Arabic Corpora for Credibility Analysis

Ayman Al Zaatari, Rim El Ballouli, Shady Elbassuoni, Wassim El-Hajj, Hazem Hajj, Khaled Shaban[†], Nizar Habash[‡], Emad Yehya

American University of Beirut, Lebanon, [†]Qatar University, Qatar, [‡]New York University Abu Dhabi, UAE

{abz02, roe04, se58, we07, hh63, ehy00}@aub.edu.lb, [†]khaled.shaban@qu.edu.qa, [‡]nizar.habash@nyu.edu

Abstract

A significant portion of data generated on blogging and microblogging websites is non-credible as shown in many recent studies. To filter out such non-credible information, machine learning can be deployed to build automatic credibility classifiers. However, as in the case with most supervised machine learning approaches, a sufficiently large and accurate training data must be available. In this paper, we focus on building a public Arabic corpus of blogs and microblogs that can be used for credibility classification. We focus on Arabic due to the recent popularity of blogs and microblogs in the Arab World and due to the lack of any such public corpora in Arabic. We discuss our data acquisition approach and annotation process, provide rigid analysis on the annotated data and finally report some results on the effectiveness of our data for credibility classification.

Keywords: Credibility, Blogs, Twitter, Crowdsourcing.

1. Introduction

The increasing popularity of social networks and blogging websites has transformed the Web into a dynamic, fast-paced and user-centered platform for sharing information; which is commonly referred to now, as the Social Web 2.0. For instance, there are more than 500 million tweets generated daily on Twitter[1]. Similarly, there are more than 54.2 million blog posts and 52.3 million comments generated every month on just one popular blogging website[2].

This immense amount of data generated on Twitter and Blogging sites has become a vital and rich source for opinion mining tasks such as sentiment analysis, pro/con classification and emotion recognition. With such a large scale of generated data, it is inevitable that the credibility of the generated data would highly vary. We adopt the Merriam Webster[3] definition of credibility that states: *credibility is the quality of being believed or accepted as true, real or honest*. In other words, a credible (micro)blog is one which holds enough evidence to be believed or accepted as true, real or honest. The presence of non-credible information can highly influence the accuracy of the tasks performed on such data and hence filtering out non-credible data would be very beneficial.

To predict the credibility of (micro)blogs, supervised machine learning can be deployed. First, a corpus of (micro)blogs must be constructed. This corpus must then be annotated for credibility (i.e., each (micro)blog in this corpus must be labeled as either credible or not, using human judges). This annotated corpus will then play the role of a training dataset (or ground truth), which can be used to build automatic credibility classifiers that can accurately predict the credibility of a given (micro)blog.

In this paper, we describe the process of creating two corpora for credibility analysis and we validate their usefulness. The first corpus consists of 175 Arabic blog posts in which each blog was manually labeled as being credible, fairly credible or non-credible by a number of human judges. Additionally, each blog was annotated for another set of features that relate to credibility. These additional features include: *reasonability*, *bias*, *sentiment*, and *objectivity*. The second corpus consists of 2,708 Arabic tweets, which again were manually labeled as either credible or non-credible. Our two corpora are the first publicly available corpora for Arabic tweets/blogs that are annotated for credibility, and can be downloaded from the resources at www.oma-project.com[4].

2. Related Work

We broadly classify the research done on credibility into: work done on blogs and that done on Tweets. Most of the developed corpora are for English content. Credibility of Arabic content has not received profound attention from researchers and as such, this area has a lot of room for improvement. In what follows we present the details of the developed corpora for credibility analysis.

For tweet credibility, most of the available corpora such as (Castillo, Mendoza, & Poblete, 2011; A. Gupta & Kumaraguru, 2012; A. Gupta, Kumaraguru, Castillo, & Meier, 2014) are in English as mentioned above. The only exception is (R. Al-Eidan, Al-Khalifa, & Al-Salman, 2010) where the authors collected 600 Arabic tweets and 179

[1] https://business.twitter.com/en-gb/basics
[2] https://wordpress.com/activity/
[3] http://www.merriam-webster.com/

[4] http://www.oma-project.com (under Resources)

