Predicting Tomorrow's Headline using Twitter Deliberations

Roshni Chakraborty
IIT Patna
India
roshni.pcs15@iitp.ac.in

Abhijeet Kharat
IIT Patna
India
abhijeet.mtcs17@iitp.ac.in

Apalak Khatua
XLRI Jamshedpur
India
apalak@xlri.ac.in

Sourav Kumar Dandapat
IIT Patna,
India
sourav@iitp.ac.in

Joydeep Chandra
IIT Patna,
India
joydeep@iitp.ac.in

Abstract

Predicting the popularity of a news article is a challenging task. Existing literature mostly focused on article contents and polarity to predict the popularity. However, existing research has not considered the users preference towards a particular article. Understanding users preference is an important aspect for predicting the popularity of news articles. Hence, we consider social media data, from the Twitter platform, to address this research gap. In our proposed model, we have considered the users involvement as well as the users reaction towards an article to predict the popularity of the article. In short, we are predicting tomorrows headline by probing todays Twitter discussion. We have considered 300 political news articles from the *New York Post*, and our proposed approach has outperformed other baseline models.

1 Introduction

Gone are those days when an office going New Yorker used to board the subway with a folded newspaper in his hand. Reading morning newspapers on New Yorks subway is becoming outdated. Things have changed drastically in recent times. Todays millennial generation is not only emotionally but also physically tied to their smartphones and tablets. This has severely affected the newspaper industry. All leading newspapers across the globe have reported a sharp drop in their print circulation. So, the future of this industry lies on the digital platform. The competition in this newspaper industry is not anymore about sending the print version to the remotest corner of the country. The challenge of this digital platform is to understand the latent psychological aspects of the users. If a newspaper fails to satisfy the user, then within the next few seconds she will switch to another news-related app. This will directly impact the ad revenue of a news outlet. Between various news related apps, and various social media platforms, users these days are spoilt for choice. Customer loyalty is a concept of a bygone era in this digital age, and the customers preferences are also not homogeneous. In brief, the phenomenal growth of online news consumption and innumerable news sources has significantly increased the competition among news media outlets. Further, the continuous influx of newsworthy events further aggravates the situation. Thus, for media outlets, the need of the hour is to develop an automated system that can help them to predict which of the todays headlines will maintain its popularity tomorrow.

Existing literature has attempted to address this. However, this stream of research mostly explored various features and contents of the articles and the title of the articles [FVC15, LWZ+17]. Prior studies considered the subjectivity and polarity of contents [FVC15, KWHR16], the sentiment of the head-

line [RBdM+15], and so on. In other words, these works focus broadly into the articulation aspects of an article. Few studies also considered the importance of an event to predict the popularity of news article [SAMA17]. One of the major shortcomings of the above approach is that hypothetically, two articles might have similar feature and polarity, but the reaction of readers might be different. It would be similar to comparing an apple to an orange even though they might be nearly similar in shape and weight. We argue that probing the social media platform can hint which is orange, and which is apple. Social media platforms can hint about the users preference.

Nowadays social media platforms, such as Twitter, generate an enormous amount of user-generated data, and many times this social media platforms become the mirror of the society. Existing literature has successfully explored the Twitter data to predict the election outcome [KKGC15], to understand social movements [KK16a], to tackle disasters and epidemic outbreaks [KK16b]. Therefore, we argue that understanding the finer nuances of Twitter deliberation can be beneficial to predict the popularity of news article on digital platforms. This paper attempts to address this research gap.

Prior studies noted that analyzing social media platform could shed light regarding the popularity [KHGPS16] and the life cycle of various news article [Cas13]. However, these studies have considered tweets, which have exclusively mentioned the URL of news article. In fact, these studies failed to probe the richness of the Twitter platform by restricting them to a very small sub-sample of tweets with the URL of a specific news article. On the contrary, our study takes a more holistic approach than these studies and considers both users involvement with a news article and user reaction towards a specific news article. However, the biggest challenge for this approach is to identify the relevant tweets for a specific news article. Therefore, we have developed an iterative and adaptive algorithm that considers both textual and semantic attributes to identify the relevant tweets. Our user involvement aspect considers various count measures, such as a total number of tweets and average number of retweets, count of hashtags, the cumulative number of unique users as well as influential users, and so on. Also, our user reaction indices consider linguistic aspects of the Twitter discussion, such as variances in sentiment and emotion for a particular news article. For the sake of robustness, we have considered various machine learning algorithms and an exhaustive set of baseline models based on prior studies. Our findings on the basis of 300 news items strongly suggest that patterns of Twitter deliberations can outperform other baseline models in predicting the popularity of the news articles.

