CbI: Improving Credibility of User-Generated Content on Facebook

Sonu Gupta[1], Shelly Sachdeva[2], Prateek Dewan[3], and Ponnurangam Kumaraguru[3]

[1] Jaypee Institute of Information Technology, Noida
gupta.sonu1607@gmail.com
[2] National Institute of Technology, Delhi
shellysachdeva@nitdelhi.ac.in
[3] Indraprastha Institute of Information Technology, Delhi
{prateekd, pk}@iiitd.ac.in

Abstract. Online Social Networks (OSNs) have become a popular platform to share information with each other. Fake news often spread rapidly in OSNs especially during news-making events, e.g. Earthquake in Chile (2010) and Hurricane Sandy in the USA (2012). A potential solution is to use machine learning techniques to assess the credibility of a post automatically, i.e. whether a person would consider the post believable or trustworthy. In this paper, we provide a fine-grained definition of credibility. We call a post to be credible if it is *accurate, clear, and timely*. Hence, we propose a system which calculates the *Accuracy, Clarity, and Timeliness* (A-C-T) of a Facebook post which in turn are used to rank the post for its credibility. We experiment with 1,056 posts created by 107 *pages* that claim to belong to news-category. We use a set of 152 features to train classification models each for A-C-T using supervised algorithms. We use the best performing features and models to develop a RESTful API and a Chrome browser extension to rank posts for its credibility in real-time. The random forest algorithm performed the best and achieved ROC AUC of 0.916, 0.875, and 0.851 for A-C-T respectively.

Keywords: Online Social Media · Facebook · Credibility.

1 Introduction

With the advent of time, OSNs have replaced traditional media like print media and television as a source of information about the latest happenings around the globe. Instant updates and easy sharing nature of OSNs have contributed to this shift. They have also become a go-to resource for journalists during news gathering. Facebook is the most popular social networking site (SNS) with 2.2 billion monthly active users on average as of January 2018.[4] Given the increasing popularity, it has emerged as a news source and as a medium to disseminate information. Therefore, OSNs witness an upsurge in user activity whenever a high impact event takes place. Users log-on to Facebook and other SNSs to

[4] https://en.wikipedia.org/wiki/Facebook

check for updates, to share posts and their opinions on these events. Albeit a vast volume of content is posted on OSNs, not all of the information is accurate and reliable. Some users intentionally post fake news while other share such posts without verifying its content. The effect of such rumors can be highly misleading and can cause panic among people. Credibility on an OSN is a matter of great concern as information spreads quickly here. Figure 1 shows an example of a fake Facebook post.[5] During the 2016 US presidential election, a satirical news website asserted that Francis endorsed Trump for president. The story was almost entirely fabricated but picked over 960,000 Facebook engagements.

Fig. 1. An example of a fake Facebook post.

On Facebook, pages are more popular than user profiles. Generally, celebrities, businesses, and organizations create Facebook pages to connect with everyone. User profiles and pages follow other pages of their interests. Various news channels have pages and keep updating latest news. There are no restrictions on the number of followers a page can have while a user profile can have a maximum 5000 friends. So, a page enjoys a broader audience than a user profile. Thus, pages are the best medium to spread any information quickly. Also, it has been observed that user-profile owners often post their opinions only and to post about news-related events they tend to share posts created by pages instead of writing one by themselves. So, for the purpose of this study, we focus on the posts created by pages.

Many researchers have studied the credibility of information on Twitter. There are a few real-time systems to detect misinformation in Twitter. But there is only a little research on the credibility of user-generated content on Facebook. Detecting misinformation on the Facebook faces more challenges than Twitter. Unlike Twitter which provides both streaming API and search API, Facebook

[5] https://www.cnbc.com/2016/12/30/read-all-about-it-the-biggest-fake-news-stories-of-2016.html

provides only Graph API.[6] Using Graph API, we cannot search for posts using a keyword. We can fetch a post directly via Graph API only if it is a public post and its post ID is known. In Twitter, we can get at most 3200 tweets for a particular user using the Twitter API, but Facebook does not provide any way to fetch posts for a specific user.[7] Facebook also restricts the information access to great extends due to its various privacy policies. Facebook has a more complex structure due to its diverse features like pages, events etc. which are not present on Twitter. Thus, the techniques used for credibility assessment on Twitter cannot be directly applied on Facebook.

