AMERICAN UNIVERSITY OF BEIRUT

CREDIBILITY MODELS FOR ARABIC CONTENT ON TWITTER

by
REEM OSSAMA EL-BALLOULI

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Department of Computer Science
of the Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
September 2014

CREDIBILITY MODELS FOR ARABIC CONTENT ON TWITTER

by

REEM OSSAMA EL-BALLOULI

Approved by:

Dr. Wassim El Hajj, Associate Professor — Advisor
Computer Science

Dr. Shady ELbassouni, Assistant Professor — Member of Committee
Computer Science

Dr. Hazem Hajj, Associate Professor — Member of Committee
Electrical and Computer Engineering

Date of thesis defense: September 12, 2014

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT RELEASE FORM

Student Name: El-Ballouli Reem Ossama

Last First Middle

● Master's Thesis ○ Master's Project ○ Doctoral Dissertation

☑ I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

□ I authorize the American University of Beirut, **three years after the date of submitting my thesis, dissertation, or project,** to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

Reem Ballouli 26-9-2014

Signature Date

ACKNOWLEDGMENTS

First and foremost, my special thanks to my advisor Dr. Wassim El Hajj for his time, supervision, and support. I would also like to thank my thesis committee members, Dr. Shady Elbassouni and Dr. Hazem Hajj, Whose assistance, guidance and constructive criticism had made this dissertation possible. Thank you all for your time and assistance.

To my parents, for your love and support, thank you for your encouraging words that kept me driving forward despite the friction. To my lovely sister and brother.

To all my friends, thank you for your support and encouragement. Special thanks to Mona Darwish for your joyful company and continuous support. I would also like to thank Ayman Al Zaatari for engaging and helping with struggles related to my thesis.

Last but not least, thank you God for giving me the strength to complete this dissertation. For all of you, I hereby dedicate this dissertation.

AN ABSTRACT OF THE THESIS OF

Reem Ossama El-Ballouli for Master of Science
Major: Computer Science

Title: Credibility Models for Arabic Content on Twitter

Microblogging websites such as Twitter have gained popularity as an effective and quick means of expressing opinions, sharing news and promoting information and updates. As a result, data generated on Twitter has become a vital and rich source for tasks such as sentiment mining or newsgathering. However, a significant portion of such data is either biased, untruthful, spam or non-credible in general. Consequently, filtering out non-credible tweets when performing data analyses tasks on Twitter becomes a crucial task. In this work, we present credibility models for content on Twitter. We focus on Arabic tweets due to the recent popularity of Twitter in the Arab world and due to the presence of a large portion of non-credible tweets in Arabic.

We build a binary credibility classifier that classifies a tweet that belongs to a given topic as either credible or non-credible. The suggested classifier relies on an exhaustive set of features extracted from both the author of the tweet (user-based) and the tweet itself (content-based). To evaluate the performance of the suggested classifier in categorizing credible vs. non-credible tweets, we compared it to several baselines and to state-of-the-art approaches. The classifier consistently surpassed the accuracy of the baseline approaches. It also outperformed the state-of-the-art approaches with an increase of 14% in F-measure.

Furthermore, we analyzed our feature set by comparing the accuracy of the classifier when trained on user-based features only versus content-based features only. Overall, user-based features only generated better accuracies than content-based features only when tested on multiple topics, indicating that features related to the tweet author are more important than features related to the content of the tweet, when it comes to deciding on the tweet credibility.

CONTENTS

Chapter

ILLUSTRATIONS

TABLES

Table Page

CHAPTER 1

INTRODUCTION

Microblogging websites such as Twitter have gained popularity as an effective and quick means of expressing opinions, sharing news and promoting information and updates. As a result, data generated on Twitter has become a vital and rich source for tasks such as sentiment mining or newsgathering. However, a significant portion of such data is either biased, untruthful, spam or non-credible in general. Up to this date Twitter does not support a clear-cut approach to limit the spread of such information. Consequently, filtering out non-credible tweets when performing data analyses tasks on Twitter becomes a crucial task.

1.1 Motivation

With the evolution of online social media, the Internet is replacing traditional media like radio and television. The web has become a treasured source of opinions, news and information about current events. In the era before the introduction of the web, people would refer to traditional media for information, but in recent years, one tends to obtain information from blogs, social networks sites and E-commerce sites. When we acquire new information, we are likely to use certain cues to help us assess the credibility of what is being received. The basic definition of credible information is one that is not fake, spam, or biased, yet there is no consensus in the literature about a definition for information credibility. Some tangible cues used for assessment could include fluency, author expertise, lifestyle, eye contact, etc. It is this set of cues that aid us to accept or block what we are hearing or reading. These cues however are not sufficient since they can only be harvested in a physical social environment. Information on the web is plain textual data where cues like physical appearance, fluency, and speaker tone are snatched away. Sometimes, even the source or the author of the information is undetermined making the processes of assessing the credibility of information found on the web tedious and unintuitive for basic users. Hence, filtering out non-credible information from the web is extremely important so as to protect readers from being misinformed, scammed, or wrongly influenced. In this work, we adopt the Merriam Webster definition of credibility that states: *credibility is the quality of being believed or accepted as true, real or honest*.

Fake Twitter accounts are becoming an up rising phenomenon. Fake accounts are being used to influence public opinion or to generate profit. Some fake accounts use the names of popular people and tend to present real content including pictures and videos. Up to this date Twitter doesn’t have a clear-cut approach to prevent such behavior, instead when the official person becomes aware of his/her fake representative, he/she will report it to Twitter and the account will be blocked by Twitter after performing some verification measures. One account gets blocked but another gets created; a total of 100,000 fake accounts can be automatically generated in 5 days. While automated bots manage some of the fake accounts, some accounts are managed by humans manage and these accounts are harder to detect since they tend to abide to a normal user behavior. According to Ammar Muhammad, a specialist in social media, 27% of followers of the top 10 Twitter accounts are reportedly fake [1]. In addition, it was also noted that a fake Twitter account impersonating Egyptian President Hosni Mubarak appeared to widely impact and shape the events during the Egyptian Arab Spring [2]. Companies, celebrities and leaders are interested in having a good reputation amongst their target audience, in fact there are websites such as *likefolllowers.org* [3] that allows its users to target the Arabic market by increasing the number of Arabic followers by offering companies and leaders the option of buying Arabic followers. The website encourages users to buy Arabic Twitter followers as it claims that's it's "a way to increase your Twitter account's social credibility" [3]. The above examples highlight the importance of filtering out non-credible Arabic content. To the best of our knowledge no supervised model has been developed to assess the credibility of Arabic content. Our research is dedicated to do so.

1.2 Problem Statement

Up to this date Twitter does not support a clear-cut approach to limit the spread of biased, untruthful, spam or non-credible tweets. Consequently, filtering out non-credible tweets when performing data analyses tasks on Twitter becomes a crucial task. Significant amount of research is dedicated to credibility classification of English content as opposed to Arabic content on Twitter. This thesis addresses the matter of building a credibility classifier for classifying Arabic tweets as credible or noncredible.

1.3 Objectives and Contribution

In this work, we present a credibility classifier for content on Twitter. Unlike previous work that focused on English content or factual tweets, our work analyses the credibility of any tweet type and targets Arabic tweets, a challenging language for NLP in general. We focus on Arabic tweets due to the recent popularity of Twitter in the Arab world and due to the presence of a large portion of non-credible tweets in Arabic. We build a binary credibility classifier that classifies a tweet that belongs to a given topic as either credible or non-credible. Our classifier relies on an exhaustive set of features extracted from both the author of the tweet (user-based) and the tweet itself (content-based). Our main contributions are as follows:

- Surveying existing work on credibility of content on Twitter.
- Proposing the first supervised model for classifying Arabic tweets as credible or noncredible.
- We provide the first public Arabic dataset annotated for credibility.
- We propose a unique interface that provides the real context for each tweet, such as the author profile and a Web search about the content of the tweet.
- Evaluation of the performance of our classifier in comparison to most recent work on Credibility of English content on Twitter.