2 Related Works

Predicting the popularity of news article is a well-researched area. However, the genesis of this research lies in the prior works on news recommender system. News recommender system research mostly focused on the personal preference of an individual user [LXG+14]. So, understanding the user-level latent political leaning, or bias towards a certain sport or a team, can help to predict the suitable news article for an individual user. For instance, prior studies considered users historical preferences [WLC+10], social network data [DFMGL12, AGHT11], user feedback [SBZ11] or combination of both user preferences and user feedback [LWL+11, LCLS10, MGÁRLGMM13], but this stream of studies struggled due to lack of adequate data. Moreover, the users interest can vary over time, and historical data is not available for new users. Thus, prior studies have attempted to mitigate the challenges by considering opinions of social influencers [LXG+14], topic or temporal [XXLZ14] relationships between news items and users [LL13] or analyzing user communities [ZLHL13]. These approaches yield better results in comparison to initial studies, but still, the accuracy of filtering news articles for a newspaper is not satisfactory.

In comparison to news recommender system for an individual user, predicting the popularity of a news article is a complex task because an efficient prediction model needs to account for the heterogeneity of users. More importantly, the summation of individual user preference would not be the proxy for societal acceptance For instance, it is easy to predict what a Democrat or Republican will prefer to read on the digital platform but the task becomes complex if we try to predict what political news will engage both Democrats and Republicans. So, a dominant stream of prior works focused on the content of the news article and article headline and employed machine learning algorithms [SAM+16] to predict the popularity of a news article. For instance, existing literature considered different features of an article [VCLDD17, KWHR16], such as textual [FVC15] and temporal features, to predict its popularity. Prior studies noted that article content features, such as the length of the article, the time of publishing, category or genre of the article, the author of the article and so on, can predict the popularity of a news article [LWZ+17]. Another set of studies also considered the linguistic attributes of an article [KMJO16, KYS+17] or the presence of important entities within an article [SS16] to investigate the issue. Existing literature also noted that the headline of an article itself [KFKN15] and the polarity within the headline [RBdM+15] could be important input variables for the predicting the popularity of a

news article.

Another set of works highlighted the event importance to predict the popularity of a news article. For instance, Setty et al. [SAMA17] ranked news articles by linking them to a chain of recent news events. Similarly, other studies tried to explore the event importance by combining articles using topic similarity from Wikipedia [MB16] or by considering the causal relationships [KVW14]. However, this approach has limitations for new upcoming events or for an event which is losing relevance among readers. In these scenarios, we argue that probing users behavioral pattern on social media platform can hint about the popularity of news article.

To the best of our knowledge, there is hardly any study which considered social media platform for predicting the popularity of news article. Some of the prior studies considered the users behavior on a news media outlet and argued that engagement of users could predict the popularity of a news article [TADAF14, TLA$^+$11]. However, it is worth noting that the news media outlet represents a minuscule of digital platform readers. This is one potential research gap in the existing literature. Popular social media platforms, such as Twitter, not only provides users an option to share their views but also allows to reply or endorse the views of others by retweeting. In other words, Twitter provides a platform for its users to engage in a deliberation. Interaction of users on the Twitter platform can shed light about users engagement with a particular news article [OCDA15]. Thus, this paper attempts to predict the popularity of news article using Twitter data. Some of the earlier works consider initial twitter reactions [MTR14, CEHPS14] or content and structural features [LZZ15]. However, as we mentioned, they have only considered tweets that has news related URLs. Thus, none of the prior studies considered the richness of Twitter data. So, this paper not only considers the user-level involvement (by using count measures of tweets, retweets, number of unique users and other parameters) but also probes user-level reaction towards a particular news article (by understanding the various linguistic aspects of their tweets).

3 Data Collection

To address our research problem, we have considered the front-page political news of the New York Post, which is one of the most popular newspapers in the United States. It has experienced a whopping 500% growth in the last five years with 331 million page views in March 2018. Predicting the popularity of political news is the most challenging in comparison to other genres of news. For instance, a news article on climate change will uniformly affect all users. Therefore, it is easy to predict the reaction of users. However, the political news might not uniformly engage and affect all users because of their ideological heterogeneity. In this study, we have considered 300 political news, from the New York Post, during the period July 2016 to September 2016.