Arabic news articles and used cosine similarities between the tweets and articles as a measure of credibility. However, this work relies on automatic annotations rather than human annotations, which is the best known method to generate ground truth for supervised learning.
For blog credibility, the authors in (Schwarz & Morris, 2011) collected 1000 webpages related to various topics using a popular search engine. Then, they asked a single member in their research group to rate the credibility of those 1000 articles on a 5 Likert scale. However, this corpus is in English, and was rated by a single person only, which makes it inapplicable in our case.
In (Weerkamp & de Rijke, 2012), the authors used the Blog TREC corpora collected over the years 2006, 2007 and 2008, to study credibility of blogs. However, the Blog TREC corpora and some later corpora also developed by TREC do not have a sufficient number of Arabic blogs. In (Juffinger, Granitzer, & Lex, 2009), the authors collected a very small set of 40 hand-picked blogs evenly distributed over several topics and languages and then annotated it for credibility. However, this dataset is very small to be used to build any robust automatic credibility classifier.

3. Data Collection

3.1 Tweet Collection

We used Twitter stream API to collect over 36 million tweets in a period of two weeks, which were then indexed using Apache Lucene (Białecki, Muir, & Ingersoll, 2012). In addition, we monitored the daily Arabic news and identified a set of highly discussed news topics, which we then used as queries over our tweet dataset. Table 1 provides the news topics we identified and a brief description of each. The count of retrieved tweets for each topic from our dataset

Query topics	Description	Tweet count
<u>قوات النظام</u>	<u>The forces of the Syrian government</u>	1791
<u>الثورة السورية</u>	<u>Syrian revolution</u>	1232
الأزمة السورية	Syrian problems and concerns related to the Syrian revolution	297
الإنتخابات الرئاسية في لبنان	The election of Lebanese president	38

Table 1: Description of trendy Arabic topics and count of retrieved tweets for each topic

can also be found in Table 1. The two topics with underlined text are the ones whose relevant tweets were annotated and later used in our experiments. We picked these two topics only as they contained the largest number of relevant tweets from our tweet dataset. In addition, we disregarded the other two topics since they were too small to make any future analysis and deductions with respect to topics.

3.2 Blog Collection

For Arabic blog credibility, 175 Arabic blog posts were collected by issuing queries to Google (Blogs) Search engine and handpicking blog posts based on relevance to the query and content type (blog posts with bare opinion and news content). The search queries included trendy topics at the time of data collection. The search queries issued to the search engine and respective blog counts from each are shown in Table 2. Our dataset is composed of 90 purely news articles and 85 articles reflecting the author's opinion on various events and topics.

Query topics	Description	Blog count
إنتخابات رئيس الجمهورية في لبنان	Lebanese President Elections	89
حكومة المصلحة الوطنية	Lebanese Parliament elections and related issues	29
الأزمة السورية	Syrian Crisis	19
المحكمة الدولية الخاصة بلبنان	Special Tribunal for Lebanon	19
أيفون 5 اس , نوت 3 سامسونج	Iphone 5s, Samsung Note 3	7
2014 كأس العالم	FIFA world cup 2014	4
مواضيع مختلفة	Other Topics	8

Table 2: Description of trendy events/topics and count of retrieved blogs for each topic

4. Corpus Annotation

While a tweet is a single sentence with a maximum of 140 characters, a blog is a long sequence of sentences. Additionally, a tweet belongs to a well-identified author whose previously published tweets can be easily retrieved; on the other hand, especially in the case of Arabic blogs, many articles are anonyms, or include the author name only without providing a link to the full user profile or previous posts. Therefore, the different nature of both mediums dictates different cues for credibility judgment and hence a different annotation scheme. For example, judging the credibility of a tweet using only its text is not enough; instead one may additionally rely on the author background, expertise and external web references. On the other hand, a blog post might contain enough cues in its text to assess its credibility.
Based on the above mentioned distinctions, the two corpora (one for tweets and another for blogs) were separately

collected and annotated. We describe in details the annotation process for each corpus next.

4.1 Tweet Annotation

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 the annotation process is and what cues they relied on in their credibility judgments. After multiple iterations and refinement over the annotation interface, we provided for each tweet, 1) 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. 2) the complete author profile 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. 3) the results of a Google search on the tweet text. The Google search was restricted to the time of the tweet creation ±5 days. Our interface left a room for annotators to decide on what links to visit and what information to rely on when deciding on the credibility of a tweet. After reading and analyzing a tweet, annotators were asked to label it as either "credible" or "non-credible". They were also given the option to select "can't decide" when they felt confused or unsure. To that end, we built a custom web-based annotation tool (http://twitter.me-applications.com/) where users logged in and completed their annotations.