Haralabopoulos et al. [10] assert that if an OSN user has a strong position in the network, it is expected for an inaccurate and poorly timed post to have a modest impact in the network. Therefore, the authors claim that *Accuracy, Clarity, and Timeliness* are the three critical parameters to define information quality on OSN. Accuracy is the condition of validating the information in the sense of being true, correct, reliable, precise, and free of errors. Clarity implies the clearness, ease of consumption, readability, and absence of ambiguity. Timeliness is used to describe up-to-date, seasonable, or well-timed information, which is a crucial factor in most online breaking news services and social networks. Following up [10], we define credibility as a function of *Accuracy, Clarity, and Timeliness*. The *pages* have a strong influence in the network given the huge number of followers. Hence, we call a post made by a *page* credible if it is accurate, clear and timely. In this paper, we propose a novel technique which ranks the posts created by *pages* for its credibility. It classifies each post on the basis of A-C-T. These classification results are used to rank the posts from 1-4. To the best of our knowledge, this is the first work which ranks the Facebook posts for its credibility in terms of A-C-T. Our major contributions of this work are:

- a ranking model to assess the credibility of posts in terms of A-C-T.
- an extensive set of 152 features.
- a RESTful API and a browser extension to assess the credibility of posts in real-time.

The rest of the paper is organized as follows: Section 2 describes the related work. In Section 3, we explain the proposed technique and discuss the methodology used for data collection and labeling in Section 4. In section 5, we discuss various features and present the classification results. In section 6, we conclude by highlighting the limitations and future work directions.

2 Related Work

Many researchers have studied the credibility of information on OSNs. There are many solutions based on both computational science and social science. The credibility score of a message on online social media can be computed by using (a) web-page dependent features like share count, likes etc., (b) or by comparing the messages with those of trusted news sources.

[6] https://developers.facebook.com/docs/graph-api
[7] https://dev.twitter.com/overview/api

2.1 Credibility Assessment of Content on Twitter

Out of all OSNs, the credibility of Twitter messages (tweets) is studied the most by the research community. Mendoza et al. characterized Twitter data generated during the 2010 earthquake in Chile [12]. They studied the response of Twitterers in this emergency situation. They observed that fake news spread rapidly and thus create chaos in the absence of real information from traditional sources. Gupta et al. distinguish between fake and real images propagated during Hurricane Sandy on Twitter using decision tree classifier [9]. Castillo et al. showed that automated classification techniques could be used to detect news topics from conversational topics and computed their credibility based on various features [2]. They were able to achieve a precision and recall of 70-80% using J48 decision-tree algorithm. Gupta and Kumaraguru applied machine learning algorithms (SVM-rank) and information retrieval techniques (relevance feedback) to compute the credibility of tweets using message based and source based features [7]. They observed that the dispersion of information differs during crisis and non-crisis events. In [1], Alrubaian et al. proposed a system to compute credibility score on Twitter using five procedures; tweet collecting and repository, credibility scoring technique, reputation scoring technique, user experience measuring technique, and trustworthiness value, the last of which is an output of the preceding three procedures. Due to the absence of network-based and entity-based features on Facebook, the above techniques cannot be applied directly to calculate the credibility of Facebook posts.

Li and Sakamoto showed that displaying both retweet count and collective truthfulness rating minimizes the spread of inaccurate health-related messages on Twitter [11]. They suggested that collecting and displaying the truthfulness rating of crowds in addition to their forwarding decisions can reduce the false information on social media. However, we believe that computing truthfulness rating on the basis of crowd's response can be a victim of a collusive attack by malicious entities and can have adverse effects. In [10], Haralabopoulos et a1. proposed three solutions to address the credibility challenge which includes community-based evaluation and labeling of user-generated content in terms of A-C-T along with real-time data mining techniques. The above solution entirely relies on community-based evaluation. Thus, it fails to generate a reliable credibility score if no/few users evaluated the post. It also depends on the critical thinking of the crowd as mentioned in [15]. So, instead of relying on the crowd, in this work, we use supervised machine learning algorithms to identify A-C-T of a post.

Credibility Assessment Tools: Researchers developed and deployed tools to compute the credibility score on Twitter in real-time. Ratkiewicz et al. created a web service called Truthy that helps in tracking political memes on Twitter [13]. It detected astroturfing, smear campaigns, and misinformation in the context of U.S. political elections. In [8], Gupta et al. presented a semi-supervised ranking model using SVM-rank for calculating credibility score. They developed TweetCred as a browser extension, web application, and a REST API, to calculate real-time credibility scores. Inspired by the TweetCred, we also developed a user-friendly browser extension to rank Facebook posts in real-time.