We have considered the Twitter platform for collecting the social media data. Twitter allows free access to approximately 1% of total tweets (in a random fashion) using the streaming API. To probe our research question, we have considered the tweets related to a particular news article. Extracting the relevant tweets for particular political news is a challenging task. To address this, we have developed an adaptive algorithm that has considered both content (similar keyword mapping) features and context (same hashtag) features of tweets to extract the related tweets of a news article. Following prior studies [CBDC17], as an initial step, we have considered a set of preliminary hashtags that have threshold keywords overlap with the representative (by top 10 TF-IDF) keywords within the news article. In other words, these preliminary hashtags, which we have initially considered to crawl tweets for a particular news article, are a bag of hashtags on the basis of the articles seed tweets [CBDC17]. However, there are certain limitations to this approach because users can use multiple hashtags for a particular news article on the social media platform but not all of them might be unique to that particular news article. For instance, social media users have used multiple hashtags for the following news article titled GOP blasts Obama 400 million dollars *secret ransom* paid to Iran as follows: *#whitehouse*, *#trump*2016, *#chicago*, *#irandeal*, *#obamabetrayus* and so on (as shown in Table 3). The last two hashtags are more specific about the news article in comparison to others. Hence, we need to consider this in our data collection as well as analysis.

To address the above concern, we have collected all hashtags related to all political news articles published in the previous one month (with respect to the publication date of the article we are considering in our analysis). This process has generated a bag of hashtags. From this bag of hashtags, we have identified a set of hashtags, which were frequently used by Twitter users and labeled them as generic hashtags. Consequently, we have labeled *#whitehouse*, *#trump*2016, *#chicago* as generic hashtags for the above article because these hashtags were used by Twitter users for other issues/news article also.

Next, we have developed an automated system for identifying hashtags specific to a news article. We have filtered out the preliminary hashtags as the hashtags those were mentioned in a tweet T, and ful-

Table 1: The table shows title of 4 news articles, sample tweets related to the news article and the hashtags(both generic(H_G) and article specific(H_A) related to the news article.

SNo	News Article	Sample Tweets	Hashtags
1	GOP blasts Obama's 400 million dollars *secret ransom* paid to Iran	1. It's not like Obama ever earned any money ... he gave 400 million in cash to Iran.... *#irandeal* 2. I strongly oppose the Raskin-supported foreign policy toward *#iran*. we must not pay ransom to a dangerous terror regime.	*#whitehouse* (H_G),#irandeal(H_A), *#pressecretary*(H_G), *#obamabetrayus*(H_A), *#trump*2016(H_G), *#chicago*(H_G), *#cnn*(H_G)
2	Melania Trump: I have never lived in the US illegally	1. What visa enabled melania trump to work in the U.S.? *#theplotthickens*, *#immigration*, *#trump* 2. If Melania Trump broke immigration laws, the best punishment is ... *#melaniaImmigration #nevertrump*	*#whitehouse* (G), *#irandeal*(H_A), *#pressecretary*(H_G), *#obamabetrayus*(H_A), *#trump*2016(H_G), *#chicago*(H_G), *#cnn*(H_G)
3	Hillary to blame for Iranian scientist's hanging, general says	1. Hillary Clinton reckless emails outed an Iranian nuclear scientist who was executed by Iran for treason *#neverhillary* 2. *#crookedhillary* server has emails discussing nuclear scientist *#executed* by iran *#shortcircuit*	*#crookedhillary*(H_G), *#neverhillary*(H_G), *#shortcircuit*(H_G), *#hillary*(H_G)
4	Obamacare hikes has families struggling to afford insurance	1. Shhh...You're not supposed2 know about health insurance rate hikes until after elections! vote *#trump #repealobamacare* 2. *#repealobamacare* Obama will take our money	*#repealobamacare*(H_A), *#trump*(H_G), *#maga*(H_G)

filled the threshold criteria of keywords matching with the news article [CBDC17]. We define article-specific hashtags as those hashtags that were mentioned in T but not in our list of generic hashtags. For instance, Thus, the article 1 (as shown in Table 3), we have labeled *#irandeal* and *#obamabetrayus* as article-specific hashtags. Similarly, the article-specific hashtag for the news article 4 (as shown in Table 3), i.e., Obamacare hikes has families struggling to afford insurance was *#repealobamacare*.

