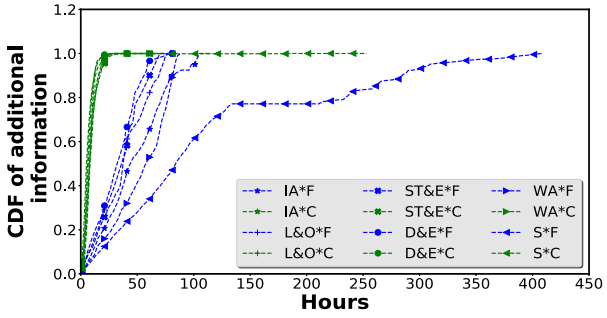


Fig. 6. Cumulative information gain.

Thus these observations indicate that Reddit comments would be more suitable to gain additional knowledge on matters related to an event, however, the major challenge lies in determining the

credibility of the news considering the extreme views and biases of the users commenting on the same.

7. Conclusion

The objective of this study was to understand the relative advantages that one social media platform can pose over other in meeting certain journalistic requirements. Our findings reveal that depending on the news category and information requirements, one platform can be better than the other. Bursty traffic at early phase makes Reddit suitable for providing a comprehensive view of opinions in the very short span of time. Reddit structure (presence of interest based subreddit) makes Reddit more suitable for finding experts' opinion. Unconstrained post length of Reddit plays important role and enriches us with more background information. On the other hand, imposition of constraint in post length helps in reducing biased and extreme views in Twitter compared to Reddit. Moreover, an event in Twitter is alive for a longer span of time that may be useful for its evolutionary analysis. Lower inter-arrival time makes Twitter suitable for getting a frequent update during an emergency or any live event. However, the behavior of the platforms are not independent of the news category, so information extraction from a social media platform must consider the news category for which the platform would be more appropriate.

Acknowledgment

This research work is sponsored by the project under the Visvesvaraya Ph.D. Scheme of Ministry of Electronics and IT, Government of India, being implemented by Digital India Corporation.