1.4 Thesis Plan

The remainder of this thesis is organized as follows. CHAPTER 2 surveys existing work on credibility of content on Twitter. CHAPTER 3 presents the methodology used and the steps involved in the creation of the credibility classifier. First, we present the dataset collection process. Second, we present the details of the annotation process, where recruited workers manually label each tweet as credible or noncredible. The labeled tweets form the training set required by the supervised credibility classifier. Third, we discuss the features used for training the classifier and details of their extraction. Finally the performance of the classifier is evaluated and discussed in details in CHAPTER 4. CHAPTER 5 concludes our work.

CHAPTER 2

LITERATURE REVIEW

There is limited work in the literature dedicated to credibility classification on Twitter. Related work can be broadly classified to credibility classification of either Arabic or English content. Credibility of Arabic content had not received profound attention from researchers and as such this area is still open for development. To the best of our knowledge there is only one research dedicated to credibility detection of Arabic content on twitter [4]. In section 2.1 we discuss the details of this research. On the other hand, credibility detection of English content received the highest attention from researchers [5-7]. We discuss related work done on English content in section 2.2. Some related work targeted credibility classification of twitter users instead of tweets relying solely on the user to determine credibility. We discuss details of such research in section 2.3.

2.1 Credibility of Arabic Content on Twitter

In this section we present related work targeting credibility classification of Arabic content on Twitter. R. Al-Eidan, H. Al-Khalifa and A. Al-Salman [4] automate the process of credibility assessment of Arabic news tweet. They collected 600 tweets over a period of two weeks. The tweets collected are a result of trendy political topics suggested by political experts as queries. For the same queries, they collected 179 news articles from Aljazera.net, Saudi press agency, and Google news. These articles are taken as sources of factual and truthful news. The documents collected, both tweets and articles, undergo text processing, where stop words are removed, and words get tagged with POS (part of speech) tags and finally stemmed. After text processing each tweet and document is now represented as a bag of words (BOW) facilitating the calculation of TFID scores for terms in the tweet. Later on, both tweet and documents are represented as vectors of TFID scores and cosine similarity is computed between the tweet and the document. This score measures the similarity of the tweet with the factual truthful article news.

To compute the three different levels of credibility (low, high, average), the authors compute the cosine similarity between the document and the tweet for three different POS tags. Any tweet having a cosine similarity for terms tagged as noun/others above a certain threshold will be labeled as low credible. On the other hand any tweet having cosine similarity score for all pos tags above the threshold is said to be a highly credible tweet. The remaining tweets are tweets labeled average credible. The cosine similarity score is the only feature used by the authors to assess credibility in the first approach. However in the authors tested another approach where they included four other features to the credibility measure; the presence of inappropriate words, presence of an authoritative URL, whether the author is verified, analytic degree from degree.com API. To evaluate their approach the authors ask three political experts to evaluate the ranked tweets (ranked by credibility score). The authors achieve an average precision of 0.52 and average recall value of 0.56 with the first approach, which outperformed the second approach.

The proposed model requires the continuous maintenance and update of a factual database that includes almost every possible news topic. If we assume that the model automatically updates its database by continuously mining information on the web and storing it, then this will allow the model to predict credibility of previously discussed topics on the web. Yet, it will still fail to assign credibility values for tweets discussing breaking events as such events may not have corresponding information on the web. Furthermore, any textual information belongs to one of two categories: factual or opinion. These two categories can broadly classify textual information. The aforementioned work focused on factual data i.e. news and ignored the credibility of other texts that includes opinions. Hence, the model fails to predict credibility of any nonfactual tweet.

2.2 Credibility of English Content on Twitter

In this section we present related work on credibility classification of English content. Some researchers build classifiers/rankers to classify/rank tweets according their credibility, while others build classifiers to classify a cluster of tweets composing a topic. We discuss work on credibility of tweets and tweet clusters in sections 2.2.1 and 2.2.2 respectively.

2.2.1 Credibility of tweets

A. Gupta and P. Kumaraguru [7] propose an automated ranking scheme to present the user with a ranked output of tweets according to the credibility of the information in them. To create the dataset, the researchers first obtain query terms by collecting trends from Twitter API every 3 hours. Next they use these trends as queries to collect tweets from Twitter Stream API. Tweets for a certain topic where collected while the topic is trending and stopped otherwise. Over a period of a month and a half they collected 35 million tweets. From the list of trending topics, the authors shortlisted 14 major events. Each event had one or more trending topics associated with it. For each event, they considered tweets containing the words in the associated trending topics to be set of tweets for that event.

A sample of 500 tweets from each topic was presented to human annotators through a web interface that the authors developed. A total of 7,000 tweets got annotated, and each tweet annotated three times. They provide the annotators with the definition of credibility as found in Oxford dictionary i.e. "A tweet is said to contain credible information about a news event, if you trust or believe that information in the tweet to be correct/true." Annotators where also provided with a brief description of the event and URL links to news articles related to the event found in premier news websites. Annotators where asked to give one of four credibility labels for each tweet; definitely credible, seems credible, definitely incredible, and I can't decide. A total of 5,578 annotations where obtained after discarding annotations on which the three annotators didn't reach consensus.

Next the authors extract both user-based and content-based features from each tweet. Some user-based features include registration age, number of tweets, number of followers etc. Content-based features include number of retweets, sentiment score, and number of swearwords etc. The authors used regression analysis over the set of features to identify the prominent features that can help assess the credibility of a tweet. The results show that features like unique characters, swear words, pronouns, presence of, number of followers and the length of username where prominent features and where indicative of credibility.

They finally used SVM ranking scheme to rank the results according to the credibility of the information contained within the tweet and learned features. The ranking model significantly surpassed the NDCG value achieved by the baseline. The baseline represents the current ranking scheme on Twitter, which ranks tweets by recency. Next they apply psudo relevance feed-back by re-ranking the tweets according to their BM25 similarity score with the most frequent unigrams from the top-k tweets. This additional step increased the NDCG to 0.73.

A. Gupta, P. Kumaraguru, C. Castillo and P. Meier [6] develop a real time credibility analyzer through a semi supervised ranking model for credibility assessment. For their study, the authors collected tweets through Twitter streaming API for six prominent events that affected a large population and generated high content during 2013. To obtain ground truth data, 3,000 tweets, 500 from each event, was presented to annotators through CrowdFlower.com. The annotators perform a first pass over the tweets labeling tweets that contain informative information about the events. 45% of the tweets where marked as informative, these tweets where used in the next step, which is credibility assessment. Annotators where presented with the same information as annotators in the work of [7]. Each tweet got annotated by three annotators and is given one of four levels of credibility definitely credible, seems credible, definitely incredible, I can’t decide. Tweets not receiving consensus from the annotators where discarded. It was noted that 23% of the tweets where labeled definitely credible and 6% where definitely not credible to annotators.

Next, they extract features from each tweet. They extract a total of 45 features, all of which can be extracted in real time. Their features set doesn't include features related to a group of tweets as in [5]. Neither does it include user-based features that are dependent on the previous tweet posts of a user. Next, the feature vectors for all the annotated tweets were given as input to SVM-Ranking algorithm as training dataset. They used the trained model as a backend for their system. When a new Twitter feed comes in real-time the rank of the tweet is predicted using the learnt model and displayed to the user on a scale of 1 (low credibility) – 7 (high credibility). They performed multiple experiments, to determine the cutoff threshold for each rating level. Furthermore, the authors developed a Chrome Extension (TweetCred) in order to allow its users to check the credibility scores of tweets embedded into their Twitter webpage.