To check the accuracy of this approach, we have provided around 130 news articles along with all the hashtags to three annotators. We have labeled a hashtag as an article-specific hashtag if the majority of annotators have marked that particular hashtag as specific to that article, or otherwise labeled it as a generic hashtag. We observed that our proposed approach yields an accuracy of 89%in identifying an article specific hashtags. (as shown in Table 3) reports a few sample news articles and corresponding article-specific (H_A) and generic (H_G) hashtags. After identifying the article specific hashtags, we use these hashtags to extract further tweets related to that news article.

Next, we have extracted the user level information. In other words, we have extracted the information related to users who had participated in the political discussion related to any of these 300 news articles. We have extracted the users name from our Twitter corpus and identified around 1 million unique users who have tweeted at least once for our sample of 300 news article. Next, we have crawled their last 3200 tweets and profile-related information. In our model, we have considered whether a user is influential or not. If a user has more than 1000 followers, then we have considered them as an influential user. To sum up, we have considered 1.8 million tweets for our 300 news articles made by around 1 million unique users.

4 Proposed Approach

For predicting the popularity of a news article, we have considered two categories of social media data namely, user involvement indices and user reaction indices. We argue that the popularity of a news article among the social media users (which is a proxy for digital platform readers) can be captured by analyzing the attention that a news article is receiving and the linguistic content of the discussion on the Twitter platform on the very day of its publication. In brief, the former category considers various user-level tweet statistics, and the latter employs natural language processing techniques to understand the linguistic aspects of the social media discussions. The following sections narrate how we have operationalized the involvement and reaction indices.

Table 2: The table shows 5 News Articles along with the whether it was published next day (P_{n+1}) and User Involvement Indices related to each article. H_P, H_G and H_A represents the number of preliminary hashtags, generic hashtags and article specific hashtags and T_t, T_r and T_f represents the number of tweets, retweets and favourites and u, qu represent the users and unique users of the tweets related to a news article.

SNo	Title of a few sample news article on n^{th} day	P_{n+1}	$(H_P/H_G/H_A)$	$(T_t/T_r/T_f)$	(u/qu)
1	Suicide bombing at Pakistani hospital kills at least 63	Yes	16/8/8	1319/42150/37150	1319/1240
2	Trump to propose big tax breaks in economic plan	Yes	20/12/8	208/44151/37929	208/180
3	Trump gives Post columnist a shout-out in economic speech	No	5/5/0	10/0/0	10/10
4	Obama commutes sentences for record-breaking 214 prisoners	No	13/1/12	13/2/0	13/13
5	Furious GOP leaders plot to get Trump on track	Yes	39/8/31	344/223/ 150	344/289

4.1 User Involvement Indices

We capture the attention of Twitter users for a particular news article by considering the user involvement through three aspects: *tweet statistics*, *user statistics* and *hashtag statistics*. Under the *tweet statistics* category, we have considered the number of tweets, the number of retweets and the number of favorites received by a particular news article on the very day of publication. These three statistics represents the user response towards a particular article. We observe that there is a significant variance in user involvement. Some news article receives hundreds of tweets and retweets whereas another news article merely receives ten to twenty tweets. Thus, we have normalized the number of tweets received by a particular article by dividing it with the maximum number of possible tweets that an article can have in a day.

Intuitively, the number of users get involved with a particular news article is a good predictor of the popularity of the news article. Furthermore, we note that some users get more involved with a particular news article, and they tweet multiple times in a day. However, it is worth noting that 10 tweets from 10 different users, in comparison to 10 tweets from 1 particular user, is a better proxy to gauge the popularity of a news article. On the contrary, if a user tweets, about a particular news article, for more than once, then it also indicates his high involvement of that user with that particular news article. So, we have considered these finer variances in our analysis. We have considered the fraction of users, who have tweeted more than once for a particular news article, as affected users.

Subsequently, it is also important to note whether a user is influential on the social media platform or not. In other words, an influential person can be an opinion leader on a social media platform. For instance, if personalities, such as Barack Obama or Donald Trump, endorse a particular news article through their personal/official twitter handle then immediately that tweet will get retweeted by hundreds of their followers. more than 1000 followers. To sum up, in addition to generic user statistics we have also considered the fraction of affected users and influential users for each news article as an input variable in our model.

