

AMERICAN UNIVERSITY OF BEIRUT

MINING FOR CREDIBLE OPINIONS IN ARABIC BLOGS

by  
AYMAN BASSAM AL ZAATARI

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  
February 2017

AMERICAN UNIVERSITY OF BEIRUT

MINING FOR CREDIBLE OPINIONS IN ARABIC BLOGS

by  
AYMAN AL ZAATARI

Approved by:



Dr. Wassim El Hajj, Chairperson & Associate Professor  
Department of Computer Science

Advisor



Dr. Shady Elbassuoni, Assistant Professor  
Department of Computer Science

Member of Committee



Dr. Hazem Hajj, Associate Professor  
Department of Electrical and Computer Engineering

Member of Committee

Date of thesis/dissertation defense: 3<sup>rd</sup> of February, 2017

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT RELEASE FORM

Student Name:

Al Zaatari  
Last

Ayman  
First

Bessam  
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.

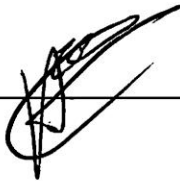
I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes after:

**One --- year from the date of submission of my thesis, dissertation, or project.**

**Two --- years from the date of submission of my thesis, dissertation, or project.**

**Three --- years from the date of submission of my thesis, dissertation, or project.**

Signature



13-02-2017

Date

This form is signed when submitting the thesis, dissertation, or project to the University Libraries

## ACKNOWLEDGMENTS

أحمد الله الكريم المنان أولاً وأخراً أن وقّفتي وأعانني على إنجاز رسالتي, وأسأله الأجر والثواب عليها, وأن يكون فيها النفع والخير.

ثم أتقدم بالشكر لو الذي الذّين صباراً عليّ ودعامنني حتى وصلتُ إلى ما وصلت إليه, فأحيل شكرهما الله عز وجل. وها أنذا أمنحهما رسالتي, ليمنحاني الرضى والدعاء.

وكذا أشكر مستشار أطروحتي د. وسيم الحاج, الذي كان يعملني كابنه طول فترة الماجستير, وكان دعمه دائماً يدفعني إلى الأمام, وإن تعثرت كان بجنبي ليقمني.

وأشكر د. شادي الباسيوني الذي تابعني في عملي, وطالما وجهني حتى أنهيت رسالتي. وكما أشكر د. حازم الحاج الذي طالما كان نقده بناءاً لمصلحتي ومصلحة أطروحتي ونجاح عملي.

وأشكر ريم البللولي التي كانت مساعدتها أساسية في عملي. وشكر خاص لشادي الحلوة, الذي كانت مساعدته لي في أحلك الظروف وأشدّها, فأشكره على ذلك وعلى طول سهره معي في العمل على أطروحتي.

وأخيراً وليس آخراً, أشكر كل معلمي واساتذتي ومشايخي الذين صبروا على تعليمي منذ صغري, حتى بلغت ما قد بلغت, وأجمل معهم كل من كان له يد في عملي هذا

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

## AN ABSTRACT OF THE THESIS OF

Ayman Bassam Al Zaatari for Master of Science  
Major: Computer Science

Title: Mining for Credible Opinions in Arabic Blogs

Blogging websites are growing globally, allowing online users to express their views and engage in discussions related to various domains such as politics, technology, entertainment, and lifestyle. Posted blog entries often reflect their authors' trustworthiness, quality, authority and believability, which vary from one author to another. While some blog posts state facts, others tend to spread rumors, state personal views, or support certain propagandas. **The aim of this work is to create models to automatically rate the credibility of Arabic blog posts in real-time**, adopting the Merriam Webster credibility definition: "*the quality of being believed or accepted as true, real or honest*".

We focus on Arabic blog posts due to their recent popularity fueled by the recent uprisings in the Arab world, and due to the scarcity of tools for assessing the credibility of Arabic blog posts. We note that Arabic Natural Language Processing (NLP) is challenging due to the natural complexity of the Arabic language and its very rich morphology, unavailability of benchmark corpora, and immaturity of its NLP tools compared to those available for English and other languages. To achieve our objective, we first compiled a set of credibility features from literature, and added other features that we believe affect the credibility of Arabic blog posts. We then selected from the web 25 Arabic blog posts, extracted these features, and annotated the posts for credibility. Afterwards, we applied feature selection, and reduced the feature space to the four features that affected credibility the most, namely: *reasonability*, *bias*, *objectivity*, and *sentiment*. Having selected the features of interest, we annotated a manually collected medium-size corpus of 273 Arabic blog posts, and created several classification models including SVM, Neural Nets, Decision Trees and others, among which we ended up using Decision Trees which achieved 74 % accuracy and F-measure score, and a 10% increase on those scores (84%) when we tested

the models on a golden subset of the blog posts i.e. the blog posts that received full annotation agreement amongst the annotators.

Finally, for full system automation, we used MadaMira -an Arabic morphological analyzer- to extract Arabic linguistic features for each document, and included them in the training of several deep learning models, such as Long Short-Term Memory (LSTM) and Gated Recurrent Neural Networks (GRNN), resulting in 63% F-measure on the full set, and 74% on the golden set. This F-measure drop is mainly due to the small size of the data set we used, with 268 documents which is considered short for building deep learning models. If trained on a larger corpus, it is certain that the deep models will produce results comparable to the ones achieved above, if not higher.

# CONTENTS

ACKNOWLEDGMENTS.....	II
AN ABSTRACT OF THE THESIS OF.....	V
ILLUSTRATIONS .....	X
TABLES .....	XII

## Chapter

INTRODUCTION.....	1
1.1    Motivation.....	2
1.2    Problem Statement .....	6
1.3    Objectives and Contribution.....	7
1.4    Thesis Plan .....	7
LITERATURE REVIEW.....	9
2.1    Blogosphere.....	11
2.2    Web Credibility.....	12
2.2.1 Website Credibility Summary .....	18
2.3    Blog Credibility.....	19
2.4    Credibility in Other Web Content Types .....	22

<b>SUMMARY OF FINDINGS FROM LITERATURE .....</b>	<b>24</b>
2.5    Definition.....	24
2.6    Features .....	24
2.7    Methods.....	24
2.8    Corpora.....	25
<b>PROPOSED APPROACH .....</b>	<b>26</b>
3.1    Small-Scale User Study.....	28
3.1.1 Small Corpus Construction.....	28
3.1.2 Start-up features.....	29
3.1.3 Small Annotation.....	31
3.1.4 Annotation Evaluation.....	32
3.1.5 Feature Selection .....	35
3.1.6 Findings .....	38
3.2    Large Scale Annotation.....	39
3.2.1 Large Corpus Collection.....	40
3.2.2 Features.....	40
3.2.3 Annotation .....	44
3.2.4 Annotation Results and Analysis.....	46
3.3    Feature Extraction .....	48
3.3.1 Reasonability Exploration .....	49
3.3.2 Objectivity Exploration .....	49
3.3.3 Bias Exploration .....	49



3.3.4 Sentiment Exploration .....	50
3.3.5 Extraction Tools .....	50
<b>PERFORMANCE EVALUATION .....</b>	<b>51</b>
4.1 Machine Learning Algorithms vs Baselines .....	51
4.2 Deep Learning Models .....	54
4.3 Discussion .....	63
<b>FUTURE WORK .....</b>	<b>66</b>
<b>CONCLUSION .....</b>	<b>67</b>
<b>REFERENCES .....</b>	<b>68</b>

## ILLUSTRATIONS

Figure	Page
Figure 3-1 Arabic Blog Credibility prediction System .....	27
Figure 3-2 Small Scale Classification Results.....	35
Figure 3-3 Modals Accuracies when the 4 features were only used .....	39
Figure 3-4 – Credibility and sentiment annotation. ....	45
Figure 3-5 – The credibility features annotation .....	45
Figure 3-6 Classification results on full data set.....	47
Figure 3-7 Classification results on golden data set .....	48
Figure 4-1 Classification Models accuracies on All documents as input, comparing all features and reduced features inputs .....	52
Figure 4-2 Classification Models accuracies on Golden documents as input, comparing all features and reduced features inputs .....	52
Figure 4-3 Classification Models Accuracies on full feature space .....	53
Figure 4-4 Classification Models Accuracies on reduced feature space .....	54
Figure 4-5 LSTM Model.....	56
Figure 4-6 LSTM + Discrete Features Model .....	57