References

[1] E. J. Bell, T. Owen, P. D. Brown, C. Hauka, N. Rashidian, The platform press: how silicon valley re engineered journalism, Technical report, Tow Center for Digital Journalism (2017).
[2] U. Hedman, M. Djerf-Pierre, The social journalist: embracing the social media life or creating a new digital divide? Digit. J. 1 (3) (2013) 368–385.
[3] N. Diakopoulos, M. De Choudhury, M. Naaman, Finding and assessing social media information sources in the context of journalism, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2012, pp. 2451–2460.
[4] M. Hasanain, M. Bagdouri, T. Elsayed, D.W. Oard, What questions do journalists ask on twitter? in: Proceedings of the SMN@ ICWSM, 2016.
[5] O. Aghili, M. Sanderson, Journalists' information needs, seeking behavior, and its determinants on social media, arXiv:1705.08598 (2017).
[6] A. Cornia, A. Sehl, R. Nielsen, Developing Digital News Projects in Private Sector Media, 15, Reuters Institute for the Study of Journalism, University of Oxford, Luettu, 2017.
[7] D.T. Nguyen, J.E. Jung, Real-time event detection for online behavioral analysis of big social data, Fut. Gen. Comput. Syst. 66 (2017) 137–145.
[8] B.E. Weeks, A. Ardèvol-Abreu, H. Gil de Zúñiga, Online influence? Social media use, opinion leadership, and political persuasion, Int. J. Public Opin. Res. 29 (2) (2017) 214–239.
[9] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media? in: Proceedings of the 19th International Conference on World wide Web, ACM, 2010, pp. 591–600.
[10] M. Tsagkias, M. De Rijke, W. Weerkamp, Linking online news and social media, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, 2011, pp. 565–574.
[11] S. Priya, M. Bhanu, S.K. Dandapat, K. Ghosh, J. Chandra, Characterizing infrastructure damage after earthquake: a split-query based ir approach, in: Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 202–209.
[12] P. Burnap, M.L. Williams, L. Sloan, O. Rana, W. Housley, A. Edwards, V. Knight, R. Procter, A. Voss, Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack, Soc. Netw. Anal. Min. 4 (1) (2014) 206.
[13] E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne, Finding high-quality content in social media, in: Proceedings of the 2008 International conference on Web Search and Data Mining, ACM, 2008, pp. 183–194.
[14] R.K. Nielsen, A. Cornia, A. Kalogeropoulos, Challenges and Opportunities for News Media and Journalism in an Increasingly Digital, Mobile, and Social Media Environment, Council of Europe Report. DGI (2016)18. London: Oxford University, Reuters Institute for the Study of Journalism, 2016.
[15] M. Knight, C. Cook, Social media for journalists: principles and practice, SAGE Publications Limited, 2013.
[16] M. Brautovic, I. Milanovic-Litre, R. John, Journalism and twitter: between journalistic norms and new routines, MediAnali: Me Unarodni Znanstveni Casopis za Pitanja Medija, Novinarstva, Masovnog Komuniciranja i Odnosa s Javnostima 7 (13) (2013) 19–36.
[17] G. Enli, C.-A. Simonsen, 'social media logic' meets professional norms: Twitter hashtags usage by journalists and politicians, Inf. Commun. Soc. 21 (8) (2018) 1081–1096.
[18] R.W. Lariscy, E.J. Avery, K.D. Sweetser, P. Howes, An examination of the role of online social media in journalists' source mix, Public Relat. Rev. 35 (3) (2009) 314–316.
[19] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining., in: Proceedings of the LREc, 10, 2010, pp. 1320–1326.
[20] R. Chakraborty, M. Bhavsar, S. Dandapat, J. Chandra, A network based stratification approach for summarizing relevant comment tweets of news articles, in: Proceedings of the International Conference on Web Information Systems Engineering, Springer, 2017, pp. 33–48.
[21] M. Brauto, I. M Litre, R. John, Journalism and twitter: between journalistic norms and new routines, MediAnali 7 (13) (2013) 19–36.
[22] M. Revers, The twitterization of news making: transparency and journalistic professionalism, J. Commun. 64 (5) (2014) 806–826.
[23] F.M. Kundi, S. Ahmad, A. Khan, M.Z. Asghar, Detection and scoring of internet slangs for sentiment analysis using sentiwordnet, Life Sci. J. 11 (9) (2014) 66–72.
[24] H.-N. Teodorescu, N. Saharia, An internet slang annotated dictionary and its use in assessing message attitude and sentiments, in: Proceedings of the International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2015, IEEE, 2015, pp. 1–8.
[25] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, K.-L. Ma, Breaking news on twitter, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12, ACM, New York, NY, USA, 2012, pp. 2751–2754.
[26] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, L. Shrimpton, Can twitter replace newswire for breaking news?, in: E. Kiciman, N.B. Ellison, B. Hogan, P. Resnick, I. Soboroff (Eds.) Proceedings of the ICWSM, 2013.
[27] G. Stoddard, Popularity dynamics and intrinsic quality in reddit and hacker news., in: Proceedings of the ICWSM, 2015, pp. 416–425.
[28] A.C. Leavitt, Upvoting the news: breaking news aggregation, crowd collaboration, and algorithm-driven attention on reddit. com, University of Southern California, 2016 (Ph.D. thesis).
[29] S. Phuvipadawat, T. Murata, Breaking news detection and tracking in twitter, in: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE, 2010, pp. 120–123.
[30] M.A. Anwar, H. Al-Ansari, A. Abdullah, Information seeking behaviour of Kuwaiti journalists, Libri 54 (4) (2004) 228–236.
[31] M.N. Ansari, N.A. Zuberi, Information needs of media practitioners in Karachi, Pakistan, Chin. Lib. Int. Electron. J. 33 (2012).
[32] N. Diakopoulos, M. Naaman, F. Kivran-Swaine, Diamonds in the rough: social media visual analytics for journalistic inquiry, in: Proceedings of the Visual Analytics Science and Technology (VAST), IEEE, 2010, pp. 115–122.
[33] M. Cha, H. Haddadi, F. Benevenuto, P.K. Gummadi, et al., Measuring user influence in twitter: the million follower fallacy. 10 (10–17) (2010) 30.
[34] R. Ghosh, K. Lerman, Predicting influential users in online social networks, arXiv:1005.4882 (2010).
[35] K. Subbian, C.C. Aggarwal, J. Srivastava, Querying and tracking influencers in social streams, in: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, ACM, 2016, pp. 493–502.
[36] T. Weninger, An exploration of submissions and discussions in social news: mining collective intelligence of reddit, Soc. Netw. Anal. Min. 4 (1) (2014) 173.
[37] M. Mendoza, B. Poblete, C. Castillo, Twitter under crisis: can we trust what we rt? in: Proceedings of the First Workshop on Social Media Analytics, ACM, 2010, pp. 71–79.
[38] S.M. Jang, T. Geng, J.-Y. Q. Li, R. Xia, C.-T. Huang, H. Kim, J. Tang, A computational approach for examining the roots and spreading patterns of fake news: evolution tree analysis, Comput. Hum. Behav. 84 (2018) 103–113.
[39] G. Singh, M. Sharma, Information seeking behavior of newspaper journalists, Int. J. Lib. Inf. Sci. 5 (7) (2013) 225–234.
[40] M. Mathioudakis, N. Koudas, Twittermonitor: trend detection over the twitter stream, in: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, ACM, 2010, pp. 1155–1158.
[41] M. Naaman, J. Boase, C.-H. Lai, Is it really about me? Message content in social awareness streams, in: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, ACM, 2010, pp. 189–192.
[42] J. Huang, R. Kornfield, G. Szczypka, S.L. Emery, A cross-sectional examination of marketing of electronic cigarettes on twitter, Tob. Control 23 (suppl 3) (2014) iii26–iii30.
[43] E.M. Clark, C.A. Jones, J.R. Williams, A.N. Kurti, M.C. Norotsky, C.M. Danforth, P.S. Dodds, Vaporous marketing: uncovering pervasive electronic cigarette advertisements on twitter, PloS One 11 (7) (2016).
[44] K. Rudra, S. Banerjee, N. Ganguly, P. Goyal, M. Imran, P. Mitra, Summarizing situational and topical information during crises, arXiv:1610.01561(2016).
[45] S. Wu, C. Tan, J.M. Kleinberg, M.W. Macy, Does bad news go away faster? in: Proceedings of the ICWSM, 2011.
[46] S. Asur, B.A. Huberman, G. Szabo, C. Wang, Trends in social media: persistence and decay., in: Proceedings of the ICWSM, 2011.