Next, we considered the number of article-specific hashtags on the Twitter platform as a metric to gauge the involvement of social media users. Intuitively, it can be argued that higher user involvement with a news article would generate higher number of article-specific hashtags. For instance, the news article Trump gives Post Columnist a shout-out in economic speech didnt generate article specific hashtags. On the contrary, the news article Trump to propose big tax breaks in economic plan has created 8 article specific hashtags (as shown in Table 2). Thus, we have considered the total number of article-specific hashtags as an input variable in our proposed model.

4.2 User Reaction Indices

As we mentioned earlier, natural language processing (NLP) techniques allowed us to go beyond various count-based user-level measures and to probe the linguistic content of Twitter deliberations to understand the cognitive involvement of users with a particular news article. This cognitive involvement of users can be a good proxy to predict the popularity of a news article. So, we have employed NLP techniques, such as sentiment and emotion analysis, to gauge the user reactions towards a particular news article. We have considered three indicators to capture the user reaction namely, sentiment variance, emotion variance and argumentativeness index.

We argue that differences of opinion would lead to higher debates and discussion on the Twitter platform. For instance, most social media users would agree with a news article such as Global warming would be a se-

Table 3: The table shows 5 news articles along with whether the article was published next day(P_{n+1}), sentiment variance(SV) and emotion variance(EV) of the tweets related to each news article.

SNo	Title of a few sample news article on the nth day	P_{n+1}	SV	EV
1	Suicide bombing at Pakistani hospital kills at least 63	Yes	0.67	43.19
2	Trump to propose big tax breaks in economic plan	Yes	0.90	20.30
3	Trump gives Post columnist a shout-out in economic speech	No	0.00	0.00
4	Obama commutes sentences for record-breaking 214 prisoners	No	0.00	0.00
5	Furious GOP leaders plot to get Trump on track	Yes	0.37	41.51

rious threat in the coming decades and it might create a discussion but not debates. However, a hypothetical news article such as President Trump is failing to take appropriate policy measures to control global warming would probably lead to a debate between the Democrats and Republicans. Republican will try to discard this view, whether Democrats will try to justify this view. Consequently, the popularity of this particular news article will also go up. We are attempting to capture this in our proposed model.

Following prior studies, such as Vader Sentiment Analyzer [HG14] and TextBlob [LKH+14], we have calculated the average sentiment score of a tweet. We have identified all tweets specific to a particular news article and classified whether the tweet is positive or negative. Next, we have considered the sentiment variance of all tweets related to a particular news article to understand the differences in opinions. We have calculated the sentiment variation (SV) as follows:

$$SV = 1 - \frac{\mid (PC - NC) \mid}{\mid (PC + NC) \mid}$$

PC is the number of positive tweets for a news article, and NC is the number of negative tweets for the same news article. The sentiment variance is highest when the count of positive tweets and negative tweets are equal for a news article, and the sentiment variance decreases when there is only (or higher number of) positive/negative tweets. In other words, having an equal number of positive and negative sentiment indicates that users are from two ideologically opposite camps. On the contrary, only positive or negative tweets indicate that users are ideologically homogeneous.

Next, we also considered the emotional content of a tweet. We employ the NRC emotion lexicon[MT10, MT13] to classify a tweet among various emotion classes such as anger, anticipation, trust, disgust, fear, joy and surprise. Similar to our sentiment variance analysis, we have considered all tweets specific to a particular news article and classify them into various categories of emotions. Intuitively, high emotional variance indicates that users are displaying different emotions towards a news article. We have calculated the emotional variance (EV) as follows:

$$EV = \frac{\sum_{i=1}^{8}(e(i) - m(e))^2}{n}$$

In the above formula, n is the number of emotion categories which is 8 [MT10, MT13], $e(i)$ is the fraction of tweets with $\mathrm{i}^{th} emotion, and the value of \mathrm{m(e)} is \mathrm{N}_{\overline{8}}$ where N is the total number of tweets related to a particular article. Here, the highest emotion variance indicates that tweet corpus for a particular news article represents multiple emotion categories. For instance, in response to the immigration issue related news, a Republican, who believes that strong immigration law would protect American jobs, might display joy. On the contrary, a social activist, who thinks otherwise, might display her anger to the same news article.