Figure 4-7 CNN + LSTM Model.....	58
Figure 4-8 CNN + LSTM + Discrete Features Model.....	59
Figure 4-9 LSTM vs CNN + LSTM Accuracies .....	60
Figure 4-10 The 4 credibility features prediction in intermediate level .....	61
Figure 4-11 Accuracies on original credibility levels vs on collapsed levels.....	62
Figure 4-12 Accuracy measure per topic, with corpus size differences .....	63

## TABLES

Table	Page
Table 3.1 List of features used for credibility assessment .....	29
Table 3.2 Self-describing features .....	30
Table 3.3 Classification results of the original and (SMOTE-ed) balanced datasets using different classification models .....	33
Table 3.4 Information Gain scores of each feature.....	36
Table 3.5 Best Subset of Credibility features using CfsSubsetEval and Information Gain .	37

## LIST OF ABBREVIATIONS AND ACRONYMS

NLP: Natural Language Processing

MSA : Modern Standard Arabic

SMOTE: Synthetic Minority Oversampling Technique

WEKA: Waikato Environment for Knowledge Analysis

## DISCLAIMER

Everything published and produced in this thesis was made for research purposes. The authors are not to be held accountable for any misuse of this research work. Any results produced by the tools in this thesis are not guaranteed by the authors to be correct. Use on your own responsibility.

# CHAPTER 1

## INTRODUCTION

The increasing popularity of social networks such as Facebook and Twitter has transformed the Web into a dynamic, fast-paced and user-centered platform for sharing information, what is commonly referred to now as the Social Web 2.0. Blogging websites were among those social sites, acting as a platform where Web users can express their views and engage in discussions about various topics including: technology, sports, religion, politics, nature, food, education, entertainment, lifestyle, and many others. With their widespread and ease of access from all internet users, millions of blog posts are being posted monthly, inducing knowledge into it's audiences, and affecting their behaviour. For example, blogs played a big role as a means of engagement in the recent uprising in the Arab world; what is commonly referred to as the Arab Spring. However, this public platform made open to all internet users lacks experts' inspections to assure that the published content is truthful; Moreover, blog author identity can be made anonyms, or even faked in some cases. In fact, in most cases, anyone can publish anything without any obstacle.

In this work, we focus our efforts on Arabic blog posts, tackling the problem of automatically predicting the credibility of Arabic blog posts. We explore many machine learning algorithms and several deep learning models to build our fully automated Arabic

blog post credibility prediction system. To our knowledge, this is the first fully automated system for this task on Arabic posts.

## **1.1 Motivation**

With this huge online platform available for public internet users, posting articles in blogs have been made very easy and accessible for all. The result of that is a vast amount of blog articles being posted daily from internet users of variant ages, education levels, languages, interests, and motivations. Some bloggers might use blogs as personal diaries, others for information seek or spreading news; however, many others might use blogs to promote certain products, spread biased news for political purposes, and include hate speech, scam, or support false information for commercial or political motives. Moreover, it is easy for bloggers to publish articles with anonyms or fake identities, and therefore, internet users must be very cautious by reading blog posts deeply, comparing information to another sources, and have objective judgments before they admit to the information they find in blogs. Therefore, it would be of high importance to provide internet users with unbiased, objective, and truthful information found online about a given topic. To that end, many researchers did user studies and experiments on various information e-sources to produce assessment tools and methodologies which can help users evaluate the credibility of the information they find online.

For example, consider a controversial topic such as the special tribunal for Lebanon (المحكمة الدولية). There are a large number of posts published on this topic ranging from mere news describing the events surrounding the trial to detailed juridical reports about the trial itself. However these posts vary highly in their quality in terms of readability,



subjectivity, biasness, cohesiveness and reasonability, among others. Moreover, some of these posts are posted by experts while others are posted by fanatics or politically charged groups. It will thus be very useful to automatically extract these properties and to use them in assessing the credibility of a blog post. Credible posts should then be ranked higher or given higher weights before presenting them to users interested in finding the different opinions about a certain topic.

For example, consider the following credible blog post:

محكمة لبنان: أداة أو نتيجة للفتنة السنية الشيعية؟

من المتوقع، أن تبدأ المحكمة الجنائية الدولية الخاصة بلبنان، أولى جلساتها في السادس عشر من شهر كانون الثاني الجاري، مع بعض الاشكاليات القانونية المتعلقة بضم ملف المتهم الخامس الى القضية. ولقد تزامن هذا الحدث مع "احتفالية" اعلانية في المحكمة، وجولة اعلامية للناطق باسم المحكمة مارتن يوسف لتسويقها داخل لبنان، ودخول اسرائيلي اعلامي على المشهد المتعلق بالمحكمة

وما يهمنا في كل هذا أمران

الأمر الأول- إعلان مارتن يوسف بأن "المحكمة غير مخوّلة اتهام حزب أو فريق أو جهة سياسية معينة بل الأفراد المنتمين للأحزاب وتحصر اهتمامها بالأدلة التي تثبت علاقتهم بالاعتداء". بالمبدأ، ما يقوله يوسف صحيح من ناحية أن المسؤولية في المحاكم الجنائية الدولية - ومنها محكمة لبنان- هي مسؤولية فردية وليست جماعية، وبالتالي هي توجّه لأفراد وليس لشخصيات معنوية. لكن يوسف لم يقل كل الحقيقة، وتعامى عن مبدأ أساسي طورته المحكمة الجنائية الخاصة ببوغسلافيا، وقد ادرجه نظام محكمة لبنان في مادته الثالثة، وهو مبدأ المشروع الجنائي المشترك، والذي يُعتبر بدعة في القانون الجنائي الدولي

وبموجب هذا المبدأ الذي أكدت عليه المادة الثالثة من نظام محكمة لبنان، يمكن للمحكمة - حين تثبت الحكم

على عناصر من حزب الله- أن تلجأ الى التوسع في التحقيق لتتهم من تشاء ضمن المجموعة التي ينتمي اليها هؤلاء، على أساس الجرم الجنائي بموجب الهدف المشترك، وحتى لو لم يكن هذا الهدف المشترك جرمياً بالأساس. ونلاحظ أن محكمة يوغسلافيا قامت- من ضمن هذا المبدأ- بالحكم جنائياً على العديد من الأشخاص لمجرد تشاركتهم بـ "النية الذهنية" مع الجاني، أو لأنهم انتموا الى نفس المجموعة التي انتمى اليها الجناة

الأمر الثاني الذي يجب لفت النظر اليه، وهو في الفقرة الثانية ضمن المادة ٣ ايضاً من نظام محكمة لبنان، أنه يمكن للمحكمة انطلاقاً من مبدأ مسؤولية الرئيس عن المرؤوس أن تطلب تحويل السيد حسن نصرالله وسائر قيادات حزب الله، سواء بسبب الفعل أو عدم الفعل، وهذا يعني أن يُقدموا الى المحاكمة في حال كانوا يعلمون بالجريمة وأمرها بها، أو أنهم علموا ولم يمنعوا ارتكابها، أو ببساطة لأنهم كان من المفترض أن يعرفوا ولم يعرفوا، أو لم يبلغوا السلطات المختصة

الأمر الثاني والأخطر، هو الدخول الاسرائيلي على المشهد من خلال بابين: أولاً تسريب إحدى الوثائق المتعلقة بالمحكمة، والثاني الاعتراف بأن اسرائيل كانت قد زودت المحكمة بأدلة تورط قادة من حزب الله بقضية اغتيال الحريري

أ- بالنسبة لتسريب ونشر وثيقة من الوثائق الرسمية التابعة للمحكمة، والادعاء بأن أحد الباحثين الاسرائيليين قد "عثر" عليها، فهذا يشير الى أمور عدّة أهمها