[47] J. Sankaranarayanan, H. Samet, B.E. Teitler, M.D. Lieberman, J. Sperling, Twitterstand: news in tweets, in: Proceedings of the 17th ACM Sigspatial International Conference on Advances in Geographic Information Systems, ACM, 2009, pp. 42–51.
[48] W. Shen, J. Wang, J. Han, Entity linking with a knowledge base: Issues, techniques, and solutions, IEEE Trans. Knowl. Data Eng. 27 (2) (2015) 443–460.
[49] B. Kleinberg, M. Mozes, A. Arntz, B. Verschuere, Using named entities for computer-automated verbal deception detection, J. Forensic Sci. 63 (3) (2018) 714–723.
[50] K.J. Stewart, Y. Zhang, Effects of hypertext links on trust transfer, in: Proceedings of the Fifth International Conference on Electronic Commerce, ACM, 2003, pp. 235–239.
[51] C. Castillo, M. Mendoza, B. Poblete, Predicting information credibility in time-sensitive social media, Internet Res. 23 (5) (2013) 560–588.
[52] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th International Conference on World Wide Web, ACM, 2011, pp. 675–684.
[53] T. Sahni, C. Chandak, N.R. Chedeti, M. Singh, Efficient twitter sentiment classification using subjective distant supervision, in: Proceedings of the Nineth International Conference on Communication Systems and Networks (COMSNETS), 2017, IEEE, 2017, pp. 548–553.
[54] S.K. Sharma, X. Hoque, Sentiment predictions using support vector machines for odd-even formula in Delhi, Int. J. Intell. Syst. Appl. 9 (7) (2017) 61.
[55] B.D. Horne, S. Adali, The impact of crowds on news engagement: a reddit case study, arXiv:1703.10570(2017).
[56] F. Vis, Twitter as a reporting tool for breaking news: journalists tweeting the 2011 UK riots, Digit. J. 1 (1) (2013) 27–47.
[57] B.D. Eugenio, M. Glass, The kappa statistic: a second look, Comput. Linguist. 30 (1) (2004) 95–101.
[58] J. Carletta, Assessing agreement on classification tasks: the Kappa statistic, Comput. Linguist. 22 (2) (1996) 249–254.
[59] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford 1(12) (2009).
[60] Z. Wei, W. Gao, Utilizing microblogs for automatic news highlights extraction, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 872–883.
[61] L. Hong, O. Dan, B.D. Davison, Predicting popular messages in twitter, in: Proceedings of the 20th International Conference Companion on World wide Web, ACM, 2011, pp. 57–58.
[62] H. Becker, M. Naaman, L. Gravano, Beyond trending topics: real-world event identification on twitter, in: Proceedings of the ICWSM, 2011, pp. 438–441.
[63] E. Bakshy, J.M. Hofman, W.A. Mason, D.J. Watts, Everyone's an influencer: quantifying influence on twitter, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, 2011, pp. 65–74.
[64] P.H.C. Guerra, W. Meira Jr, C. Cardie, R. Kleinberg, A measure of polarization on social media networks based on community boundaries., in: Proceedings of the ICWSM, 2013.
[65] Y. Gong, X. Liu, Generic text summarization using relevance measure and latent semantic analysis, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, ACM, New York, NY, USA, 2001, pp. 19–25.
[66] G. Erkan, D.R. Radev, Lexrank: graph-based lexical centrality as salience in text summarization, J. Artif. Int. Res. 22 (1) (2004) 457–479.
[67] P. Khosla, M. Basu, K. Ghosh, S. Ghosh, Microblog retrieval for post-disaster relief: applying and comparing neural ir models, arXiv:1707.06112(2017).