The system works as follows first, a user logs in to his Twitter account, and as soon as a tweets start loading, the chrome extension sends the Ids of those tweets to the backend web server that hosts the credibility model. Next the server requests the complete JSON representation of the tweet from Twitter API. After obtaining the JSON, all the features are extracted and then credibility score is computed using the prediction SVM-Rank model. Finally, the webserver sends back the score to the user's browser where the score is displayed alongside each tweet.

Credibility scores are cached for 15 minutes. Hence, if a tweet receives multiple credibility requests from multiple users then the tweet credibility score is retrieved from the cache if the request time is bounded to 15 minutes. On the other hand, if the cached value is older than 15 minutes then a new request for credibility score is issued to TweetCred. Flushing the cached results every 15 minutes accounts for any changes in values of some features such as follower, followees count, retweet count etc.

Many users downloaded TweetCred chrome extension and tested its performance. A total of 717 unique users have used TweetCred over a period of three weeks. TweetCred predicted credibility scores for 1.1 million tweets. Some users gave feedback on the predicted scores and they agreed with 43% of the predicted scores. They used feedback from users to improve their model.

Unlike TweetCred our classifier relies on features that are dependent on previous tweet posts of a twitter user. Such features include expertise, tweeters activity etc. With the inclusion of such features our classifier outperformed TweetCred with 14% increase in f-measure. A detailed comparison of the performance of our work to TweetCred is presented in section 4.2.

2.2.2 Credibility of tweet clusters

C. Castillo, M. Mendoza and B. Poblete [5] focus on building an automatic model for assessing the credibility of a given *set* of tweets. Their analysis is mainly focused on information credibility of news tweets propagating through Twitter. They first use

TwitterMonitor as a tool to extract trendy and bursty topics found on Twitter. *TwitterMonitor* is an online-monitoring tool which detects sharp increases of frequency of sets of keywords found in tweets. For every burst detected, *TwitterMonitor* releases the query keywords for tweets on the trendy topic using Twitter API. A total of 383 trendy topics with the corresponding tweets are collected. Then, 7 participants on Amazon Mechanical Turk (AMT) where asked to distinguish and classify news topics from chat topics by considering 10 tweets from each topic. Next, a decision tree supervised classifier was built and trained on the human collected annotations. Their classifier determined whether a given *set* of tweets belongs to a news topic or not with a precision and recall value 0.922, and 0.972 respectively.

The authors define credibility to be believability: "offering reasonable grounds for being believed". Hence, to complete the credibility classifier, AMT participants where further asked to state whether they believe tweets from news topics that have been previously classified as news are likely to be true or false by viewing 10 tweet samples from this topic. Labels for each topic were decided by majority, requiring agreement of a least 5 evaluators. The labeled data is used to extract relevant features that humans use to assess the truthiness of *set* of news tweets. They used the best-fit selection method to detect the features that decide on the credibility of a *set* of news tweets. The best-fit selection method starts with the empty set of attributes and searches forward. The selected features, 15 in total, where used to build the classifier which assesses the level of credibility of a news tweet.

In conclusion, the authors were able to automatically separate news topics from other types of conversation with an accuracy of 89%. It was also noted that among other features, news topics tend to include URLs and have deep propagation trees. They also built an automatic tool to assess the level of credibility of news topics with an accuracy of 86%. Results show that among other features, credible news propagate through authors that have previously written a large number of tweets, and have many re-posts.

The aforementioned work detects rumor and noncredible topics, but fails to detect non-credible tweets within a credible topic.

2.3 Credibility of Twitter users

While some researchers focused on credibility classification at the tweet level, others have focused on credibility of tweet authors instead relying solely on the author to determine cerdibility. For example the work K. R. Canini, B. Suh and P. L. Pirolli which we discuss in detail in this section.

The work of K. R. Canini, B. Suh and P. L. Pirolli [8] is motivated by the goal of deciding whose updates to subscribe to in order to maximize the relevance and credibility of information received. To address this problem, the authors first explored the factors on a Twitter page that affect users’ credibility judgment about an author on Twitter. In their experiment they presented 98 participants with 30 different generated Twitter pages, where every page includes a user name and an icon, number of followees, followers and tweets, 40 latest tweets by the user, and a word cloud summarizing all tweets by the user. Each Participant was asked to judge the price of a car before and after viewing a Twitter page.

The shift in the judgment made by the participant was used as a measure of implicit credibility judgment for the author of the Twitter page. To construct 30 different pages, three variables where manipulated:

- Author domain. To accomplish this WeFollow[1] was used as a tool to extract the top 10 experts in a certain categories. Five different categories where considered including car, investing, wine, dating, fantasy football. The remaining accounts where picked at random and didn't belong to any category. Authors in the car domain where considered *on-topic* with respect to the target task which is judging a cars' prices. The other domains were *cross-topic,* and the remaining random accounts where considered *off-topic.*
- Social status. For each generated Twitter page the social status determined the number of followers, followees and tweets for that author. When the social status is high, the count of followers, followees and tweets is high.
- Visualization. The generated page either included tweets, or word clouds, or both.

Results of the study showed that the category from which the Twitter authors where selected had strong influence on the credibility judgment made by the participants.

[1] WeFollow is an online tool that ranks and groups twitter users into categories in order to allow its users to find people and follow prominent people by surfing through a certain category. It ranks all users based on how many users in the same category are following him/her.

Authors from the car domain (on-topic) scored higher credibility ratings than those from other domain (cross-topic). Social status and visualization factors had smaller influence.

Based on the results above the authors designed an automatic tool to rank social network users based on their credibility. They defined explicit information credibility of a *source* by expertise of the author, which is defined by support and nomination of peers, and topical relevancy of the discussion topic. Hence the ranking of Twitter users is based on their relevance and expertise for a given topic. To rank Twitter users, the authors first defined a set of users called *voters*; voters are users gathered using the standard Twitter search for a given query. The set of users whom the voters follow constitute the *candidates* set. For each *candidate* u, they retrieved the number of *voters* who follow user u, called f_u.

The total number of followers of a *candidate* is detonated by F_u. The authors use these two values, f_u and F_u, to compute the social status for each *candidate*. Four different formulas relating these two values where tested to calculate the social status. It was found that the best ranking was given by the formula that measures the proportion of one's followers who are in the search results, while appropriately penalizing unpopular users given by: $\frac{f + \propto}{F + \propto + \beta}$

where $\propto$, and β are parameters that need to be optimized for each query. After applying a social filter, the authors implemented a method to re-rank the previous ranked results based on a topic modeling analysis. They collected the entire tweet histories of the ranked candidates and ran Latent Dirichlet allocation (LDA), a topic modeling algorithm, on the corpus. The LDA results provide a way of determining the topical similarity of any candidate to a search query based on the content of the candidate's tweets. The resulting set of candidates is a re-ranked set based on topical relevance of each candidate to a query.

Five different queries where used to evaluate the system (biking, medicine, Photoshop, teaparty, and wine). For each query the authors compiled the top 20 users ranked by their implementation, and WeFollow. The ground truth data is generated by AMT participants. The author's implementation often produced results competitive with WeFollow which only ranks Twitter users based on the number of followees. These results suggest that incorporating a content-based analysis improves the results in comparison to rankings that only depend on explicit expertise of the source (trust and nomination of peers).

The aforementioned work determines the expertise of the author by relying only on his social network. Yet, some authors with high social network status may be noncredible when discussing certain topics that they are not experts in. Authors have expertise in certain topics and their expertise should not solely depend on their social network but must include their historical background. In fact this is one of the features that we included in our credibility classifier. Unlike the aforementioned work, we compute expertise of author u on a query q using his previous historical tweets. Using the previous tweet posts we determine the author's background and his interest, which dictate his expertise. We discuss the details of expertise extraction in section 3.3.3. The above-mentioned model may be able to rule out noncredible authors but will not be able to detect noncredible tweets from credible authors.