محاولة اسرائيل تثبيت الاتهام على حزب الله اعلامياً حتى قبل تثبيته قضائياً وشحن الأجواء المذهبية في ظل الحرب السنية الشيعية الدائرة في المنطقة. بالإضافة الى تشويه صورة المحكمة أكثر مما هي مشوهة بالأساس، وفقدانها ما تبقى لها من مصداقية هشة. ولعل عبارة "عثر عليها"، تشير الى أن ابسط معايير العدالة لم تطبق في هذه المحكمة، فكيف يمكن لوثائق مصنّفة سرّية، وتدخل في صلب القرار الاتهامي بأن يعثر عليها باحث وكأنها مرمية في المهملات، أو موضوعة في أرشيف علني، وهذا لا يجوز اطلاقاً

ب- بالنسبة للإعلان الاسرائيلي بأن اسرائيل قد زوّدت المحكمة بالأدلة عن تورط عن حزب الله، فهذا

ليس بالأمر الجديد، فقد كشف مساعد بلمار في وقت سابق بأن اسرائيل هي مصدر من مصادر القرار الاتهامي، وذلك بالرغم من أن جميع المحققين كانوا قد أقرّوا بأن اسرائيل لم تتعاون مع المحكمة بالنسبة لما طلبه المحققين منها من معلومات. إذًا، اسرائيل قدمت الدليل الاتهامي، وهذا يطرح شكوكًا كبيرة على القرار الاتهامي بحد ذاته المبني على الاتصالات بين المتهمين، وهذا الاعتراف الاسرائيلي يصعب المهمة على المحكمة ويجعلها مطالبة بمزيد من الشفافية والصدقية، فكيف يمكن الوثوق بقرار اتهامي يدين حزب الله، قد قدمته عدوته اسرائيل، علمًا أن اسرائيل مدانة دوليًا باختراق الشبكة الخلوية والهاتفية اللبنانية

بكل الأحوال، من المتوقع أن يتم تأجيل جلسات المحاكمة مرة أخرى، وكما معظم المحاكم الدولية في العالم، ستكون المحكمة الدولية الخاصة بلبنان أداة بيد الدول الكبرى، تستخدمها في محاولة تغيير الموازين القوى في الداخل أو لسحق المهزومين، وبدون أن ينسحق أو ينهزم لن تستطيع أن تقوم بمهمتها تلك

As you can see, the author of this article took care of her writing eloquence, stated a clear thesis statement, included arguments in clear order, and supported her arguments by citing the Lebanese constitution, and stating facts, etc. Of course, those elements in her post would make it be very credible and authentic.

On the other hand, consider the following non-credible blog post:

المحكمة الدولية الخاصة بلبنان  
الم يظهر زيف تلك المحكمة المزعومة بعد ؟؟ انا اراها قد دفنت نفسها بنفسها !!! فكل كذبة سيأتي عليها  
يوما وتتكشف بشكل فاضح !! ولكن السؤال : لماذا نرى بعض اللبنانيين والسياسيين بشكل خاص يتمسكون بتلك  
المحكمة رغم معرفتهم الوثيقة بها وبالغرض الذي انشأت لاجله ؟ فهم امها وابوها والقائم مقامها !!! الم يستشعروا  
الخطر بعد ؟؟ الم يفكروا ولو قليلا بمستقبل هذا البلد الذي لم يرى الراحة والازدهار منذ زمن بعيد ؟؟؟ الا يعرف  
هؤلاء ان محكمتهم تلك يمكنها فعل كل شئ الا الوصول للحقيقة المزعومة ؟ الا تأخذهم رحمة ورفقة بتراب وطنهم ؟

كيف يمكنهم فعل ذلك بلبنان !!! ام انهم على ثقة من قدراتهم المادية ويعرفون مسبقا ان مكروها لن يصيبهم !!  
معتمدين على اموالهم وارصدتهم البنكية التي تتيح لهم مغادرة لبنان وتركه يحترق ويتمزق بفتنة طائفية تذهب به  
وبتاريخه؟! فتنة مذهبية هم اخبر الناس واعرفهم بخطورتها على لبنان؟! كيف يجروون؟؟؟ سؤال برسم المنطق!  
!فهل من مجيب؟!؟

Obviously, this blog post isn't at all as well-written as the previous one, contains no arguments, and is full of unanswered questions. Of course, one wouldn't want to see such content in search results for example.

## 1.2 Problem Statement

Arabic text credibility assessment hasn't come to a converging point yet, as we shall see in our literature review. Additionally, there hasn't been built any fully automated system that can deal with Arabic text and blogs, to assess credibility. Therefore, we aim in this thesis to explore the language based features of Arabic blog posts to determine the credibility of the beheld opinions, and to produce an automated tool to analyze and assess the credibility of the blog posts. We adapt the Merriam Webster credibility definition: the *quality of being believed or accepted as true, real or honest* [15]. Our work will focused on finding the best answers for the following two questions: ***“What are the features which most affect credibility in Arabic blog posts?”*** and ***“What is the best automated strategy to assess opinion credibility in Arabic blog posts?”***

### **1.3 Objectives and Contribution**

In our work, we focus on building a credibility classifier for content of Arabic blogs. In contrast with previous work that focuses on English (or other languages) and only proposes high level solutions for the Arabic content credibility problem, our work proposes a complete solution, and presents a fully automated system overcoming the challenging nature of the Arabic language for NLP in general. Our system relies on a set of morphological and syntactic features automatically collected and fed to a three-level classifier for credibility.

Our main contributions can be summarized as follows:

- A thorough literature review on credibility in general, and for blogs in specific
- A proposed set of features, and first to be, highly correlated with Arabic blog Credibility
- An Arabic Blog corpus annotated for credibility and related features, and first to be published on line for the research community
- A fully automated system for Arabic blog content credibility assessment

### **1.4 Thesis Plan**

The remainder of this thesis is organized as follows. CHAPTER 2 surveys existing work on credibility of content. CHAPTER 3 presents the methodology used and the steps

involved in the creation of the credibility classifier. First, we present the dataset collection process, and feature engineering step. Second, we present the details of the annotation process, where recruited workers manually label each blog for credibility and its related features. Third, we discuss the expansion of our data set, and finally the extraction of the suggested features. After that, we discuss the different models used for the credibility prediction task, and evaluate them in in 0. **Error! Reference source not found.** and **Error! Reference source not found.** conclude our work.

## CHAPTER 2

### LITERATURE REVIEW

Credibility of web content has been an interest of many researchers and an active research area for years. Although many papers have been published in this field, most previous work focused on one aspect of credibility or another. Some research was done on determining the factors affecting the user perception of credibility, while others used this to propose checklist for internet users that can be used to help them decide on the credibility of a given website. Another group of researchers compiled a list of features that can be used later in an automated manner to predict credibility, while others tried to extract those features from web pages and content and built tools that read the feature values and automatically decide on the credibility of the content.

Although the collective work of those research teams might seem complementary and have many common grounds, yet it can be differentiated by several aspects. One difference between the proposed models by different researchers is the content type each of them addresses when evaluating credibility, where some researchers addressed the credibility of the whole website in general, others addressed blogs in specific, some addressed the individual blog posts (individual articles as opposed to the blog, which is the collection of articles), some addressed author credibility, other researchers focused on twitter tweets and users in specific. In addition, some authors focused on specific domains in each of the above content types, including health, news, events, and others. Another difference between the researches done on credibility is the target language of the content,

which was majorly addressing the English language compared to like Arabic, German or other languages. Analysing English text in general has been addressed more in research since the web is majorly composed of English content, and the resources and tools available for English NLP are much richer and more mature than those available for non-English languages, making non-English text analysis very challenging and less targeted by researchers compared to that on English.

The definition of credibility was also a variant between one work and another, depending on the nature of the web content they are working with or the domain of the topics studied; however, most agree that credibility is a metric of truthfulness, trustworthiness and expertise.