To ensure reliable annotations, a tutorial session was completed by five annotators including a qualification test at its end. Three annotators passed the test and were recruited to complete the full task and received monetary compensation for their annotations. Furthermore, "gold tweets" and "repeated tweets" were injected into their annotation assignments to help us better assess the quality of annotations. All annotators passed our gold tweets and were generally consistent with their annotations across repeated tweets. Finally, we obtained an inter-rater agreement of 0.43 between our three annotators using *Fleiss' kappa*. While there is no precise rule for interpreting *kappa* scores, the work in (Viera & Garrett, 2005) suggests that such a *kappa* score translates to having a moderate agreement between the annotators.

Topic	Description	Credible	Non-credible
قوات النظام	The forces of the Syrian government	1131	510
الثورة السورية	Syrian revolution	439	628

Table 3: Distribution of credible and noncredible annotations for every query in the datasets

Table 3 shows the distribution of credible and non-credible tweets as annotated by our annotators for each of the two topics we picked. A majority vote was used to decide on the final labels of the tweets

4.2 Blog Annotation

A similar process was initiated for the blog corpus, where we asked the annotators to annotate Arabic blogs for credibility and for four additional features that we believed to have a great effect on credibility, namely: *reasonability*, *bias*, *objectivity* and *sentiment (Merriam Webster*[5] *was again used to define each feature)*. These features were identified based on a literature review (R. M. B. Al-Eidan, Al-Khalifa, & Al-Salman, 2009; Flanagin & Metzger, 2000; Fogg et al., 2003; Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, Harald Schoen and Peter Gloor, Daniel, Castillo, Mendoza, & Poblete, 2013; M. Gupta, Zhao, & Han, 2012; Metzger, 2007; Nakamura, Suzuki, & Ishikawa, 2013; Olteanu, Peshterliev, Liu, & Aberer, 2013; Ulicny, Baclawski, & Magnus, 2007), and by a pilot study executed by the research group. In this pilot study, 25 Arabic blogs were annotated for credibility and a large set of other features including the latter four. Finally, feature ranking based on Information Gain was performed on the feature set using the WEKA classification platform (Hall et al., 2009), which resulted in the four features mentioned earlier. Discussions with the research group members who participated in the pilot study also confirmed that those features helped them decide on the credibility of the blog posts the most.

We hypothesize that credible blog posts tend to present a highly reasonable content with a clearly justified stance on regards of the topic being discussed; without being overly biased to a group or party; with an objective presentation of facts; and in tone of writing far from being strongly positive or strongly negative, but rather neutral. Non-credible blogs posts on the other hand tend to miss a majority of these characteristics and therefore suffer from a drop in credibility.

Similar to tweet annotation, a custom-built annotation interface (http://annotate.me-applications.com/) was used for blog annotation, and a training session was held for a group of annotators before they started annotating the blog posts. All annotators were familiar with the blog topics in the corpus. In the training session, the problem was clearly explained, and credibility, reasonability, objectivity, bias and sentiment were all described separately, and example blog articles were also presented to better clarify each feature and its possible values. Afterwards, 50 participants completed their annotation tasks in a 4 weeks period on the

[5] http://www.merriam-webster.com/

web-based annotation interface, and received monetary compensation for their annotations. We collected at least 4 annotations from different annotators for each blog. For each article, annotators had to annotate for credibility first, then for sentiment, reasonability, bias and objectivity, each on relevant nominal scale. To ensure the reliability of the obtained annotations, "gold blogs" were injected-as in the case of tweets-and the duration of each blog post annotation per annotator was also saved seamlessly and later used to detect and remove effortless annotations (the annotators were not told about that). Finally, we obtained an inter-rater agreement of 0.3 using *Krippendorff's alpha* measure which is suitable in our setting as it handles the ordinal nature of our data, accepts missing values (the can't decide ratings), and works for any number of annotators. The achieved *k-alpha* score is acceptable as (Stoyanov & Cardie, 2008; Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010) suggest, since the features have several nominal levels and not numeric (distance between levels wasn't taken into consideration by the *k-alpha* algorithm which drastically reduces the *k-alpha* result), and the nature of opinion annotation task is hard and varies from person to person.