5 Data Analysis

5.1 Preparation of Gold Standard

For our analysis, we need to know whether a particular news article of the n^{th} day is followed by another subsequent article on $(n+1)^{th}$ day. It is important to note that on (n+1) the day the title or the content of the subsequent article can differ significantly from the previous day. For instance, on nth day, the hypothetical title of a news article can be: Why Brexit matters for the American Corporate Sector? However, on the $(n+1)^{th}$ day the issue will continue, but the title can be: American Corporates are reluctant to invest in the UK. So, it requires a contextual understanding to prepare the database for our analysis. Thus, we employed three annotators and provided them with a particular news article of nth day and all the news articles of $(n+1)^{th}$ day for manual annotation. We have asked our annotators to mark a news article either as 1 if the same news gets covered on the subsequent or $(n+1)^{th}$ day and 0 otherwise. For our analysis purpose, we have considered the labeling on the basis of the majority of the annotators. We have done this for all 300 news articles that we considered for our final analysis.

Table 4: The table shows the baseline models along with the features considered in each of the baseline models

SNo	Baseline Models	List of Features/Input Variables
1	Article Content +Article Polarity	no of words in the article; the rate of non-stop words; day of the week on which it got published; published on weekend or not;no of entities in the news article; average word length of the article
2	Article Polarity	Polarity score of the article; the rate of positive and negative words per 100 words; the rate of positive and negative words per 100 words with non-neutral words, the average polarity of positive and negative words; min. and max. the polarity of positive and negative words
3	Title Content +Title Polarity	no of words in the title; the rate of non-stop words in the title;no of entities in the title; the average word length of the title
4	Title Polarity	Polarity score of the title; the rate of positive and negative words in the title; the rate of positive and negative words with non-neutral words in the title; average polarity of positive and negative words in the title; min. and max. the polarity of positive and negative words in the title
5	Event importance	no of days a news article related to the event was published, no of articles of the event was published

Table 5: The comparison of the proposed approach with the baselines

	RFC	SVM	CART
Proposed Approach			
Precision	91.4	94.6	83.3
Recall	84.2	83.3	85.7
F1-Score	87.6	88.6	84.9
F1-Score of Baseline Models			
Article Content+Article Polarity	86.8	86.9	74.9
Article Polarity	86.8	84.14	81.7
Title Content + Title Polarity	86.0	87.8	82.5
Title Polarity	85.3	84.9	83.5
Event Importance	81.5	87.4	84

5.2 Baseline Models

As we discussed in our literature review, a plethora of studies have tried to predict the popularity of news article [KMJO16, KFKN15, KYS$^+$17, RBdM$^+$15, SS16]. However, this stream of literature is broadly classified into five categories as follows: article content and polarity, title content and polarity, and event importance categories(as shown in Table 4). We have considered all these five prediction models as our baseline models.

Following the prior studies [KFKN15, RBdM$^+$15, KVW14], we have extracted both the content and polarity of the article to predict whether the article will get published on the next day or not. Here, we employed NLP techniques to the understand the overall sentiment of the article, usages of positive or negative words within the article, the length of the article, and so on. Similarly, we have also considered the content and polarity of the title of the article to predict its popularity. Here, our analysis is restricted only to the title of the article. Prior studies portray certain differences in terms of the number of features. However, in our studies, we have tried to consider an exhaustive set of features for first four baseline models: article content and polarity, title content and polarity. The final baseline model is the event importance [SAMA17]. The event importance tries to capture the dominance of a topic/issue in comparison to others. So, the event importance of a news article is calculated by the numbers of similar articles that get published on consecutive days. A pair of news article will be considered as a similar article if it crosses the threshold among the list of entities and bi-grams features between two articles. Prior studies noted that event importance is also an indicator to predict the popularity of a news article.

6 Results and Discussions

We have employed Random Forest Classified (RFC), Support Vector Machine (SVM), Gradient Boosting Classifier (GBC), and Classification and Regression Trees (CART) algorithms for our analysis. We have applied these four classifiers on our article dataset both for the five baseline models as well as our proposed models. We have considered ten-fold cross-validation for our analysis. We have repeated our experiments multiple times and found our results are consistent. We have reported the same in Table 5. Our proposed model has outperformed all five baseline models for all three classifiers. Our F1-score for SVM and RFC classifiers are marginally better than the CART classifier. Broadly, the SVM classifier has outperformed other classifiers not only for our proposed model but also for baseline models.