Next, we provide quick history on blogosphere, followed by an overview of the most related work done on predicting credibility of web content and position our work with respect to it. We divide the spectrum of this literature review into 3 main categories:

1. Work on *website credibility* which focuses on predicting the credibility of the whole website
2. Work on *blog credibility* which is the closest to our work.
3. Work on predicting *credibility in other web content types* such as Twitter tweets or news.



## 2.1 Blogosphere

Blogs are web like journals in which authors publicly publish up-to-the-minute posts representing their personality, passions and point of views [1, 2] about various topics of interest. Blogs were first introduced into the web as a log of a list of links to internet sites surfed by Jorn Barger [3] and therefore was named “*weblogs*”. Later, the term was converted into “wee-blog”, and finally the term *blog* was used to describe the act of posting new articles on a *blog* which is the collection of published articles by a given author, aka. : blogger [4, 5]. Blogs are today’s journal paper, made online and in compliance with Social Web 2.0. There are several types of blogs published online, including: Personal blogs, business blogs, schools, non-profit organizations, politics, technology, religion, photography and art, fashion, sports, entertainment, fitness and health, gaming, and many others [6, 7] Blogs now not only allow bloggers to publish their content, and internet users to read, but also to interact with the published material by posting comments, sharing and liking. The number of published blogposts every day is estimated to be 2 Million [8, 9] and more than 2 Million comments, with a total above 400 million online blogs with the end of 2015[10-12]. microblogging (similar to Twitter ), photo blogs (similar to Flickr ), art blogs, podcast blogs and video blogs (aka. Vlogs, similar to YouTube ) [13, 14], each of which has revolutionized the blogosphere world in its own way. In this thesis, our target type will be the original text based Arabic blogs; other types of blogs required different approaches, and might be addressed later.

## 2.2 Web Credibility

One of the first major works done on credibility was the work done by BJ Fogg in [16-18]. In [16], the author studies dimensions of credibility in computers by dissecting it into 4 types: (1) *Presumed Credibility*, which describes the degree of believability the perceiver has in mind on content or the writer because of general assumptions he holds; this can be induced by the domain identifier liker *.gov* and *.edu*; (2) *Reputed Credibility*, which is the degree of credibility the perceiver acquires on the content or writer as a result of what third parties have reported about them; this can be induced by comments, or search engine rank and awards; (3) *Surface Credibility*, which is the credibility acquired after simple inspection; this can be induced by the website professional design; (4) *Experience Credibility*, which is acquired with first-hand experience and knowledge of topic or person. After that, the author revisions possible strategy for evaluating the credibility of a subject matter, and discusses three possible models: (1) *Binary Evaluation of Credibility*, which given a subject matter one of only two possible credibility evaluations: *credible* or *non-credible*; this strategy is usually used when there is low interest in subject, low ability to analyse the information, and no reference point for comparison; (2) *Threshold Evaluation of Credibility*, that includes upper and lower thresholds for credibility; this model is used in moderate cases in terms of interest and knowledge in subject; (3) *Spectral Evaluation of Credibility*, which is considered the most difficult and sophisticated strategy since it doesn't give a black or white result, but rather shades of grey of credibility; this strategy is usually used when users have high interest and familiarity with the subject, high ability to analyse the data with favourable cognitive and situational factor, and various sources to compare

with. In a different work [17], the same author proposes the *Prominence Interpretation Theory* in which he posits that two steps need to occur in order to make a credibility judgment on a given website: noticing something (*Prominence*), and then making judgment about it (*Interpretation*). The author considers credibility to be simply *believability* in essence, and agrees with the consensus of all others that it is a joint metric of trustworthiness and expertise. The author proceeds in his research by discovering the factors inducing prominence and those affecting the interpretation, and then uses this data to design a user experiment – in [18] – involving 2500 participants who were asked to rank the credibility of 100 websites presented as pairs based on a set of proposed features. The experiment results indicates that website design had the strongest effect on credibility, followed by information design/structure, information focus on a specific topic throughout the website, company motive, usefulness and accuracy of information present, website reputation, absence of advertisements, information bias, tone of writing, disclosure of site sponsor identity, functionality of site, customer service, past experience with the site, information clarity, readability and affiliations.

In [19], the authors, defining credibility as *believability*, suggested that users assess the credibility of a website by building impressions based on site appearance, usability of website and typographical mistakes first; then users evaluate the actually content by checking the reliability of both the source and the content itself. Finally, based on their current cognitive state, information need, and prior knowledge in the subject, users produce their credibility judgment.

The author in [20] adapted *believability of some information and/or its source* as the definition of credibility to review work done on web credibility by other authors and describe cognitive models for online information evaluation. Her extensive review shows that researchers generally found five main elements influencing credibility: (1) the accuracy and reliability of the information provided, and the absence of false information; (2) the authority of the website or website author as it would be author profile or organization description and affiliations; (3) the objectivity of the content by checking whether the content describes facts or opinions, and whether the website has some malicious intent; (4) the currency of the present data; and finally (5) the coverage and depth of the information provided on the site. In turn, those credibility elements can be studied and evaluated by checking a set of factors including: plausibility of arguments, professional quality and clear writing, source citations, presence of author profile and contact information, absence of advertisements, presence of privacy policies, professional website design, paid-access to the information on the website, and website domain and rank.

Augmenting on others' work, authors of [21] compiled a list of credibility features and prepared user experiments to study the impact of each on the end user credibility assessment for webpages. In their definition of credibility, they considered website credible "*if one can accept the information present there as true without looking elsewhere*". To launch their experiments, they first selected 5 main topics based on Open Directory Project ([www.dmoz.org](http://www.dmoz.org)) that would have a large amount of credible and non-credible material, namely: Health, Politics, Finance, Environmental Sciences and Celebrity news. Next, they built a query set of 5 trendy queries per topic, submitted it to a popular search engine and

collected the 40 top results from each query, building up a data set of 1000 URLs. After that, the 1000 URLs were annotated for credibility by an experienced member of the group on a five level Likert scale, and then a sample was cross-checked by another member, and an expert from each domain, achieving a correlation between the annotations high enough to consider the 1000 URLs annotation reliable. Then, they used the credibility feature values they have collected from each URL, and calculated the *Spearman's rho correlation* metric for each feature with credibility score to study the impact of each feature on credibility and identify the most important credibility features. Then, they used those results to run another experiment in which they asked a group of paid participants to assess the credibility of a set of webpages, with the credibility feature scores they collected augmented on the webpages in order to aim users in their credibility judgment, and computer the effectiveness of that on the confidence of the users about their credibility judgments. Their experiment results show that among: (1) *Off-Page features*: Awards, Alexa Rank, PageRank, had high correlation, while sharing had a lower correlation; (2) *On-page feature*: spelling errors, advertising, and domain type didn't have high impact on credibility; (3) *Aggregate features*: general and expert popularity (based on visits) along with geographic reach had high correlation with credibility, while dwell time (loading time), and re-visitation patterns had lower impact on credibility. In addition, the authors found other features like factual correctness, title, look and feel, and author information to have high importance and usefulness in the credibility assessment process.