2.2 Credibility Assessment of Content on Facebook

Saikaew and Noyunsan studied the credibility of information on Facebook [14]. They developed two chrome browser extensions. The first extension was used to measure the credibility of each post by asking users to give a score from 1(the lowest value) to 10 (the highest value). These post evaluations were used to train an SVM model. The second extension was used to automatically evaluate the real-time credibility of each post using the SVM model. Also, the model was trained using only 8 features such as likes, comment counts etc. In order to solve this issue, we train a model with our proposed extensive feature set to rank the Facebook posts for its credibility in real-time using a Google chrome-browser extension. To the best of our knowledge, our work is the first research work which ranks the Facebook posts for its credibility in terms of A-C-T.

3 Credibility Assessment of User-Generated Content on Facebook

We propose a supervised machine learning technique to improve the credibility of user-generated content on Facebook. For this, we obtain ground truth and train classification models on it. The resulting models are used to develop a RESTful API and a browser extension which ranks the Facebook post for its credibility on the basis of A-C-T. Each step is explained in detail in the following sections.

3.1 Proposed Credibility Assessment Model

We define credibility as a function of A-C-T. Therefore, we require three independent classification models, i.e. each for A-C-T. Figure 2 describes the technique proposed in this paper to assess the credibility of a Facebook post. The first step is to collect data using Facebook's Graph API. It returns data in a semi-structured form, and wherefore we store it in a NoSQL database. For our work we use MongoDB. Each document stores data of a post. From all the posts we collected, we randomly sample η posts to curate the ground truth dataset. We host an annotation portal on AWS and with the help of human annotators, we collect ground truth for A-C-T. With the help of previous work and our analysis, we curate a feature set. We use it to train various supervised binary classification models each for A-C-T. We rank each post on the basis of classification result for A-C-T. If a post is accurate, clear and timely, we rank it as *1*. If a post has positive results for any two of A-C-T, it is ranked as *2*, whereas if it has a positive result for anyone, then it is ranked as *3*. A post would be ranked as *4* if it is neither accurate, clear nor timely, i.e. it has negative results for A-C-T.

It is worth mentioning that we experimented with the standard ranking models like SVM-rank[8], but couldn't achieve decent results due to the lack of ground truth data. For the same reason, we were unable to use unsupervised learning models. Therefore, we preferred to use standard supervised machine learning algorithms.

[8] https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

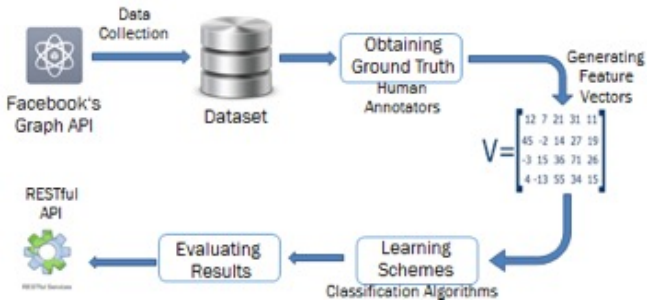

Fig. 2. Describes our proposed technique to assess credibility of user-generated content on Facebook.

3.2 Credibility Assessment Tools

Our aim in this research work is to provide a user-friendly tool which can compute credibility scores for Facebook posts in real-time. Using the best performing models, we develop a RESTful API which receives the input from the trained classification models. On top of the API, we develop a Google Chrome browser extension that displays the results in real-time on user's Facebook news-feed in the form of an alert symbol. In this section, we describe the working of the tools.[9]

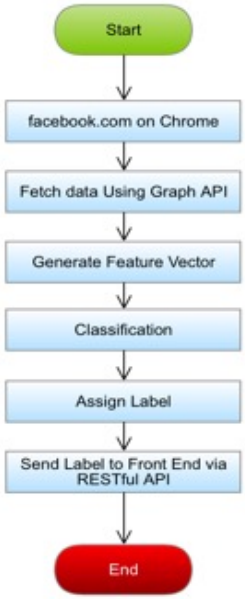

Fig. 3. Data flow steps of the CbI extension and API.

[9] both the tools are in the development stage; hence, they are not available online.

Credibility Investigator API: Figure 3 shows the data flow step of the CbI extension and the API. Credibility Investigator API is a RESTful API written in python using the Flask framework.[10] Due to Facebook's API restrictions, CbI API works only on public posts which are accessible through Facebook's Graph API. The API provides a POST method to submit public post's ID from a user's news feed for analysis. Once a post is submitted to the API, it generates feature vectors which are given as input to our A-C-T pre-trained models. Since the objective is to provide a real-time alert to the Facebook user, we need to minimize the time taken in feature extraction and classification. In order to achieve this goal, we implemented multiprocessing such that features are extracted simultaneously, which helps in saving a lot of processing time. Hence, the output of the models is the rank of the post.