CHAPTER 3

PROPOSED APPROACH

To achieve our objective, we divide the objective into 5 steps illustrated in Figure 3.1. First, we collect Arabic tweets using Twitter stream API. Second, we monitor Lebanese news to come up with widely discussed topics in Lebanon. These topics are used as queries to query the tweet dataset. Third, we create an index for the dataset in order to easily retrieve tweets related to specific queries. The index will contain words and the corresponding tweets that contain each word. Forth, we retrieve tweets for three topics about the Syrian revolution and give the retrieved tweets to annotators to annotate. Fifth, we extract both user-based and content-based features form each tweet. Finally we use the extracted feature vector and the annotation labels for each tweet for training our classifier. Sections 3.1-3.3 describe the details of each step.

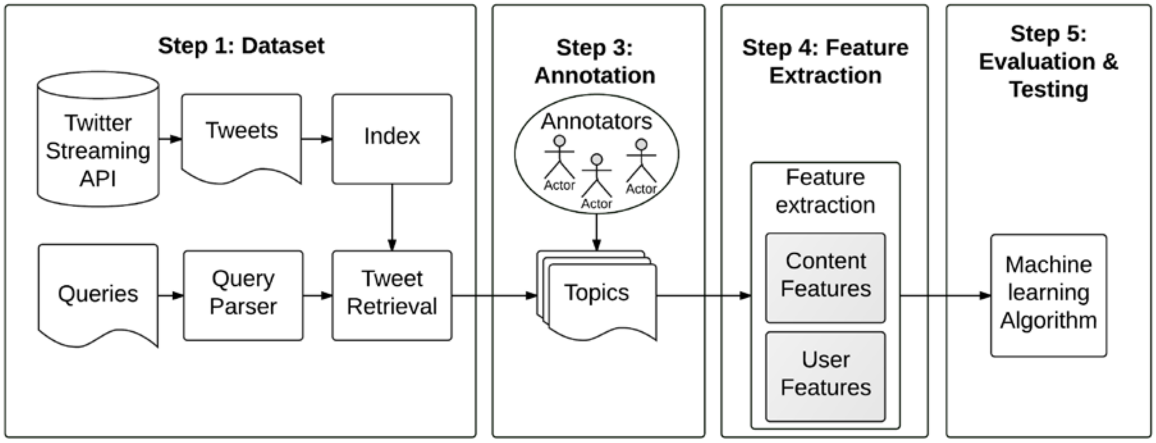

Figure 3.1 High level view of the process of creating our credibility classifier.

3.1 Dataset

In this section, we present the process of Arabic tweet dataset collection. We first introduce the tweet collection method in section 3.1.1 and then we present the query dataset that will be used to query the tweet dataset in section 3.1.2. Next we provide the details of the indexer and the query parser in sections 3.1.3 and 3.1.4. The indexer creates an index for the dataset, which allows easy retrieval of tweets for a given query. The query parser parses the queries before using them for the retrieval. Finally, we highlight the results obtained from querying the index with our set of queries in section 3.1.5.

3.1.1 Arabic Tweet Dataset

Twitter API was used to connect to a stream of Arabic tweets. Certain query terms must be given to the API in order to restrict the stream of tweets received. Hence, providing query terms to the API will collect tweets only containing such terms. In order to avoid this restriction and to expand our collection to as many Arabic tweets as possible, we use a list of Arabic stop words composed of 162 words as queries to the stream API. This allowed us to collect almost all Arabic tweets. Using this method we collected around 36 million tweets, 36,155,670 to be exact for a period of two weeks. The JSON representation of each tweet is stored in a file. The JSON, a human friendly textual representation of information, stores all metadata about the tweet such as author name, count of retweets, and date of creation etc. We resorted to JSON representation to reduce future API calls to Twitter API that is already very restrictive. The JSON will be queried instead for any information related to the tweet.

3.1.1.1 Dataset Cleaning

Upon manual inspection of tweets in the dataset, it was noted that there are a lot of retweets. To increase the variety of our data and to have a clean ground truth data, we filtered out retweets from the collection. If a tweet is a retweet, i.e. a repost of a tweet written by another author, then its JSON representation contains a value for “retweeted_status”. Every tweet that contains such a value is then removed from the dataset. The dataset is now composed of around 16 million (16,127,627). 44.6% of the tweets where retweets. Despite removing retweets, we kept the count of retweets for a tweet as a feature in our feature set, since it could be an indicator of credibility.

3.1.2 Query Dataset

We monitored daily Lebanese news to come up with a set of highly discussed news topics in Lebanon. These topics will form our query dataset and will be used later on to query the dataset. This will allow us to retrieve Arabic tweets corresponding to hot Lebanese news. Table 3.1 provides the news topics and a brief description of each

Table 3.1 Description of trendy lebanese topics

Query topics	***Description***
قوات النظام	The forces of the Syrian government
الإنتخابات الرئاسية في لبنان	The election of Lebanese president
الثورة السورية	Syrian revolution
الأزمة السورية	Syrian problems and concerns related to the Syrian revolution
عون رئيس جمهوريجة	The election of Michel Aoun
المحكمة الدولية الخاصة بلبنان	International Tribunal for Rafic Haririri's assassination

3.1.3 Index Creation

In order to Query the tweet Dataset, an index was created using Apache Lucene, a free/open source information retrieval software library that allows the creation of index and querying of the index. In order to create an index we went through 6 steps that are illustrated in Figure 3.2. The description of each step is described in details as follows:

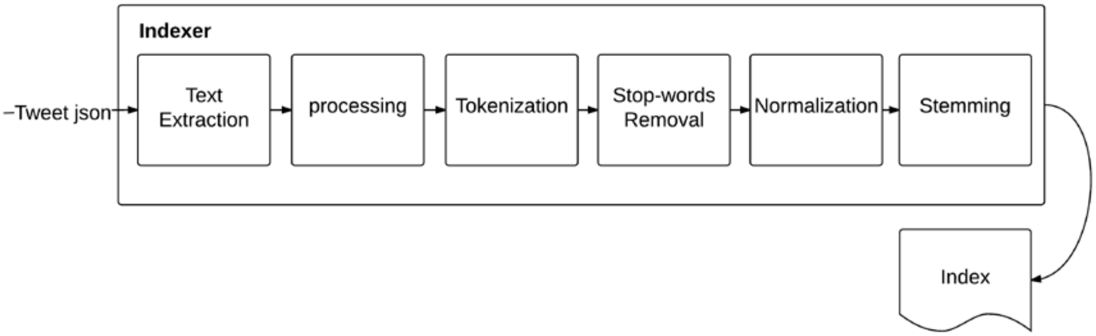

Figure 3.2 Process of creating an index

- **Step 1.** Extraction of raw tweet text from JSON representation. Since we only want to index the content of the tweet itself and not its metadata information present in the JSON.
- **Step 2.** Removal of mentions and URLs from the tweet since we are not interested in retrieving tweets with certain URLs or mention tags.
- **Step 3.** Tokenization of tweet text using Lucene standard tokenizer, where the tweet is split into words by whitespace, punctuations etc.
- **Step 4.** Stop-words where filtered from the tweet using lucenes' Arabic stop word list.
- **Step 5.** Tweet text is normalized, where diacritics are removed and dotted letters such as ي become dot-less character ى.
- **Step 6.** Tweet text is stemmed and attached definite article, conjunction, and prepositions are removed from each word in the tweet.

The result of the above steps is then fed to the indexer, which creates an index for the collected tweets. The index will contain a list of stemmed words and the corresponding tweets in which the stemmed word occurred. The index structure will allow the retrieval of tweets that contain certain query terms.