Moving on to a new level, the authors of [22] proposed an automated model for assessing the credibility of the webpages corpus created in [21] by applying machine

learning algorithms on it. Similar to other approaches, the authors first compiled a large list of 37 candidate features they deemed to contribute to website credibility, and then used several measures (like *Spearman's rho*, *Chi-square test*, & *ANOVA*) to extract the best features. The credibility features they picked included *Content features* (text, appearance, & meta-info) and *Social features* (social popularity, general popularity, & link structure). The annotated corpus composed of 1000 URLs created in [21] was used to estimate the importance of each feature using the pre-mentioned metric, and the results show that the following stood out as the best features: (1) *Content features*: number of exclamations, number of questions, overall polarity of document, number of negative sentences, number of subjective sentences, informativeness of the page (a metric for the uniqueness of the page content compared to that of others in same domain), smog value (statistical measure of text readability difficulty), number of CSS definitions, number of adverbs, and the domain type; (2) *Social features*: *Facebook* social metrics count like share, likes, comments, and clicks, number of *tweets* mentioning the page, count of *bitly* clicks, number of delicious bookmarks, Alexa rank, and number of Alexa linkings, and Google PageRank. Other features collected by the authors but weren't shown to have a high correlation with credibility include: number of spelling mistakes, count of subjective and objective statements, count of nouns, verbs, adjectives and determiners, count of ads, and some others. After analysing the features, the authors inspected the plausibility of automating credibility prediction by experimenting with several machine learning algorithms “such as support vector machines (SVM), decision trees, and extremely randomized trees (ERT), and naïve bayes for classification; and SVM and ERT's variants for regression” by splitting their data set in an 80-20 training and testing cross-validation sets. ERT in both

classification and regression gave slightly higher results than other schemes which all had similar results. The obtained accuracy reached 75% for classification, and an improvement of 53% over the random baseline in regression. Besides accuracy, other metrics were also used for evaluation including precision, recall and F1-score. Those metrics showed good scores for *credible* webpages, but relatively lower for *non-credible* webpages indicating that automated classifier might be “optimistic and assesses *non-credible* pages as *credible*”. Evaluation metrics also gave evidence that “content features generally yield better performance”, and that specifically PageRank and Alexa rank scores from social features can also have high impact on credibility.

Another approach for evaluating website credibility and information quality was the checklist approach and iterative filtering based approach [23-26]. In their works, the authors discuss the possibility of supply internet users with a list of question on a checklist that they have to fill out as they surf a website, to help them decide on the website credibility. The checklists question mainly tried to verify currency, comprehensiveness (coverage), objectivity, accuracy, organization, citations and author profile and expertise [19, 26]. Although the checklist approach might seem to be a straightforward and easy method for evaluating web content, it was shown that checklists are usually lengthy, annoying to use for all websites, can be tricked by webmasters, has a lot of questions that are hard to answer and evaluate, and needs lots of training and practice[25, 27]. Therefore, it was an urge to propose better strategies for evaluating website credibility, without requiring users to have training sessions, and spend much time in the evaluation process.

### 2.2.1 Website Credibility Summary

The work discussed above serves as a snapshot of what is mainly done in literature on website credibility. We can see how different authors approached the credibility problem differently, like studying the dimensions of credibility and evaluation methods as discussed in [16], presenting the Human Computer Interaction perspective previewed in [17], and explaining the human credibility assessment strategy in [19]. Other research work executed experiments and studies to dig deeper into the problem and indicate the web features making pages credible, and finally provided wide titles for credibility features and shined the positive impact of exposing those features to humans for better credibility assessments of webpages [18, 20, 21]. Although these studies had major impact in mushrooming the research on credibility, it can only serve as fingers pointing to the direction the research should move through, and the dimensions of credibility researches should take care of when solving the web credibility assessment problem. Other researchers aimed to help users easily evaluate information credibility on the web by building checklists and applying filtering-approaches [23-26], but their work wasn't fruitful enough to solve the problem. More advance work was done when authors built a public corpus for annotated for credibility, and collected actual values for credibility features mentioned in literature [21], and most importantly attempting to automate the credibility classification of webpages by teaching machine learning algorithms [22]. Yet, we consider that the credibility problem isn't completed yet and needs more inspection and deeper studies. Blog credibility can have different features that need to be explored in addition to those indicated for website credibility in general. In our work we focus on content credibility alone and we



use most of the content-based features identified by previous work. We consider website credibility to be an orthogonal problem. We are aware that website credibility can affect the credibility of the content itself but we deem this to be easily incorporated once a content-based credibility prediction model is developed. Moreover, we are interested in predicting the credibility of Arabic blog posts and most of the Arabic blog sites lack many of the properties that deem them credible when compared to English ones. For instance, most Arabic blog sites contain a significant amount of advertisements, have poor design and lack author information. We thus believe that content-based features are the decisive factor in determining the credibility of Arabic blog posts. We also point out that work done above wasn't targeting Arabic content, and therefore might miss many Arabic specific features, and might also not be applicable for Arabic content analysis. Tools and language resources used were also English based, and alternative Arabic based resources must be either adopted or built.

### **2.3 Blog Credibility**

Next, we review the most relevant work that addressed credibility predication problem in blogs sites specifically, to discover models and strategies that can be helpful in solving our problem, predicating credibility in Arabic blog posts.

The authors of [28, 29] explored credibility features for blog authors (bloggers) by examining expert and average bloggers. They considered credibility to be “the perceived quality for someone being accurate and/or persuasive”. By the use of real world examples from online blogs, they verified that the use of reception features (in-link and out-link to blog using PageRank and HITS algorithm) isn't sufficient enough to accurately predict

credibility. The authors also noted that professional authors provided full name, affiliation, published genuine content, cited their references and received readers' comments in significantly higher percentages than average bloggers. However, the minority of online bloggers were considered professional bloggers, and therefore, most blogs are authored by average bloggers with low credibility indications, making the credibility assessment even harder. Therefore, the authors proposed a collection of features collected from (1) the Source and (2) the Message to make the blog credibility assessment plausible. The source features include: (a) exposure of identity, (b) posting location, (c) linking to a resume, (d) number of posts per month, (e) original text to ads ratio, and some others. On the other hand, message features include: (a) the number of original sentences in a post, (b) mention of a source and/or a URL, (c) the lack of spelling mistakes and profanity, (d) writing from personal experience, (e) inclusion of a documentary photograph or video, (f) mentioning proper names, among other features.

Credibility features for weblogs were also suggested in [30] that can be used as candidates for NLP and Machine learning credibility assessment techniques. Their credibility assessment technique included 4 major factors including: (1) Blogger's expertise and offline identity disclosure, (2) Blogger's trustworthiness and value system, (3) Information quality, and (4) appeals and triggers of a personal nature. The feature list of each factor is similar to those discussed in most of previous work.

The approach to blog credibility assessment was different in [31], where they decided to rely on factual truthfulness rather than purely on perceived judgment. To that end, they collected a German news corpus from the Austrian Press Agency

(<http://www.apa.at>) which they considered all its articles credible. Next, they collected a set of blogs for credibility assessment, and used the Dynamic Time Warping algorithm to align the different time series between the news corpus and the blog corpus, since their analysis indicated that a strong correlation was observed between the time series of news articles and that of blogs with high quality. This helped them “sort out blogs with a negative influence on the correlation and that are the blog with a completely different distribution over time”, however, it will keep blogs with wrong content at the right time. To solve this, they computed the centroid Cosine similarity of the tf-idf term vectors in the Vector Space Model, using nouns (to cover the thematic information of the blog posts of each blog) and another one using verbs and adjectives to cover the association within the topic; then, they used those values with some threshold based equations to rank blogs, and categorize them into “highly credible”, “unspecified/average credible” or “little credible” classes of credibility. When the system was tested on 14 blogs, the average precision yielded was 83%.

For Arabic blogs, the authors of [32] proposed a skeleton for a system that measures their credibility. The credibility definition was adapted from literature as *believability*, and was considered a perceived quality combining multiple dimensions such as trust, quality, authority, persuasiveness and popularity. Their credibility features were distributed on two levels: (1) Blog level (author identity and number of comments), and (2) Post level (spelling, emoticons, spamming, punctuations, post length, good/bad word, and similarity with verified content). Then they proposed a high level concept of a credibility system that basically started by extracting features from a collection of blogs, then they

would train and test the system, and finally build their classifier that would extract the features and decide on the credibility of the blog accordingly. The authors also pointed out several challenge in Arabic language processing like identifying proper nouns, finding writings with diacritics, the fact that the Arabic language is highly inflectional and derivational language, and that the tools available for Arabic NLP are still immature compared to those available for English. This however, the authors didn't implement any system or built any resources for it yet.