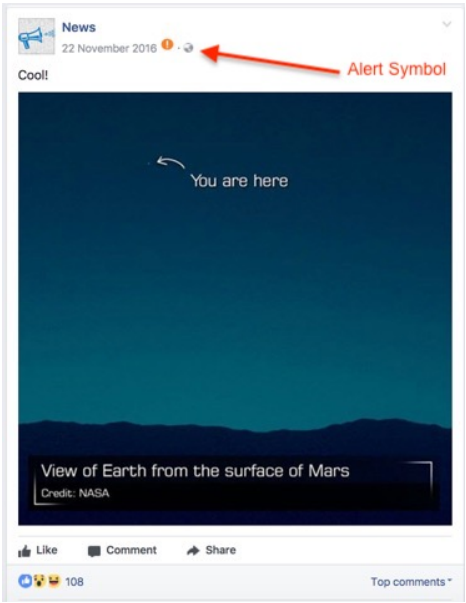

Fig. 4. (Best viewed in color print) Screenshot of the News Feed of a Facebook user when the extension is enabled. An *orange* alert symbol appears next to the post timestamp which indicates that the post is *maybe not credible*.

Browser Extension: A large no. of people use browsers to access Facebook. Therefore, we developed a Google Chrome browser extension, named, Credibility Investigator (CbI). Once installed and enabled, it seamlessly integrates credibility ranks with the user's news feed. CbI loads every time a user logs in to Facebook. It extracts post IDs of all the public posts from user's news feed. The post IDs are sent to the RESTful API. If the API returns rank *4*, a *red* alert symbol is displayed with the post, which indicates that the post is *not credible*. If the rank is *3*, an *orange* alert symbol is displayed, which indicates that the

[10] http://flask.pocoo.org

post is *maybe not credible* and for rank *2*, a *yellow* alert symbol is displayed on the user's news feed along with the post. The yellow alert symbol is to show that the post *maybe credible*. This helps the user to take an informed decision that whether to trust the post or not. In order to minimize the change in user's news feed, we prefer not to display any alert symbol if the post is ranked *1*. Figure 4 shows a screen-shot of the news feed of a Facebook user when the extension is enabled. An orange alert symbol appears next to the post's time-stamp. When we hover a mouse pointer on the *orange* alert symbol, it displays a message - *maybe not credible*.

4 Data Collection and Labeled Dataset Creation

In this section, we describe how we collect data from Facebook for analysis and to build a true dataset of credible and non-credible Facebook posts.

4.1 Data Collection

We collected data using Facebook's Graph API search endpoint. The API returns only the public posts. Using a search endpoint, we cannot directly query posts for a specific event. But we can query pages using a keyword. The response consists of all the pages related to the given keyword. Figure 5 shows our data collection procedure. Our dataset consists of posts created by various news-related pages. We used the keyword 'news' to collect page IDs of various news-related pages. Using page ID, we collected most recent 100 posts and their details from each page. For the purpose of this study, we considered only those pages that have more than 5000 likes on them. Also, there are several news-related pages which are marked as verified by Facebook, for instance, @thehindu, @washingtonpost. We assumed all the posts created by verified pages to be credible. Hence, we have not included such posts in our dataset. We also excluded all the posts that were not in the English language. Thus, we collected an initial dataset of 10,416 public posts published by 107 unique pages on Facebook.

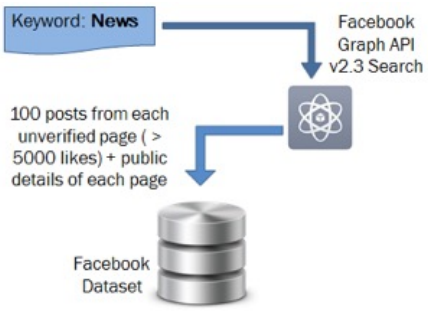

Fig. 5. Describes the data collection procedure used.