Shalini Priya is a research scholar student in the department of Computer Science and Engineering at Indian Institute of Technology, Patna. She has completed her M.tech from NIT Durgapur, West Bengal. Her current research interest includes Online Social Network, Data mining and Information Retrieval.

Ryan Sequeira was born on August 16th, 1990 in Mumbai, India. He received his B.E, in the field of Information Technology, from the University of Pune in 2012. He completed his M.Tech, in the field of Computer Science at Indian Institute of Technology Patna where he was awarded the Institute Silver Medal for securing the highest CPI in his department. His research interests lie in the area of complex networks and currently his work has focused on studying the prevalence of prescription drug abuse content on social networks.

Joydeep Chandra is an Assistant Professor in the Department of Computer Science and Engineering at Indian Institute of Technology, Patna, India. He received his Ph.D. from Indian Institute of Technology, Kharagpur, India in 2012. He was also a research fellow at the Chair of Systems Design at ETH Zurich. His research interest includes mining social media, modeling of social networks, studying diffusion of information and identifying influentials. Application domains include Journalism, Disaster, Healthcare and Crimes on the Web.

Sourav Kumar Dandapat is an assistant Professor in the department of Computer Science and Engineering of IIT Patna, India. He has completed his Ph.D. in 2015, from Computer Sc. and Eng. department of IIT Kharagpur, India. His current research interest includes Online Social Network, Mobile Computing, Human Computer Interaction.