## **2.4 Credibility in Other Web Content Types**

Besides webpages and blog posts credibility, several work has been done on popular social networks like Twitter to assess the credibility of users and published content. In [33], the authors ranked Twitter user by their credibility which they predicted by their social network status and the content they published. Again, credibility was defined as “a combination of expertise and trust, supported by the nomination of other professionals”. First, they designed user experiments by asking people to evaluate users, which helped them identify the factors that affect user credibility on Twitter. This included various social status signals like followers count (users following the target user), followees count (the users that the target user follows), count of tweets and count list memberships. In addition, content signals were collected by applying LDA and tf-idf models on the tweets history which helps identify domains discussed in tweets and therefore, expertise. For classification, several models where used, and BetaBinary probabilistic distribution seems to have best results in ranking Twitter users based on their credibility.

Other works on Twitter platform addressed the credibility of tweets rather than Twitter users, similar to that of [34] where the authors' objective was to assess the credibility of tweets on time sensitive news events. To that end, they launched a user annotation experiment on a collection of tweets related to an actual earthquake event, where participants were asked to mark the truthfulness of those tweets on a 4 level scale. Then, they collected 68 features related to (1) the Message, (2) the User, (3) the Topic and (4) the Propagation of the tweet, and used them to build classification and regression models for automated tweet credibility prediction. Result show that Logistic regression, Random Forest, and Meta-Learning based on clustering performed the best with 70% accuracy, and sentiment scores, URL presence, user mentions, specific punctuation marks, emoticons and depth of propagation trees were the features affecting credibility the most, along with few others. Another work was done on Arabic tweets by [39], where a binary classifier was built making use of exhaustive set of features, and accomplishing a 75% accuracy.

Wikipedia was also targeted in literature where the authors of [35] in which they clustered Wikipedia editors based on their biases (positive, negative, or even) "to aid users judge the credibility of each description", by observing agreement and disagreement behaviours. They used this data to build clustered editors network graphs, and run analysis experiments which supported their hypothesis "that text that remain beyond many edits are credible", and therefore, allowing "users to refer to the credibility of each text".

## SUMMARY OF FINDINGS FROM LITERATURE

In this section, we summarize the findings from our literature review that had a useful effect on our work

### **2.5 Definition**

We adapted the Merriam Webster [15] credibility definition: the quality of being believed or accepted as true, real or honest. In our context, a credible blog is one which contains enough cues to appeal as authentic and trustworthy, and not the one that just states facts.

### **2.6 Features**

The nature of Arabic blogs gives us no choice but to drop author related features, since most of the time author profiles are not provided. Additionally, we are working with the text only, so we can't make use of domain related features. All what is left for us would be linguistic features.

### **2.7 Methods**

No complete work has been done on Arabic, so there isn't a specific method to adapt from previous work. Additionally, work done in English blogs makes use of features not available in Arabic blogs. But in general, most approaches collected a set of features, and used them in a classification model which they trained and tested for credibility prediction.

## **2.8 Corpora**

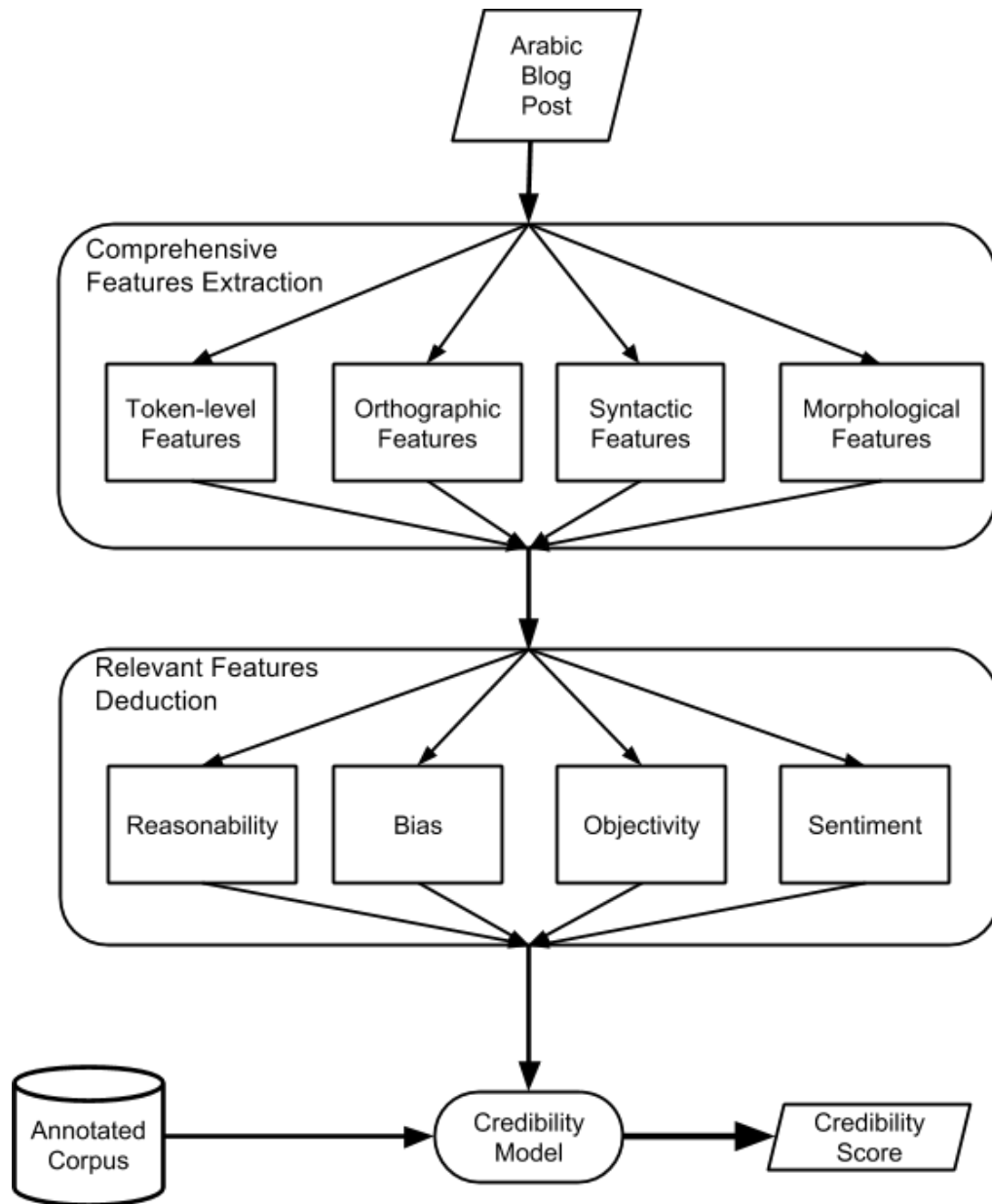
No Arabic corpora were found in literature.

## CHAPTER 3

### PROPOSED APPROACH

We intend to build a system for predicting credibility of Arabic blog posts as depicted in Figure 3-1. The input for the system would be an Arabic blog post which the user is interested in evaluating its credibility. The system would pass this blog post through a series of steps to extract the features needed for the final credibility assessment using NLP extensions for Arabic. Four types of feature sets are to be extracted varying in complexity and depth from very shallow token-level features, to orthographic, then morphological and syntactic features. The extracted features will then be used to deduce and generate a score for each of the four contextual features that affect credibility the most, namely: *reasonability*, *bias*, *objectivity*, and *sentiment*. Finally, a prebuilt Credibility model, trained on an Arabic blog post corpus annotated for credibility, will read the contextual features scores and predict the credibility score of the input Arabic blog post.





**Figure 3-1 Arabic Blog Credibility prediction System**

To build such a system, we conducted several experiments in several steps.

First we started by collecting an exhaustive set of possible credibility features.

Next, we built and annotated a small credibility corpus. Using this annotation experiment,

we were able to specific a small set of credibility features. Another time, we annotated a large corpus for the specified features. Finally, we explored different models to automatically extract those features and automatically predict credibility.

Next, we discuss the series details of each step on in the experiments we have done.