4.2 Labeled Dataset Creation

To train our model, we require a labeled dataset. To label the Facebook posts, we followed a similar approach as used by authors in [2], [7], [8] to label tweets from Twitter for credibility. Unlike [8], we curated ground truth for A-C-T instead

of credibility. We took help from human annotators to obtain the ground truth regarding the A-C-T of information present in the posts. As shown in figure 6, we developed a web interface for labeling the dataset. We hosted the annotation portal on Amazon Web Services (AWS) EC2 instance.[11] Annotators were asked to sign-up on the portal and then sign in using the same credentials. All the annotators were the frequent users of Facebook. The average age of an annotator was 22. They were given a set of instructions in which the definitions of A-C-T was mentioned. We provided them with the links to Facebook posts. For each post, they had to choose from 6 options for A-C-T individually.

We asked them to select one of the following options for each post: C1. Definitely accurate/clear/timely
C2. Maybe accurate/clear/timely
C3. Neutral accurate/clear/timely
C4. Maybe not accurate/clear/timely
C5. Definitely not accurate/clear/timely
C6. Skip

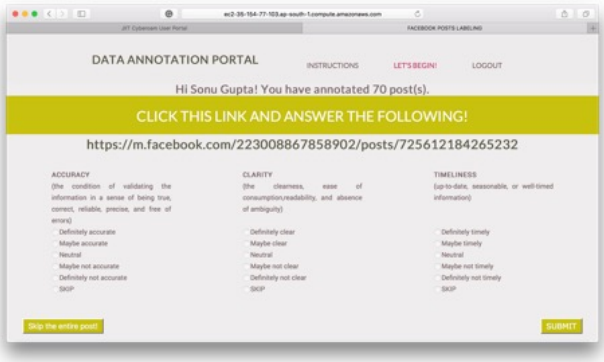

Fig. 6. Screenshot of the web interface used by human annotators to label Facebook posts for A-C-T.

We also provided an option to skip the entire post if they were not sure of their response. We obtained labels for 1,056 Facebook posts selected randomly from 10,416 posts. Each post was evaluated by three annotators to maintain the confidence in the labels, and the mode was calculated to give the label to that post. If all the users had different answers, we calculated median and mean for all such posts. During experiments we found median values give better results than mean values. So, we choose median values over mean values to give the label to that post. To input this data to the binary classifiers, we created two classes with this annotated data, each for A-C-T. Class 1 contains all the posts with the score greater than 3. And remaining posts constitutes class 2. Thus, it becomes a binary classification problem. Table 1 shows the description of our final dataset.

[11] https://aws.amazon.com/ec2/

Table 1. Descriptive Statistics of Facebook Dataset.

1.	No. of Facebook Pages	500
2.	No. of Verified Facebook Pages	223
3.	No. of pages with likes > 5000	483
4.	No. of pages with likes > 5000 + verified = true	223
5.	No. of pages with likes > 5000 + verified = false and language = English	107
6.	No. of posts which are posted by pages with likes > 5000 + verified = false and language = English	10,416

5 Automatic Credibility Assessment

Our goal is to develop a system for ranking Facebook posts on the basis of credibility. Our system classifies each post on the basis of A-C-T and then ranks the posts for its credibility. We adopt supervised machine learning algorithms for classification. First, we perform feature extraction from the posts. Second, we train multiple classification models for A-C-T using several feature sets. Lastly, we compare the accuracy of different machine learning algorithms, using the training labels obtained in section 4. In this section, we also describe Facebook's current technique to identify fake news.

5.1 Facebook's Current Techniques to Identify Fake News

Facebook relies on the community to identify fake news. As shown in figure 7, users can report fake news with the help of a report button. If many people report a story, then Facebook sends it to third-party fact-checking organizations. If the fact checkers agree that the story is fake, users see a flag on the story indicating that it has been disputed, and that story may be less likely to show up in News Feed. Users can still read and share the story, but now there a flag which indicates that the post is fake. The drawback of this technique is that this it is slow and the post would be flagged only after going viral.

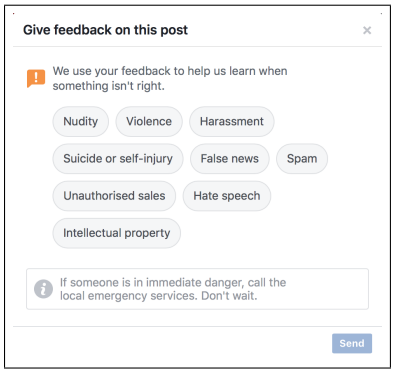

Fig. 7. Facebook's current technique to take community feedback to combat fake news.