7 Conclusions

The advent of information and communication technology has affected the newspaper industry severely in last few decades. Seamlessly connected various communication channels are generating a huge volume of information. Moreover, the digital platform is becoming a crowded place. Multiple news outlets are struggling to grab a larger share of this platform. Therefore, selecting a potentially popular news article is becoming a daunting task for the journalists. This leads to the requirement of an automated system, which can efficiently select the news article that will most likely draw the maximum attention of users on the digital platform. To address this, prior studies focused mostly on the content and polarity of the news article to predict the popularity of news articles. However, these studies failed to capture the latent psychological aspects of users. Thus, our proposed approach is trying to gauge the users perception from the social media discussions. We have consideredthe Twitter platform for our study. Our proposed model has incorporated users involvement and reaction towards a particular news article. In short, as our title suggests that we are trying to predict tomorrows popular headline by considering todays discussion on Twitter platform.We have employed various machine-learning algorithms to test the accuracy of our proposed approach. We observe that our proposed approach ensures higher accuracyin comparison to other baseline models. Considering Twitter discussion for predicting the popularity of news article is the core contribution of this study. However, there are certain shortcomings of our proposed mode which future research needs to address. Firstly, we have considered a small sample of 300-news article for a relatively shorter period. Future studies in this area should consider a larger sample and longer time horizon. Secondly, we have considered the political news of the New York Post. In other words, we have tested our proposed approach in the political sphere of the United States. So, future studies need to probe the efficacy of our model for other genres of news in other countries. The biggest challenge will be to extrapolate this approach to a context where native language is not English. Finally, we have considered a few fundamentalmachine-learning algorithms. Future studies need to consider advanced deep learning based models to see the accuracy of our model.

References

[AGHT11] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. *User Modeling, Adaption and Personalization*, pages 1–12, 2011.

[Cas13] Carlos Castillo. Traffic prediction and discovery of news via news crowds. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 853–854. ACM, 2013.

[CBDC17] Roshni Chakraborty, Maitry Bhavsar, Sourav Dandapat, and Joydeep Chandra. A network based stratification approach for summarizing relevant comment tweets of news articles. In *International Conference on Web Information Systems Engineering*, pages 33–48. Springer, 2017.

[CEHPS14] Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 211–223. ACM, 2014.

[DFMGL12] Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 153–162. ACM, 2012.

[FVC15] Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez. A proactive intelligent decision support system for predicting the popularity of online news. In *Portuguese Conference on Artificial Intelligence*, pages 535–546. Springer, 2015.

[HG14] Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[KFKN15] Sawa Kourogi, Hiroyuki Fujishiro, Akisato Kimura, and Hitoshi Nishikawa. Identifying attractive news headlines for social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1859–1862. ACM, 2015.

[KHGPS16] Danielle K Kilgo, Summer Harlow, Víctor García-Perdomo, and Ramón Salaverría. A new sensation? an international exploration of sensationalism and social media recommendations in online news publications. *Journalism*, page 1464884916683549, 2016.

[KK16a] Apalak Khatua and Aparup Khatua. Leave or remain? deciphering brexit deliberations on twitter. In *2016 IEEE 16th international conference on data mining workshops (ICDMW)*, pages 428–433. IEEE, 2016.

[KK16b] Aparup Khatua and Apalak Khatua. Immediate and long-term effects of 2016 zika outbreak: a twitter-based study. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6. IEEE, 2016.

[KKGC15] Aparup Khatua, Apalak Khatua, Kuntal Ghosh, and Nabendu Chaki. Can# twitter_trends predict election results? evidence from 2014 indian general election. In *2015 48th Hawaii international conference on system sciences*, pages 1676–1685. IEEE, 2015.

[KMJO16] Joon Hee Kim, Amin Mantrach, Alejandro Jaimes, and Alice Oh. How to compete online for news audience: Modeling words that attract clicks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1645–1654. ACM, 2016.

[KVW14] Erdal Kuzey, Jilles Vreeken, and Gerhard Weikum. A fresh look on knowledge bases: Distilling named events from news. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1689–1698. ACM, 2014.

[KWHR16] Yaser Keneshloo, Shuguang Wang, Eui-Hong Han, and Naren Ramakrishnan. Predicting the popularity of news articles. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 441–449. SIAM, 2016.

[KYS+17] Nagendra Kumar, Anusha Yadandla, K Suryamukhi, Neha Ranabothu, Sravani Boya, and Manish Singh. Arousal prediction of news articles in social media. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 308–319. Springer, 2017.

[LCLS10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[LKH+14] Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*, 2014.

[LL13] Lei Li and Tao Li. News recommendation via hypergraph learning: encapsulation of user behavior and news content. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 305–314. ACM, 2013.