3.1.4 Query Parser

In order to simulate Twitter search functionality our queries are formatted as phrase queries, where the words in the query have to occur in the tweet in order with some vicinity from each other. In order to match for a query, it is necessary that the query passes through the same steps as the tweet. Steps 3-6 described in section 3.1.3 are also repeated for each query as illustrated in Figure 3.3. The query is first tokenized and divided into query terms. Next, stop-words are removed. Finally normalization and stemming are performed on the query terms. At the end of this step each query will be decomposed of a set of stemmed words that will be used for querying the index.

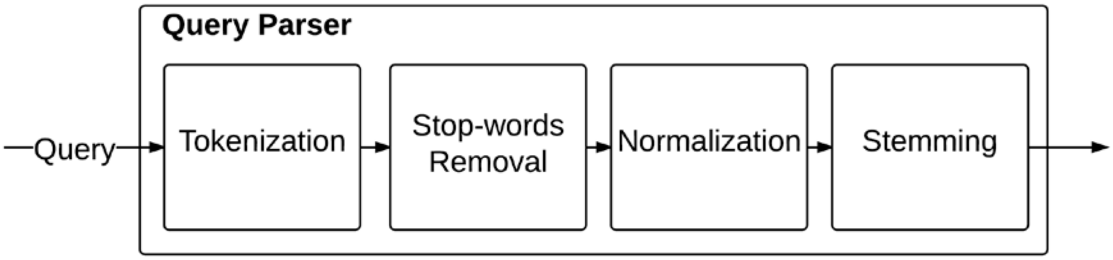

Figure 3.3 Process of creating a query parser

3.1.5 Index Querying Results

In this step, we use the index and the query parser to parse each query in the dataset and retrieve tweets related to each query. Table 3.2 presents the count of tweets retrieved for each query in the query dataset.

Table 3.2 Count of retrieved tweets for each query

Query topic	***Tweet count***
قوات النظام	1785
الإنتخابات الرئاسية في لبنان	38
الثورة السورية	1222
الأزمة السورية	286
عون رئيس جمهوريجة	22
المحكمة الدولية الخاصة بلبنان	22

3.2 Annotation

In this section, we present the details of the annotation process where native Arabic speakers are recruited and asked to label (annotate) each tweet as credible or noncredible. The collection of labeled tweets will be used as ground-truth data for training our classifier. First, we present the dataset that will be annotated in section 3.2.1. Second, we describe the process of picking annotators to complete the annotation task in section 3.3.2. Third, we describe the unique interface that was designed to avoid biasing our annotators in section 3.2.3. Finally, we present the results in section 3.2.4.

3.2.1 Annotation Dataset

The tweets forming the annotation dataset are the result of retrieving tweets for three queries from our query dataset. The selection of queries was dependent on the count of retrieved tweets. The three queries, whose tweets are used for annotation, retrieved the highest tweet count. The tweet count for each query in the dataset is illustrated in Table 3.2.

The queries with the highest tweet count where interestingly all related to Syrian revolution and are listed below:

- قوات النظام
- الثورة السورية
- الأزمة السورية

These queries retrieved a total of 3,393 tweets altogether, and it is this set of tweets that is given to annotators to annotate.

3.2.2 Annotators

Five annotators were exposed to a tutorial session before starting the annotation and each was given a sample task before being recruited to complete the full annotation task. The sample task is used to check the quality of the annotation. It included 20 tweets, one of which is a *gold tweet* allowing us to test how annotators perform on this task. A *gold tweet* is a tweet whose annotation is very evident. The amount of time each annotator needed to complete the task was recorded. This helped us rule out annotators that where haphazardly annotating. In the end, Three out of five annotators passed the sample task and were recruited to complete the full annotation task.

3.2.3 Annotation Interface

Our first attempt to design the user annotation interface included a URL linking to the tweet as displayed on Twitter. We distributed a sample of this interface with some tweets to our research group and asked for annotations. We received feedback from the group as to how easy it was to annotate and what cues they relied on. These suggestions and comments altered the user annotation interface and the form was sent back to the group. After multiple iterations and refinement over the annotation user form, we settled on providing annotators with three URL links. The first link provided the annotators with the tweet text as displayed on Twitter. This option provided annotators with cues such as count of retweets and favorites that the tweet received, author screen name etc. The second link provided a complete author profile view as found on Twitter. The author profile is rich with cues that annotators can use to make their decisions. These cues include the follower count, previous tweet posts, author profile image, and brief description about the author found on his/her profile etc. The third link provided results of a Google search on tweet text. The Google search was restricted to the time of tweet creation ±5 days. Since the user form linked to online websites (Twitter and Goggle) the form was completely interactive. This left room for annotators to decide on what links to visit and what information to rely on when deciding on the credibility of a tweet. Annotators were asked to either label a tweet as “credible” or “noncredible”. They were also given the option to select “cant decide” when they feel confused and unsure. We also added the option “tweet deleted” since some tweeters delete their tweets after posting them and as such annotators will not be able to view the tweet.

Moreover, we adopted common quality measures used in the literature to control the quality of annotations. *Gold tweets* where injected in the final task to assess the quality of annotations. To verify that the selected tweets have evident annotations, we distributed the selected *gold tweets* amongst members of our research group and verified that the tweets received the same annotation from all members. All three annotators passed the gold tweets. In addition, we repeated tweets such that 10% of the total tweets annotated by each annotator is composed of repeated tweets. The annotation of these repeated tweets where used to track the consistency of the annotations over the complete task. Overall, the annotators were mainly consistent with their annotations.

3.2.4 Annotation Results and Analysis

Table 3.3 includes the distribution of credible and noncredible annotations in each query. Tweets that only have either credible or noncredible labels are used for training the credibility classifier. Tweets of Query قوات النظام, الثورة السورية, and الأزمة السورية will henceforth be labeled as topic 1, 2, and 3 respectively. Topic 1 and 3 have similar distributions where the count of credible annotations outweigh that of noncrdible annotations. Topic 1 and 3 are composed of 32% and 24% credible tweets respectively. Topic 2, on the other hand, is balanced having 55% of its tweets annotated as credible and the rest noncredible.

Table 3.3 Distribution of credible and noncredible annotations in each query

Query	*Credible*	*Noncredible*
قوات النظام	1190	556
الثورة السورية	511	623
الأزمة السورية	183	57

3.3 Feature Extraction

In this section, we discuss the content-based and user-based features that were extracted from each annotated tweets. Our set of features is composed of the union of features used in related work mentioned in CHAPTER 2. A total of 48 features constitute our feature vector. We first describe the set of extracted features in section 3.3.1. Next we examine and illuminate the extraction details of sentiment and expertise whose extraction is complex. We present the details of their extraction in sections 3.3.2 and 3.3.3 respectively.

Features

A total of 48 features are extracted from each tweet. The features are broadly categorized into content-based and user-based features as illustrated in Figure 3.4. Content-based features are features extracted from the tweet itself. On the other hand, user-based features are extracted from the tweet author.