5.2 Credibility Assessment Features

The data is collected using Facebook's Graph API. It includes posts, users reactions to that post and the source page details. Generating feature vectors from the posts is an important step that affects the accuracy of the statistical model trained from this data. Here, we use a collection of features from previous work on user-generated contents on Facebook [4], [5], [6]. Along with that, we curate some new features to enhance the model. The features can be broadly divided into three groups (G1) Page-based, (G2) Post-based, and (G3) Page history-based. These three groups of features are used to classify posts on the basis of A-C-T which in turn are used to rank the posts from *1-4*. If a post is *accurate, clear, and timely,* the post is ranked as *1*. If a post is *either not-accurate, not-clear or not-timely,* it is ranked as *2*. If a post is classified as either of two from following, *not-accurate, not-clear, and not-timely,* it is ranked as *3*. And if a post is *neither accurate, clear nor timely,* it is ranked as *4*.

Page-Based Features (G1): Source of the post is an important factor to measure the trustworthiness of the post. With the larger audience, pages play an important role in the dissemination of information. Users share posts created by the pages, and it thus accelerates the diffusion of information in the network. Our system focuses on posts that are either created by pages or are the shared posts with a page as a source. Table 2 presents the page based features. We have three kinds of features, (a) Boolean, (b) Numeric, and (c) Nominal.

Table 2. Page-based Features.

Feature Set	Features
Boolean (20)	Affiliation, birthday, can post, cover picture, current location, working hours, description present, location, city, street, state, zip, country, latitude, longitude, personal interests, phone number, public transit, website field, founded
Numeric (36)	Average sentence length for description, average word length for description, parking capacity, category list length, check-ins, no. of email IDs in description, fraction of HTTP URLs in description, description length, fraction of URLs shortened, fraction of URLs active, likes, page name length, no. of subdomains in URLs, path length of URLs, no. of redirects in URLs, no. of parameters in URLs, [no. of exclamation marks, no. of question marks, no. of alphabets, no. of emoticons, no. of English stop words, no. of English words, no. of lower case characters, no. of upper case characters, no. of newline characters, no. of words, no. of unique words, no. of sentences, no. of total characters, no. of digits, no. of URLs] in description, description repetition factor, talking-about count, were-here count
Nominal (2)	Category, description language

Post-Based Features (G2): Some researchers have shown that post based features are useful to access information trustworthiness [2], [8], [6]. We have used features from these previous works along with some new features. Table 3

presents post-based features. The post based features include text-based features and message link-based features.

Table 3. Post-based Features.

Feature Set	Features
Numeric (38)	audience engaged, [average no. of upper case characters, average length, average word length, no. of English words, no. of English stop words] for description, message, and name fields, description, message, name, no. of comments, no. of reactions (love/like/sad /wow/anger/haha), no. of shares, no. of URLs, total no. of unique domains, has http, the no. of Para, no. of redirects, the number of subdomains, is a shortened URL, are emoticons present, no. of urls, positive, negative and neutral sentiment of users 100 comments in [chronological, ranked, reverse- chronological] order
Nominal (2)	status type [added photos, added video, created event, created note, mobile status update, published story, shared story, wall post], type [event, link, music, note, offer, photo, video, status]

Page-history Based Features (G3): On Facebook, pages are often used to disseminate the information. Every page is associated with a category. It is assumed that the topic of posts created by the page is in accordance with the page category. In our preliminary analysis, we found that this is not the case every time. There are several pages with *News* in their category but they do not or rarely post anything related to news. Thus, page history plays an important role in assessing the credibility of the post especially when it comes to news. Its a background check for the post before believing in its content. To calculate the page-history of a page, we considered the last 100 posts or posts from last 7 days, whichever is lesser. Table 4 presents the page history-based features.

In [1], researchers found that the messages with the least credibility are associated with negative social events. Such messages also contain strong negative sentiment words and opinions. So, it can be concluded that sentiment history is a good indicator of the trustworthiness of the user. We can calculate sentiment history of a page also. To do that, our system finds the sentiment of all the posts collected for the source page. We estimate the frequency of posts with negative, positive and neutral sentiment and use them as features.

Table 4. Page-history Based Features.

Feature Set	Features
Numeric (54)	Daily activity ratio, audience engaged, [average no. of upper case characters, average length, average word length, no. of English words, no. of English stop words] for description, message, and name fields, no. of posts containing the field [description, message, name], no. of comments, total reactions. No. of reactions (haha/like /love/wow/sad/angry), no. of shares, no. of posts with status type [added photos, added video, created event, created note, mobile status update, published story, shared story, wall post], no. of posts with type [event, link, music, note, offer, photo, video, status], total no. of URLs, total no. of unique domain, no. of posts with positive, negative, neutral sentiment.