### **3.1 Small-Scale User Study**

After our extensive literature review, we intended to discover Arabic blog posts credibility indicators from live blogs, and put that align with the findings from literature. To accomplish that, we launched a small scaled user study involving a small set of Arabic blogs and a candidate set of features. The details of the user study are discussed next.

#### **3.1.1 Small Corpus Construction**

A small corpus composed of 25 Arabic blog posts handpicked based on relevance and content type was manually constructed by submitting queries to Google Blog Search Engine (<http://blogsearch.google.com>). The queries used to build the corpus were related to “*Lebanese politics*” and “*Technology*”, which were trending topics at the time (January 2014); the topics were specifically related to “*the special tribunal related to Lebanon*” (المحكمة الدولية الخاصة بلبنان) and “*iPhone 5 & Galaxy Note 3*” (أيفون ٥ والنوت ٣). All collected blogs were in Modern Standard Arabic (MSA), and were selected to contain bare opinions and news related to the mentioned topics.

### 3.1.2 Start-up features

We collected from literature several features correlating with credibility. Most of these features were mainly identified for non-Arabic blogs or webpages, and might include many irrelevant or unavailable features when it comes to Arabic blogs. We, next, list the features we used in our small scale user study in Table 3.1 based on our findings from the literature review, and include a description of each feature and the scale of possible values that can be assigned to it.

Upon our inspection of Arabic blogs, it seemed that a feature was affecting our credibility decision strongly which we didn't find in literature reviewed to that date; we call this feature "*Reasonability*", and use it as a measure of reasonable arguments and logical flow of points a blog post behold. The *Reasonability* feature was included in this small scale user study to analyse its importance to credibility by experiment.

**Table 3.1 List of features used for credibility assessment**

Feature	Definition	Scale
Author Expertise	Does the author have special skill and knowledge in the topic? This can be evaluated by checking the Author profile.	3 point scale (not-expert, amateur, expert) , Na otherwise
Author consistency	Is the author consistent in the post with his previous posts?	3 point scale (not-consistent, semi-consistent, consistent), Na otherwise
Comments Conformity	Does the author have a general support from the blog reader, or he is generally opposed?	3 points scale (doesn't confirm, mixed, confirm) , Na otherwise

Feature	Definition	Scale
Citations	Does the author provide citations and link to other information sources to back up his content?	3 points scale (none, low, high)
Objectivity	Does the author express or deal with facts or conditions as perceived without distortion by personal feelings, prejudices, or interpretations?	3 point scale (low-objectivity, medium objectivity, high objectivity)
Bias	Does the author show a tendency to believe that some people, ideas, etc., are better than others that usually results in treating some people unfairly	3 points scale (not-biased, fairly biased, biased)
Sentiment	What is the author attitude and tone of writing?	5 point scale (very positive, fairly positive, neutral, fairly negative, very negative)
Reasonability	How reasonable is the author in his judgments and arguments? Does he provide a logical reasoning for his stance and sentiment?	5 point scale (exaggerated, magnified, acceptable, sensible, reasonable)
Credibility	Do you consider the content of the article credible?	3 point scale (not credible, fairly credible, not credible)

In addition to those features, we list Table 3.2 an additional set of credibility features which, however, can be easily extracted from the blog without the need of human assessment.

**Table 3.2 Self-describing features**

Other features
Votes count by readers (likes – dislikes) (high, low, none)
Author profile URL
Author previous work URL
Article topic/title
Date of publishing
Reshare count on other social media (high, low, none)

### 3.1.3 Small Annotation

After we collected a blog corpus and compiled our list of credibility features, we asked a group of annotators to annotate the blog set for credibility, and include a score for each of the features presented in Table 3.1. The values for the features in Table 3.2 were extracted by the researchers who built the corpus, and were presented to the annotators as extra information about the blogs since it is likely that those features have a role in the credibility score of the blog post. For example, a blog post with a high number of likes and social network re-shares seems more credible than one that has none. The same logic applies for the author profile which includes details about the author profession and expertise, and to others.

All annotators who participated in this experiment had are Native Arabic speakers, have a general background on the topics of the articles, are working in the NLP field, and

have been given clear instruction for the annotation process and background on the project goals.

In addition to evaluating the credibility features, annotators were given a free space to include their personal comments regarding the factors that contributed to their assessments, and to include their findings and comments on their annotation experiment experience.

Each annotator was provided with an average of 8 pseudo-randomly selected blog posts to annotate, and each of the 25 blog post was evaluated by at least 3 annotators.

#### 3.1.4 Annotation Evaluation

The annotations were then collected from the participants, and summarized based on majority vote to end up with a credibility annotated corpus of 25 blogs, 15 from which are credible (60%), 8 fairly-credible (32%) and 2 non-credible (8%). After that, the annotation agreement measure between the annotators was calculated to validate that blog credibility can be actually evaluated. If the annotators had a satisfactory agreement score on credibility scores, it can be hypothesized that there is a certain consensus on credibility between people in general and therefore can be building an automated system for credibility prediction would be adequate. Fleiss' kappa was used as an agreement measure since it works for a fixed number of annotators (greater or equal to 2) when evaluating items based on categorical values [36]. When calculated, the kappa score turned out to be about 0.583 with 75% pairwise agreement, indicating that the hypothesis is actually feasible [37].

Furthermore, the collected annotations were used to build a test model serving as a “proof of concept” for automatic blog credibility prediction. Several models were actually tested using WEKA classification platform [38]. As mentioned before, the annotated corpus is imbalanced on the classes, which might bias the classification models into classifying most articles to the most frequent class (credible); therefore, a *Synthetic Minority Oversampling TEchnique (SMOTE)* was applied on the corpus to overcome this problem, to give birth of a new balanced corpus from the original one which was used on the classification models for training and testing. We summarize in Table 3.3 the results of the classification when trained and tested based on the 10 fold cross validation method, on both the original and data set generated by SMTOE filter. The best results were achieved with Bayes Net and Decision tree classification models, with accuracy and f-measure close to 90%, supporting the feasibility of a system that decided blog post credibility automatically.

**Table 3.3 Classification results of the original and (SMOTE-ed) balanced datasets using different classification models**

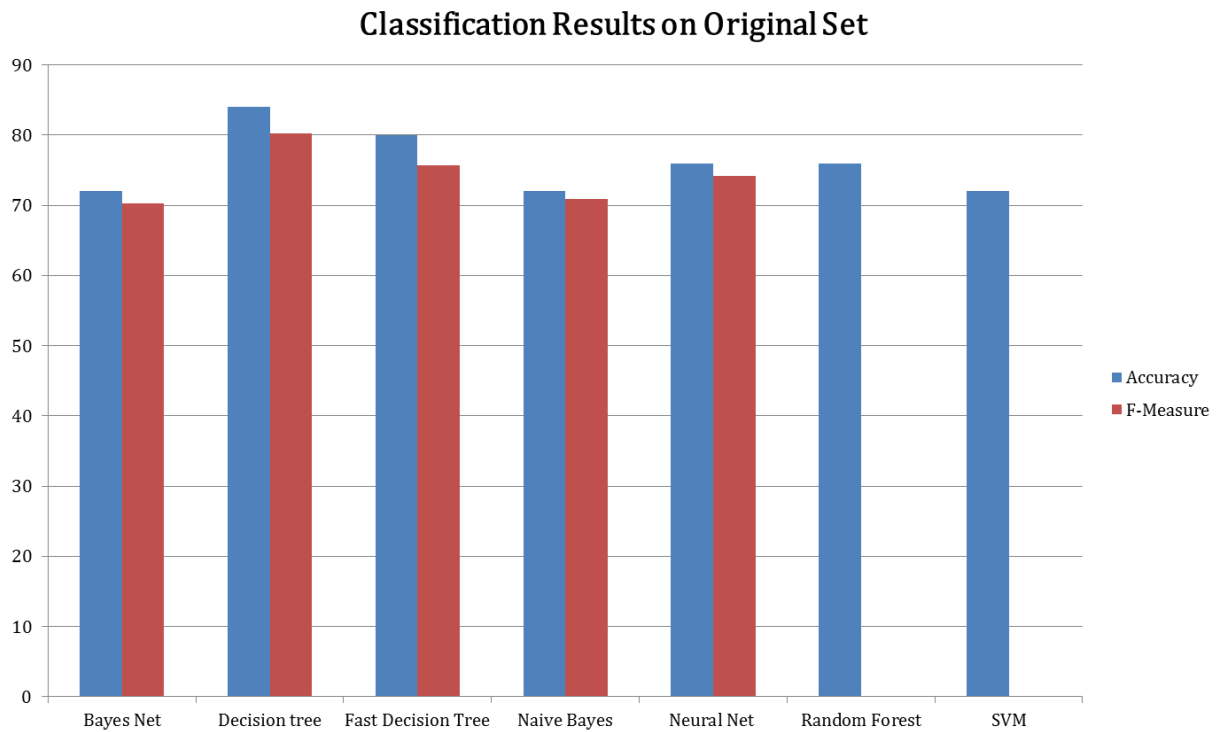
Classifier	Original Data Set	SMOTE-ed Data Set
------------	-------------------	-------------------