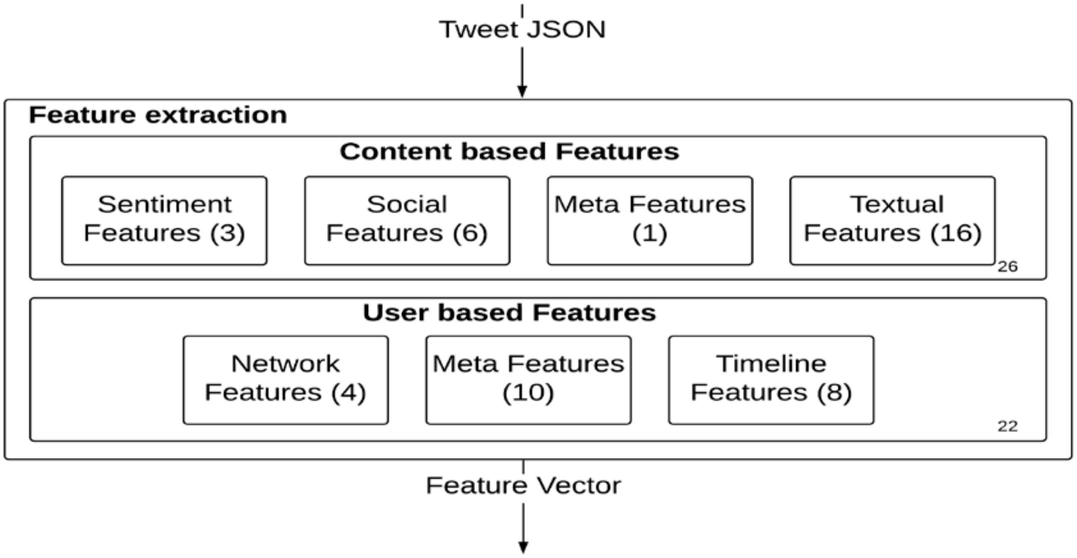

Figure 3.4 Decomposition of of content-based and user-based features

Content-based features are composed of 26 features. These features are further grouped into four subcategories, which are sentiment, social, meta, and textual features. The sentiment category is composed of the total positive sentiment, negative sentiment and objectivity score of the tweet. Sentiment has been previously shown to be an indicator of credibility in the literature and hence we included it in our features. The details of the sentiment extraction are discussed in section 3.3.2. The Social category captures the relation between the tweet and other tweeters. It includes features that reveal this connection. For example, the count of mentions in a tweet highlights the number of tweeters that are related to this tweet. The meta category is composed of a single feature which is the day at which the tweet is posted. This features is extracted directly form the JSON of the tweet. The textual category includes features such as count of exclamation mark, count of unique characters etc. These features are extracted solely form the textual content of the tweet and not from the JSON representation of the tweet. The complete set of content-based features from each category is shown in Table 3.4.

On the other hand, user-based features are composed of 22 features. These features are further grouped into three subcategories, which are network, meta and timeline features. Network features include features that capture the connectivity between the tweet author and other tweeters. For example, the count of followers and friends highlights the popularity of the tweet author. Meta features are features extracted from the JSON, such as registration age of the author, and whether he/she is a verified twitter user etc. Finally, timeline features are features that are extracted from the author's previous tweet posts, also called the timeline. From the author's timeline we can determine the expertise of the author to a given query or topic by counting the number of tweets in his history that are related to the query or topic. The extraction of expertise is discussed in details in section 3.3.3. We also capture the activity of the tweet author by monitoring the time stamp of his previous tweet posts. The complete set of user-based features from each category is illustrated in Table 3.5.

The extraction of most features requires simple computations. However, the extraction of sentiment and expertise is complex and we discuss in details in section 3.3.1 and 3.3.2.

3.3.1 Sentiment Extraction

In this section, we present the sentiment extraction process. To extract sentiment we used ArSenL [9], an Arabic sentiment lexicon. Four existing resources where used in the creation of ArSenL: English WordNet (EWN) [10], Arabic WordNet (AWN) [11], English SentiWordNet (ESWN) [12] and the Standard Arabic Morphological Analyzer (SAMA) [13]. ArSenL contains Arabic lemmas with their corresponding positive, negative and objective score. The process adopted is depicted in Figure 3.5. For each tweet we remove all non-Arabic word characters such as Urls, mentions, and hashtags. Next the tweet is tokenized and feed into MADAMIRA [14] to obtain the lemmas for each word in the tweet. MADAMIRA is a morphological analysis tool for Arabic text developed by Colombia University. Finally using the lemma for each word in the tweet we extract its corresponding positive, negative and objective score from the ArSenL lexicon. To compute the positive score of the whole tweet we sum up the positive scores for each word in the tweet. The same approach is done to obtain the tweet's total negative and objective score.

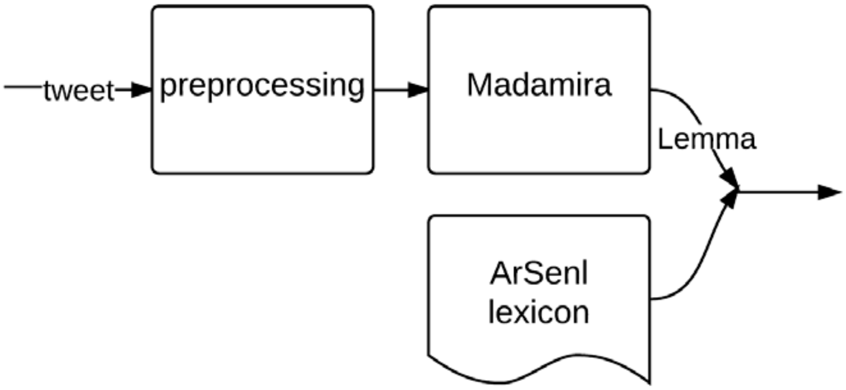

Figure 3.5 Process of computing sentiment of a tweet

3.3.2 Expertise extraction

To compute the expertise of each author, we retrieved previous tweet posts using Twitter API for every author of a tweet in our dataset. We then computed the count of his/her tweets that were on the query topic, and used this as a measure of the user expertise on the query topic. Hence, the expertise of author *u* on query topic *q* becomes the count of tweets in *u's* timeline that are related to query topic *q*. The process of expertise computation is depicted in Figure 3.6. First of all we created a Language model (LM) from all of the tweets we retrived for each of the three queries that have been annotated earlier. A langugage model is a probality distribution over the words in the set. Next, a language model was created for each tweet in the author's time line. Finally, we measured the Jensen-Shannon distance (JS-distanc) between the two language models. The Jensen-Shannon distance is a profound statistical measure ferquently used when measuring the distance between two probability distributions. In other words, we are able to determine how similar a tweet is to a query by measuring the JS-distance between the tweet's LM and the query's LM. Hence, we considered a tweet to be on query topic if the JS-distance between the tweet's language model and the query language model was below a learnt threshold.

To determine the cutoff threshold below which a tweet is considered to be on the query topic, we created a fake user timeline. The timeline is composed of 200 tweets, 100 of which are manually annotated to be on query topic, and the remaining 100 tweets are manually annotated to be off topic. The set of 200 tweets have been carefully chosen to include tweets from diverse topics. Next, we computed the JS-distance between each tweet's Language model in the fake timeline and the query language model. Given the manually on-topic/off-topic annotations and the JS-distance for each tweet we use decision tree to learn the threshold, which was found to be 0.9. Hence, any tweet that is at distance of 0.9 from a query topic is considered similar and related to the query topic.