5.3 Classification Algorithms

The dataset which we obtained after the labeling was unbalanced. In the real-world scenario, the datasets are always imbalanced. We experimented with both over-sampling and down-sampling techniques. For over-sampling, we used a well-known technique called Synthetic Minority Over-sampling Technique (SMOTE) [3]. But it further decreased the performance of the classifier. We also experimented with down-sampling, but it resulted in over-fitting of the model. Therefore, we adjusted the weights by changing the weight parameter of the classifier to balanced. We went ahead with this and performed all the experiments.

We tested and evaluated various classification algorithms to classify the data on the basis of A-C-T: Naive Bayesian, K-nearest neighbors, decision tree, random forest, gradient boosting classifiers, artificial neural networks and support vector classifiers. Here, we reported only the best results from all the algorithms. We trained three classifiers each for *Accuracy, Clarity, and Timeliness*. Table 5, 6, 7 shows the accuracy and ROC AUC values for various classification algorithms that we applied on our feature set for *Accuracy, Clarity, and Timeliness*. We also trained the models for the individual type of features and one with all the features combined. All the training models were evaluated using 10-fold cross-validation. To perform all the experiments, we used scikit-learn, a machine learning library for the python programming language.[12] We achieve the best results when all the features are used together to train the model for A-C-T respectively. Random Forest algorithm out-performed other algorithms with the best ROC AUC score of 0.916, 0.875, and 0.851 for A-C-T respectively as shown in Table 5, 6, 7.

Also, it is worth stating that even though we used the same feature set to train models for A-C-T, the accuracy peaked for different subsets of the feature set. It means, for A-C-T each feature have different feature importance. For instance, feature which is important for *accuracy* may not be as important to *clarity*. This can be seen from the results. For *accuracy* and *clarity*, page-based features performed the best whereas for *timeliness* the combination of all the features gave the best results. For the same reason, the performance decreases in some cases on the addition of more features.

Comparison with baseline model: In the best of our knowledge, there is only one study on the credibility of the Facebook posts [14]. Saikaew and Noyunsan developed two chrome browser extensions. The first extension was used to measure the credibility of each post by asking users to give a score from l(the lowest value) to 10 (the highest value). These post evaluations were used as a data to train an SVM model. They trained the model on mere 8 features; likes count, comments count, shares count, URL count, images count, hashtag count, video count, is location present. The second extension was used to automatically evaluate the credibility of each post-real-time using the SVM model. On a validation set of 1,348 posts, they report an accuracy of 81.82%. Due to a better feature

[12] http://scikit-learn.org

Table 5. Results for supervised learning experiments for *Accuracy* for four classifiers over four different feature sets. Random Forest performed the best with Page-Based features.

Classifier	Feature set	Acc(%)	ROC AUC
Random Forest	G1+G2+G3	84.52	0.893
	G1	85.57	**0.916**
	G2	83.46	0.875
	G3	82.25	0.862
SVM	G1+G2+G3	76.69	0.810
	G1	75.71	0.806
	G2	77.97	0.829
	G3	78.20	0.838
Logistic Regression	G1+G2+G3	77.21	0.820
	G1	80.33	0.848
	G2	74.69	0.799
	G3	76.18	0.805
Naive Bayesian	G1+G2+G3	65.30	0.706
	G1	67.54	0.727
	G2	63.71	0.695
	G3	62.64	0.681

Table 6. Results for supervised learning experiments for *Clarity* for four classifiers over four different feature sets. Random Forest performed the best with Page-Based features.

Classifier	Feature set	Acc(%)	ROC AUC
Random Forest	G1+G2+G3	81.62	0.859
	G1	83.41	**0.875**
	G2	79.38	0.841
	G3	80.19	0.850
SVM	G1+G2+G3	75.99	0.812
	G1	74.86	0.801
	G2	78.25	0.836
	G3	74.20	0.794
Logistic Regression	G1+G2+G3	65.33	0.706
	G1	67.89	0.728
	G2	75.93	0.810
	G3	69.04	0.758
Naive Bayesian	G1+G2+G3	63.13	0.691
	G1	59.81	0.632
	G2	62.42	0.680
	G3	62.56	0.681

Table 7. Results for supervised learning experiments for *Timeliness* for four classifiers over four different feature sets. Random Forest performed the best with a set of all features.