Our annotated corpus consisted of 175 blog posts with 100 posts annotated as credible (57.1%), 57 as fairly credible (32.6%), and 18 as non-credible (10.3%). A majority vote was used to decide on the final labels of the tweets. We also studied the correlation between each of the four features and the credibility scores in the annotated corpus. The correlations scores clearly show that credible articles tend to be reasonable, objective, not biased and neutral with a correlation score greater than 85%; and that non-credible articles lack reasonability, objectivity, contains clear author bias or negative tone with a correlation score greater than 75%. Detailed correlation for each feature was also computed and studied, and deductions were made on credibility annotation patterns between the annotators.

5. Case Studies

5.1 Tweet Credibility Classification

We use the annotated tweet corpus and an exhaustive set of features to evaluate the effectiveness of the corpus for credibility classification. The chosen features are extracted from both, the tweet itself and its author. We extracted 22 user-based features extracted directly or indirectly from the author's history (tweet timeline), for example, author expertise on the topic being discussed and rate of activity of the tweet author. In addition, 26 content-based features are extracted including sentiment, count of retweets, count of URLs etc. To extract the sentiment we tokenized the tweet using MADAMIRA (Pasha et al., 2014) to obtain the lemma for each word in the tweet. MADAMIRA is a morphological analysis and disambiguation tool for Arabic text.[6] Next, using the lemma of each word in the tweet we extract its corresponding positive, and negative score from the ArSenL lexicon (Badaro, Baly, Hajj, Habash, & El-Hajj, 2014). Finally, to compute the positive and negative score of the whole tweet we sum up the positive and negative scores for each word in the tweet respectively. Next, We trained our classifier, which will be called CAT (Credibility of Arabic Tweet) henceforth, using multiple machine-learning algorithms such as Naive Bayes, SVM and J48 Decision Tree, however, we only report the results of the highest attaining algorithm in terms of Weighted Average F-measure (WAF-measure), namely the Random Forest Decision Tree. The results are shown in Table 4 with the highest achieved metric values highlighted in bold. Using a 10-fold cross validation, our classifier achieved a WAF-measure of 69.5% and 78.8% when tested on topics 1 and 2, respectively. To check the performance of our classifier on diverse topics, we combined topics 1 and 2 into one set and re-did the classification using cross validation. CAT achieved a WAF-measure of 70.1% when classifying topics 1 and 2 combined.

We compared the performance of CAT to three common baselines. The first baseline is the Stratified baseline, where the classifier makes random predications in accordance to the distribution of credible and non-credible tweets in the training set. Hence, if the training set includes 80% credible and 20% non-credible tweets, the stratified baseline randomly predicts 80% of the test set to be credible and 20% to be non-credible. The second baseline is one that makes uniform predictions such that both credible and non-credible classes are equally likely. The third baseline is the

Topic	**Classifier**	**Weighted average Precision**	**Weighted average Recall**	**Weighted average F-measure**
Topic 1	CAT	69.6%	71.7%	**69.5%**
	Stratified	56.6%	56.8%	56.7%
	Uniform	59.0%	52.3%	54.1%
	Majority	47.5%	68.9%	56.2%
Topic 2	CAT	79.0%	79.1%	**78.8%**
	Stratified	50.5%	51.1%	50.7%
	Uniform	52.4%	50.8%	51.2%
	Majority	34.6%	58.9%	43.6%
Combined Topics	CAT	70.1%	70.3%	**70.1%**
	Stratified	33.6%	58.0%	42.6%
	Uniform	52.3%	51.2%	51.5%
	Majority	51.6%	51.8%	51.7%

Table 4: CATs' performance against baseline classifiers

[6] For more information on the challenges of Arabic natural language processing, see (Habash, 2010).

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 will be labeled credible. Table 4 presents the weighted average Precision, Recall, F-measure results for this comparison. CAT consistently surpassed the WAF-measure of the baseline approaches indicating that the user-based and content-based features we used are worthy indicators of credibility. When considering the highest WAF-measure of every baseline, CAT surpasses, on average, the baselines by 27%.

5.2 Blog Credibility Classification

Similarly, we build a credibility classifier for blogs (CAB) using cross validation to check the appropriateness of our dataset. In the table 5, we show the results from both Decision Table and Naïve Bayes classification algorithms when trained and tested on the original dataset (first column), and when trained and tested on the balanced dataset (second column) after applying the SMOTE filter (Synthetic Minority Over-sampling Technique), which is a filter used to overcome classes in balance in datasets. As observed in the table, the two classification algorithms had close results in both configurations, with 73% accuracy and 72% weighted average F-measure showing a good classification result. However, non-credible articles were few compared to credible and fairly credible articles, which biased the classification towards the latter 2 classes. This was resolved in the balanced set (second column), achieving a higher accuracy and F-measure, with values equal 80%. We also compared the classifiers' results to several baselines classifiers like Stratified classifier, Uniform Classifier, and Majority class classifier. CAB with Decision table and Naïve Bayes, as can be seen in Table 5, showed much better results than all the baselines.