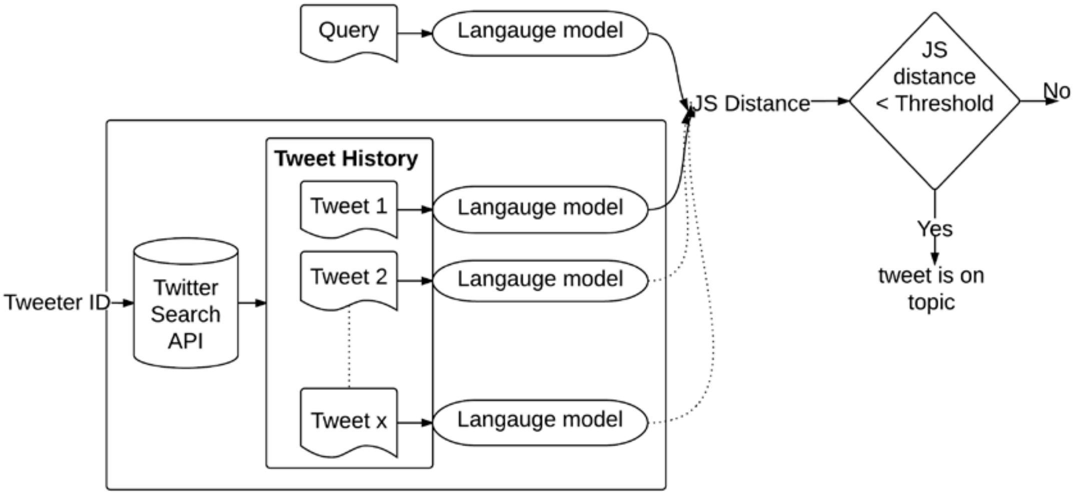

Figure 3.6 Process of computing expertise of a tweet author

Table 3.4 Detailed description of all content-based features

Content-based features		
Feature		***Description***
Sentiment features	Positive sentiment	Sum of positive score of words in the tweet
	Negative sentiment	Sum of negative score of words in the tweet
	objectivity	Sum of objectivity score of words in the tweet
Social features	Count of mentions	Count of user mentions found in the tweet
	has user mention	True when tweet mentions another user on twitter
	count of retweets	Count of retweets the tweets has received
	Tweet is a retweet	True if the tweet is a retweet
	Tweet is a reply	True if tweet is a reply
	retweeted	True if tweet has been retweeted
Meta Features	day of week	Day of the week the tweet is posted
Textual features	Length of tweet in words	Count of words that constitute the tweet
	count char / count words	Ratio of length of tweet in character to length in words
	count of urls	Count of urls found in tweet
	Length of tweet in chars	Count of characters that constitute the tweet
	count of hashtags	Count of hashtags found in the tweet
	count of unique words	Count of unique words that constitute the tweet
	count of unique chars	Count of unique characters that constitute the tweet
	has hashtag	True when tweet contains at least one hashtag
	has url	True when tweet contains at least one url
	Count of ?	Count of question marks found in the tweet
	Count of !	Count of exclamation marks found in the tweet
	has !	True when tweet contains at least one exclamation mark
	has ?	True when tweet contains at least one question mark
	Count of ellipses	Count of ellipses found in the tweet
	has stock symbol	True if tweet contains stock symbols
	count of special symbols ($!)	Count special symbols found in the tweet
	Used url shortner	True if urls found in the tweet are shortened

Table 3.5 Detailed description of all user-based features

User-based features		
Feature		***Description***
Network features	count of followers	Count of Twitter followers
	count of friends	Count of Twitter followees
	fo/fe	Ratio of follower count to followee count
	fe/fo	Ratio of followee count to follower count
Meta features	is verified	True when user is verified to be an authentic user by Twitter
	has description	True when user provides a textual description on his profile
	length of description	Length of description found on users profile
	has url	True when user provides a url on his profile
	has default image	True when twitter default image is users profile image
	length of screen name	Length of user screen name in characters
	Registration Age	Number of days since user registered on Twitter
	Listed count	Count of lists user is listed in
	status count	Count of tweets user has posted since registration
	favorites count	Count of tweets user has favorited since registration
Timeline features	Tweet time spacing	Average Time in seconds between posts in user history
	status retweet count	Count of retweets in user previous tweet posts for the user
	Retweet fraction	Fraction of tweets that are retweets in the users previous tweets posts for the user
	Average tweet length	Average length of previous tweet posts for the user
	Average urls/mentions ratio in tweets	Average ratio of url count over mention count found in previous tweet posts for the user
	Average # hashtags	Average count of hashtags found in users previous tweet posts
	Average Tweet length	Average length (in words) of users previous tweet posts
	Focus of user on topic	Determined by looking at previous tweets of the author and checking

CHAPTER 4

PERFORMANCE EVALUATION

In this chapter, we present the performance evaluation of our credibility classifier. We used the annotated data and the extracted features to train a random forest decision tree classifier using Weka [15]. We validate the applicability of our classifier by doing three different experimental setups. First, we compare the accuracy of the classifier with common baselines in section 4.1. Next, we compare our classifier's performance to TweetCred [6], the most recent work done on credibility classification, in section 4.2. Finally we analyze our feature set by comparing the accuracy of the classifier when trained on user-based features only versus content-based features only in section 4.3.

4.1 Our Classifier vs. Baselines

Recall that we annotated 3 different topics as discussed in section 3.2.1. For the first experiment we used each topic on its own and we merged these topics together and created a combined set. Each of these set have been used separately as a training set to train the classifier. We compare our performance to three common baselines using 10-fold cross validation. The first baseline is the stratified baseline, where the classifier makes random predications according to the distribution of credible and noncredible tweets in the training set. Hence, if the training set includes 80% credible and 20% noncredible tweets, then the stratified baseline randomly predicts 80% of the test set to be credible and 20% to be noncredible. The second baseline is the majority class baseline. Such a classifier predicts all tweets to belong to a single class and this class is the majority class in the training set. Hence, if the training set is mostly composed of credible tweets then each instance in the test set is labeled credible. The third baseline is one that makes uniform predictions such that both credible and noncredible classes are equally likely.

Figure 4.1presents the average f-measure of our classifier in comparison to the three baselines. As shown in Figure 4.1, our classifier consistently surpassed the accuracy of the baseline approaches. This indicates that the user-based and content-based features we used are good indicators of credibility.

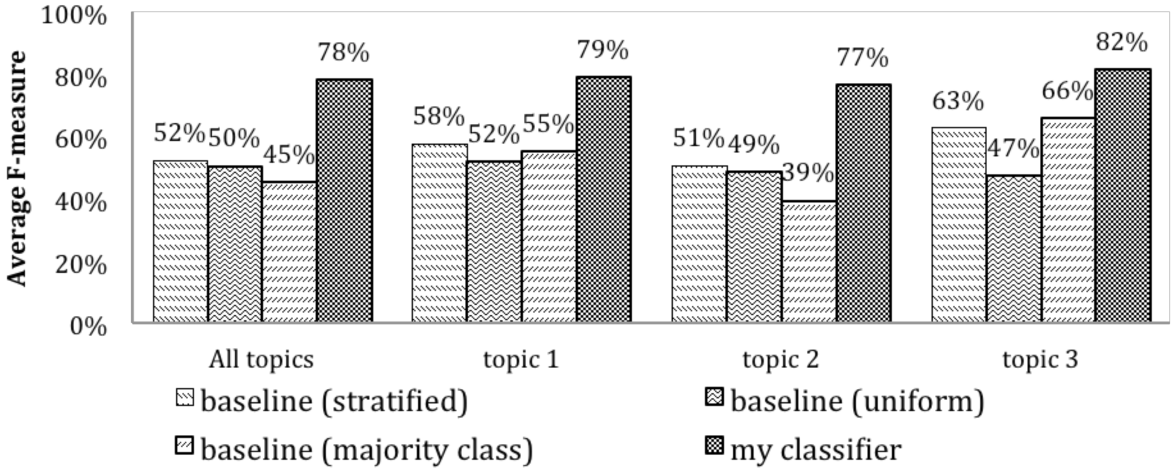

Figure 4.1 Performance evaluation of our classifier vs. baselines

4.2 Our Classifier vs. TweetCred [6]

For the second experiment, we trained our classifier on topic 2 and used topic 1 as test set to test its performance. Topic 2 was chosen to be the training set because it's the only balanced set out of the three annotated topics. Training on a balanced set avoids having a biased classifier. Furthermore, We obtained the credibility scores for each tweet in topic 1 using TweetCred [6]. In the end we have two scores for each tweet in topic 1, the first score is obtained from our classifier and the second score is obtained from TweetCred. Given these two scores we compare the performance of our classifier to TweetCred.