[LWL+11] Lei Li, Dingding Wang, Tao Li, Daniel Knox, and Balaji Padmanabhan. Scene: a scalable two-stage personalized news recommendation system. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 125–134. ACM, 2011.

[LWZ+17] Caiyun Liu, Wenjie Wang, Yuqing Zhang, Ying Dong, Fannv He, and Chensi Wu. Predicting the popularity of online news based on multivariate analysis. In *Computer and Information Technology (CIT), 2017 IEEE International Conference on*, pages 9–15. IEEE, 2017.

[LXG+14] Chen Lin, Runquan Xie, Xinjun Guan, Lei Li, and Tao Li. Personalized news recommendation via implicit social experts. *Information Sciences*, 254:1–18, 2014.

[LZZ15] Qian Liu, Mi Zhou, and Xin Zhao. Understanding news 2.0: A framework for explaining the number of comments from readers on online news. *Information & Management*, 52(7):764–776, 2015.

[MB16] Arunav Mishra and Klaus Berberich. Leveraging semantic annotations to link wikipedia and news archives. In *European Conference on Information Retrieval*, pages 30–42. Springer, 2016.

[MGÁRLGMM13] Alejandro Montes-García, Jose María Álvarez-Rodríguez, Jose Emilio Labra-Gayo, and Marcos Martínez-Merino. Towards a journalist-based news recommendation system: The wesomender approach. *Expert Systems with Applications*, 40(17):6735–6741, 2013.

[MT10] Saif M Mohammad and Peter D Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[MT13] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[MTR14] Nuno Moniz, Luís Torgo, and F Rodrigues. Improvement of news ranking through importance prediction. In *Proc. KDD Workshop on Data Science for News Publishing (NewsKDD)*, page 6, 2014.

[OCDA15] Alexandra Olteanu, Carlos Castillo, Nicholas Diakopoulos, and Karl Aberer. Comparing events coverage in online news and social media: The case of climate change. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, number EPFL-CONF-211214, 2015.

[RBdM+15] Julio Reis, Fabricio Benevenuto, P Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. Breaking the news: First impressions matter on online news. In *ICWSM15: Proceedings of The International Conference on Weblogs and Social Media*, 2015.

[SAM+16] R Shreyas, DM Akshata, BS Mahanand, B Shagun, and CM Abhishek. Predicting popularity of online articles using random forest regression. In *Cognitive Computing and Information Processing (CCIP), 2016 Second International Conference on*, pages 1–5. IEEE, 2016.

[SAMA17] Vinay Setty, Abhijit Anand, Arunav Mishra, and Avishek Anand. Modeling event importance for ranking daily news events. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 231–240. ACM, 2017.

[SBZ11] Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. Personalised rating prediction for new users using latent factor models. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, pages 47–56. ACM, 2011.

[SS16] Pedro Saleiro and Carlos Soares. Learning from the news: Predicting entity popularity on twitter. In *International Symposium on Intelligent Data Analysis*, pages 171–182. Springer, 2016.

[TADAF14] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias De Amorim, and Serge Fdida. From popularity prediction to ranking online news. *Social Network Analysis and Mining*, 4(1):174, 2014.

[TLA+11] Alexandru Tatar, Jérémie Leguay, Panayotis Antoniadis, Arnaud Limbourg, Marcelo Dias de Amorim, and Serge Fdida. Predicting the popularity of online articles based on user comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, page 67. ACM, 2011.

[VCLDD17] Steven Van Canneyt, Philip Leroux, Bart Dhoedt, and Thomas Demeester. Modeling and predicting the popularity of online news based on temporal and content-related features. *Multimedia Tools and Applications*, pages 1–28, 2017.

[WLC+10] Jia Wang, Qing Li, Yuanzhu Peter Chen, Jiafen Liu, Chen Zhang, and Zhangxi Lin. News recommendation in forum-based social media. In *AAAI*, 2010.

[XXLZ14] Zhengyou Xia, Shengwu Xu, Ningzhong Liu, and Zhengkang Zhao. Hot news recommendation system from heterogeneous websites based on bayesian model. *The Scientific World Journal*, 2014, 2014.

[ZLHL13] Li Zheng, Lei Li, Wenxing Hong, and Tao Li. Penetrate: Personalized news recommendation using ensemble hierarchical clustering. *Expert Systems with Applications*, 40(6):2127–2136, 2013.