Classifier	Unbalanced data		Balanced data	
	Accuracy	**F-Measure**	**Accuracy**	**F-Measure**
CAB (Decision Table)	**73%**	**72%**	**80%**	77%
CAB (Naïve Bayes)	71%	70%	**80%**	**80%**
Stratified	30%	32%	34%	36%
Uniform	31%	34%	34%	36%
Majority	57%	42%	57%	42%

Table 5: Classification results on full dataset

6. Conclusion and Future Work

In this paper, we presented two Arabic corpora for credibility analysis, one composed of 2708 Tweets[7] and one of 175 Blog posts[8]. To verify the usefulness of our corpora we built machine-learning models using an exhaustive list of features, and while utilizing the annotated corpora for training and testing. We also analyzed the correlation between credibility and other features that affect it positively or negatively. Our annotated corpora are the first of their kind and will serve as a valuable resource for future studies on the credibility of Arabic content.

7. Acknowledgements

This work was made possible by grant NPRP 6-716-1-138 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

8. Bibliographical References

Al-Eidan, R. M. B., Al-Khalifa, H. S., & Al-Salman, A. S. (2009). Towards the measurement of arabic weblogs credibility automatically. *Proceedings of the 11th International Conference on Information Integration and Web-Based Applications & Services*, pp. 618-622.

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

Badaro, G., Baly, R., Hajj, H., Habash, N., & El-Hajj, W. (2014). A large scale arabic sentiment lexicon for arabic opinion mining. *ANLP 2014, 165*

Białecki, A., Muir, R., & Ingersoll, G. (2012). Apache lucene 4. *SIGIR 2012 Workshop on Open Source Information Retrieval*, pp. 17-24.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th International Conference on World Wide Web*, pp. 675-684.

Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly, 77*(3), 515-540.

Fogg, B., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., & Tauber, E. R. (2003). How do users evaluate the credibility of web sites?: A study with over 2,500 participants. *Proceedings of the 2003 Conference on Designing for User Experiences*, pp. 1-15.

Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, Harald Schoen and Peter Gloor, Daniel, Castillo, C., Mendoza, M., & Poblete, B. (2013). Predicting information credibility in time-sensitive social media. *Internet Research, 23*(5), 560-588.

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

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). TweetCred: A real-time web-based system for assessing

[7] http://oma-project.com/res/tweet_corpus
[8] http://oma-project.com/res/blog_corpus

credibility of content on twitter. *arXiv Preprint arXiv:1405.5490,*

Gupta, M., Zhao, P., & Han, J. (2012). Evaluating event credibility on twitter. *SDM,* pp. 153-164.

Habash, N. (2010). Introduction to Arabic Natural Language Processing, Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18.

Juffinger, A., Granitzer, M., & Lex, E. (2009). Blog credibility ranking by exploiting verified content. *Proceedings of the 3rd Workshop on Information Credibility on the Web,* pp. 51-58.

Metzger, M. J. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology, 58*(13), 2078-2091.

Nakamura, A., Suzuki, Y., & Ishikawa, Y. (2013). Clustering editors of wikipedia by editor's biases. *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences On, , 1.* pp. 351-358.

Olteanu, A., Peshterliev, S., Liu, X., & Aberer, K. (2013). Web credibility: Features exploration and credibility prediction. *Advances in information retrieval* (pp. 557-568) Springer.

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., et al. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. *In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland,*

Schwarz, J., & Morris, M. (2011). Augmenting web pages and search results to support credibility assessment. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 1245-1254.

Stoyanov, V., & Cardie, C. (2008). Annotating topics of opinions. *LREC,*

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544-2558.

Ulicny, B., Baclawski, K., & Magnus, A. (2007). New metrics for blog mining. *Defense and Security Symposium,* pp. 65700I-65700I-12.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Fam Med, 37*(5), 360-363.

Weerkamp, W., & de Rijke, M. (2012). Credibility-inspired ranking for blog post retrieval. *Information Retrieval, 15*(3-4), 243-277.