TweetCred is the most recent work available on credibility classification on Twitter. Details of TweetCred are presented in section 2.2.1. The scores obtained from TweetCred API range from 1 to 7, where 1 indicates low credibility and 7 indicates high credibility. To fairly compare our classifier to TweetCred we must compress their score range to two values either credible or noncredible. To determine the cut-off threshold below which a tweet is noncredible using TweetCred, we used our annotations and TweetCred scores to train a decision tree. The cut-off threshold is determined to be 3.5. Hence, any tweet receiving a TweetCred score above 3.5 is credible and noncredible otherwise.

Figure 4.2 depicts the average f-measure of TweetCred vs. our classifier. TweetCred achieves an f-measure value of 55%, while our classifier achieves an f-measure value of 69%. Our classifier outperformed TweetCred by 14%. TweetCred classifier relies solely on features that can be extracted in real time. Hence, features that depend on the user timeline such as expertise are not used. It could be that using real-time features only is not sufficient enough to evaluate for credibility. It's also worth noting that in this experimental setup we used a different training and test set and our classifier performed just as well as when evaluating our classifier using cross-validation indicating that our features are profound indicators of credibility across multiple topics.

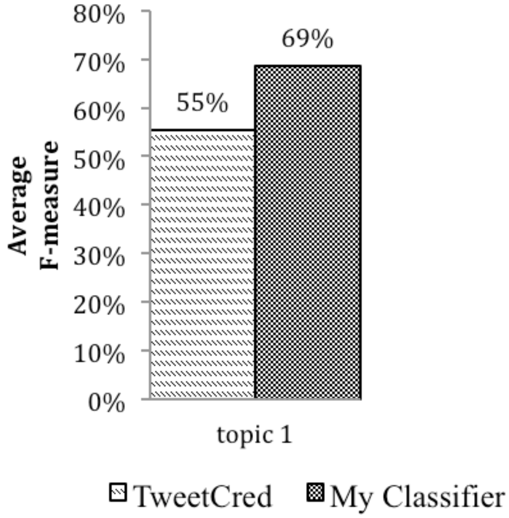

Figure 4.2 Performance evaluation of our classifier vs. TweetCred

4.3 Feature Analysis

In this section, we present the comparison of training our classifier using content-based vs. user-based features. Figure 4.3 illustrates the results of this experiment. We trained our classifier using user-based features only and performed 10 fold cross validation. This is performed on each topic in our annotated dataset and on the combined set, which contains all 3,393 annotated tweets from the three topics. We repeated the above steps but using content-based features only. As shown in Figure 4.3, using user-based features improves the performance of the classifier as opposed to using content-based features. This observation is consistent across all topics used with the exception of topic 2. In topic 2, using content-based features outperforms user-based features. This may be due to the fact that two different annotators annotated topic 2. We intend to investigate this behavior further in future work. To avoid this issue in the future and to make our classification model more robust, each topic will be annotated with at least 3 people and then a majority vote will be taken to collect the final ground

truth.

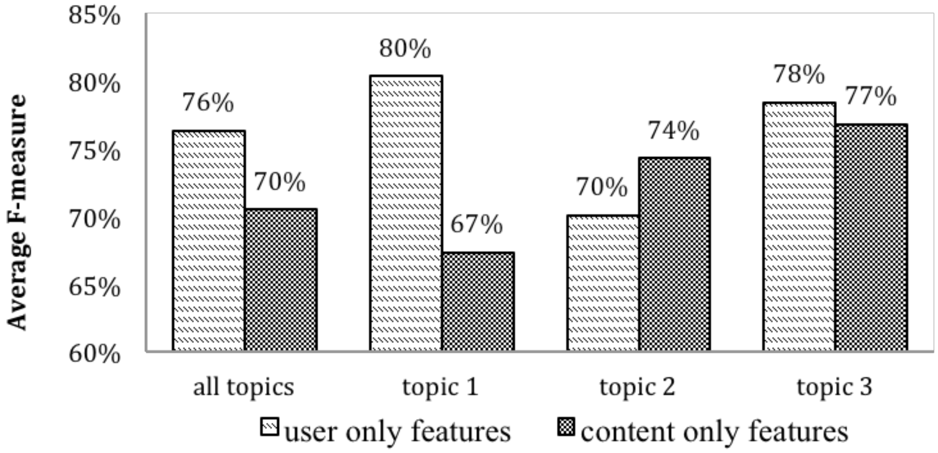

Figure 4.3 Performance evaluation using content-based features vs. user-based features

CHAPTER 5

CONCLUSION

In this work, we presented the first supervised credibility classifier for Arabic content on Twitter. We focused on Arabic tweets due to the recent popularity of Twitter in the Arab world and due to the presence of a large portion of non-credible tweets in Arabic.

We built a binary credibility classifier that classifies a tweet that belongs to a given topic as either credible or non-credible. The classifier relied on an exhaustive set of features extracted from both the author of the tweet (user-based) and the tweet itself (content-based). We evaluate the performance of the suggested classifier in categorizing credible vs. non-credible tweets, by comparing it to several baselines and to state-of-the-art approach. The classifier consistently surpassed the accuracy of the baseline approaches. It also outperformed the state-of-the-art approach with an increase of 14% in F-measure.

Furthermore, we analyzed our feature set by comparing the accuracy of the classifier when trained on user-based features only versus content-based features only. Overall, user-based features only generated better accuracies than content-based features only when tested on multiple topics, indicating that features related to the tweet author are more important than features related to the content of the tweet, when it comes to deciding on the tweet credibility.

REFERENCES

[1] Muhammad, Fatima, *Fake Twitter accounts 'a threat'*. (2013). [online]. Available: http://www.saudigazette.com.sa/index.cfm?method=home.regcon&contentid=20130723174449.

[2] Schroeder, Rob, *Mining Twitter data from the ARAB SPRING*. (2012). [online]. Available: https://globalecco.org/mining-twitter-data-from-the-arab-spring.

[3] *Buy Arabic Twitter Followers.* (2013). [online]. Available: http://likesfollowers.org/buy-twitter-followers/buy-arabic-twitter-followers/.

[4] R. Al-Eidan, H. Al-Khalifa and A. Al-Salman, "Measuring the credibility of arabic text content in twitter," in *Digital Information Management (ICDIM), 2010 Fifth International Conference On,* 2010, pp. 285-291.

[5] C. Castillo, M. Mendoza and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web,* 2011, pp. 675-684.

[6] A. Gupta, P. Kumaraguru, C. Castillo and P. Meier, "TweetCred: A Real-time Web-based System for Assessing Credibility of Content on Twitter," *arXiv Preprint arXiv:1405.5490,* 2014.

[7] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media,* 2012, pp. 2.

[8] K. R. Canini, B. Suh and P. L. Pirolli, "Finding credible information sources in social networks based on content and social structure," in *Privacy, Security, Risk and Trust*

(Passat), 2011 Ieee Third International Conference on and 2011 Ieee Third International Conference on Social Computing (Socialcom), 2011, pp. 1-8.

[9] Anonymous. *This paper is anonymized in submission*.

[10] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. J. Miller, "Introduction to wordnet: An on-line lexical database*," *International Journal of Lexicography,* vol. 3, pp. 235-244, 1990.

[11] W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease and C. Fellbaum, "Introducing the arabic wordnet project," in *Proceedings of the 3rd International WordNet Conference (GWC-06),* 2006, pp. 295-299.

[12] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of LREC,* 2006, pp. 417-422.

[13] M. Maamouri, D. Graff, B. Bouziri, S. Krouna and S. Kulick, "LDC Standard Arabic Morphological Analyzer (SAMA) v. 3.1," *LDC Catalog no.LDC2010L01.ISBN,* pp. 1-58563, 2010.

[14] A. Pasha, M. Al-Badrashiny, A. E. Kholy, R. Eskander, M. Diab, N. Habash, M. Pooleery, O. Rambow and R. Roth, "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic," in *In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland,* 2014, .

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter,* vol. 11, pp. 10-18, 2009.