Classifier	Feature set	Acc(%)	ROC AUC
Random Forest	G1+G2+G3	80.40	**0.851**
	G1	78.82	0.839
	G2	76.91	0.806
	G3	75.94	0.828
SVM	G1+G2+G3	73.76	0.784
	G1	71.42	0.763
	G2	67.50	0.727
	G3	69.51	0.759
Logistic Regression	G1+G2+G3	67.63	0.729
	G1	64.16	0.692
	G2	65.93	0.708
	G3	65.35	0.706
Naive Bayesian	G1+G2+G3	61.23	0.652
	G1	60.71	0.636
	G2	53.72	0.581
	G3	59.16	0.629

selection, our models are performing better than this model with the accuracy of 85.57%, 83.41%, and 80.40% for A-C-T respectively. Also, our definition of credibility is more fine-grained which makes our results more promising.

6 Conclusion, Limitations, and Future Work

In this paper, we propose a system which computes the credibility of Facebook posts created by Facebook pages in real-time. Here, we define credibility as a function of *accuracy, clarity, timeliness*. We experiment with 1,056 posts created by 107 *pages* that claim to belong to news-category. We propose a set of 152 features based on post content, source-page and page-history. We use this feature set to train binary classification models each for A-C-T using supervised algorithms. We use the best performing models to develop a RESTful API and a Google Chrome browser extension, named Credibility Investigator (CbI), to rank Facebook posts for its credibility in real-time. To the best of our knowledge, this is the first research work which ranks the Facebook posts for its credibility in terms of A-C-T. The random forest algorithm performed the best and achieved a maximum ROC AUC value of 0.916, 0.875, and 0.851 for A-C-T respectively. There are a few limitations in our proposed system. We do not claim that our dataset represents the entire Facebook news related pages. Facebook does not provide any data about what fraction of information is returned by its API. Due to the Facebook's Graph API restrictions, we can only access public posts. Also, Facebook supports multiple non-English languages too. As of now, our system works only on posts in the English language. In the future, we would like to address this problem. Also, we would like to explore various graph-based techniques to detect the presence of fake pages on the Facebook network.

References

1. Alrubaian, M., Al-Qurishi, M., Hassan, M., Alamri, A.: A credibility analysis system for assessing information on twitter. IEEE Transactions on Dependable and Secure Computing (2016)
2. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th international conference on World wide web. pp. 675–684. ACM (2011)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
4. Dewan, P., Bagroy, S., Kumaraguru, P.: Hiding in plain sight: Characterizing and detecting malicious facebook pages. In: Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on. pp. 193–196. IEEE (2016)
5. Dewan, P., Bagroy, S., Kumaraguru, P.: Hiding in plain sight: The anatomy of malicious pages on facebook. In: Social Network Based Big Data Analysis and Applications, pp. 21–54. Springer (2018)
6. Dewan, P., Kumaraguru, P.: Towards automatic real time identification of malicious posts on facebook. In: Privacy, Security and Trust (PST), 2015 13th Annual Conference on. pp. 85–92. IEEE (2015)
7. Gupta, A., Kumaraguru, P.: Credibility ranking of tweets during high impact events. In: Proceedings of the 1st workshop on privacy and security in online social media. p. 2. ACM (2012)
8. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: Tweetcred: Real-time credibility assessment of content on twitter. In: International Conference on Social Informatics. pp. 228–243. Springer (2014)
9. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd international conference on World Wide Web. pp. 729–736. ACM (2013)
10. Haralabopoulos, G., Anagnostopoulos, I., Zeadally, S.: The challenge of improving credibility of user-generated content in online social networks. Journal of Data and Information Quality (JDIQ) **7**(3), 13 (2016)
11. Li, H., Sakamoto, Y.: Computing the veracity of information through crowds: A method for reducing the spread of false messages on social media. In: System Sciences (HICSS), 2015 48th Hawaii International Conference on. pp. 2003–2012. IEEE (2015)
12. Mendoza, M., Poblete, B., Castillo, C.: Twitter under crisis: Can we trust what we rt? In: Proceedings of the first workshop on social media analytics. pp. 71–79. ACM (2010)
13. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., Menczer, F.: Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of the 20th international conference companion on World wide web. pp. 249–252. ACM (2011)
14. Saikaew, K.R., Noyunsan, C.: Features for measuring credibility on facebook information. International Scholarly and Scientific Research & Innovation **9**(1), 174–177 (2015)
15. Tanaka, Y., Sakamoto, Y., Matsuka, T.: Toward a social-technological system that inactivates false rumors through the critical thinking of crowds. In: System Sciences (HICSS), 2013 46th Hawaii International Conference on. pp. 649–658. IEEE (2013)