	Accuracy	F-Measure	Accuracy	F-Measure
Bayes Net	72	70. 9	91 .3043	91. 2
C4.5 decision tree (j48Grafted)	84	80. 2	91 .3043	90. 8
Naive Bayes	72	70. 9	89 .1304	89. 1
Neural Net	76	74. 2	89 .1304	89. 1
Random Forest	76	-	89 .1304	-
SVM	72	-	89 .1304	-

It is also worth mentioning that the corpus collection and annotation process highlighted some challenges mainly related to the structure and layout of Arabic blogs. It was observed that most Arabic blogospheres rarely included a profile page for the blog authors with their credentials and expertise, compared to English blogospheres where it is more available. Additionally, social features like comments, votes, and re-shares were very poor for Arabic blogs, and sometimes found null. Therefore, it was important to intend on building a



credibility prediction system that can run with the absence of those features.



**Figure 3-2 Small Scale Classification Results**

### 3.1.5 Feature Selection

In addition to verifying that building an automated credibility prediction system for Arabic blogs, the aim of the small scale user study was also to identify the features contributing in Arabic blogs credibility assessment the most and that can be collected and automatically evaluated. Annotators mentioned in their comments *reasonability*, *author expertise*, *bias* and *perceived overall website credibility* helped them decide on the blog post credibility the most. We use our literature review, our annotation experiment results, and the users input to decide on our feature set by calculating the Information Gain for each feature on WEKA. Information Gain scores can help in identifying the most relevant

features to the credibility score decision. We show in Table 3.4 the most relevant features based on Information gain ranking. We also try the CfsSubsetEval method provided also by WEKA to find the subset of features most relevant with credibility as an affirmation for Information Gain, as observed in Table 3.5.

**Table 3.4 Information Gain scores of each feature**

Attribute	Average Rank	Average Merit
Reasonability	1 +- 0	1.166 +- 0.042
Objectivity	2.2 +- 0.4	0.788 +- 0.03
Bias	2.8 +- 0.4	0.751 +- 0.041
Sentiment	4 +- 0	.642 +- 0.051
Author Consistency	7 +- 0	0.358 +- 0.037
Comments Conformity	8.1 +- 0.3	0.129 +- 0.023
Citations	8.9 +- 0.3	0.102 +- 0.013

As Table 3.4 shows, *reasonability*, *objectivity*, *bias* and *sentiment* ranked the highest among other features, and with a large gap of 3 ranks between *sentiment* and *author consistency*, and also a large difference of 0.3 in average merit between the two features,

compared to that between the prior 3 features where the average merit difference was 0.1. This clearly suggests that the four aforementioned features had the highest impact on the credibility scores.

**Table 3.5 Best Subset of Credibility features using CfsSubsetEval and Information Gain**

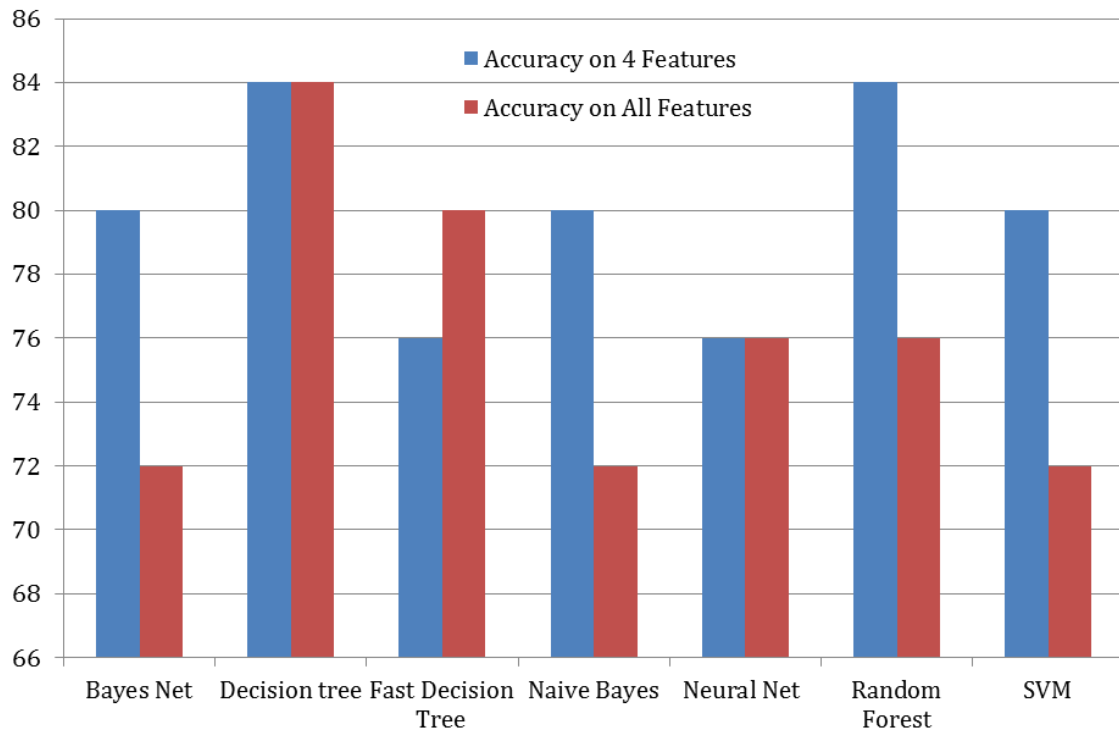
Data Set	CfsSubsetEval	Information Gain
Original	References To External Sources/Citations	Reasonability
	Biasness	Biasness
	Reasonability	Objectivity
		Document Sentiment
SMOTE-ed	Author Consistency	Reasonability
	References To External Sources/Citations	Biasness
	Biasness	Objectivity
	Reasonability	Author Expertise

Table 3.5 also suggests that those four features are the most relevant to credibility, but also shows *references to external source/citations*, *author consistency* and *author expertise* features among the best subsets. Although we find this strange given that in most blogs those features weren't available; however, we speculate that this is because annotators directly marked a high credibility score whenever they found the author to be an expert and consistent or used a lot of citations, and therefore making those features

prominent features for credibility when available. Yet, we didn't include those 3 features (*expertise, consistency and citations*) in our final credibility features set since our user study and collected corpus strongly shows that those features are very poorly or even not available in most Arabic blogs; and when available, it is hard to automatically extract those features due to the inconsistent design between Arabic blogs. Additionally, the other 4 features (*reasonability, objectivity, bias, and sentiment*) were found among the most relevant features in almost all experiments, which encouraged us to use those features instead.

#### 3.1.6 Findings

The main finding from our small scale user study was the feature set we finally arrived to which includes: *reasonability, objectivity, bias, and sentiment*. Those 4 features will be extracted from Arabic blogs and used as input for Arabic Credibility Models to output a predicted score for credibility.



**Figure 3-3 Modals Accuracies when the 4 features were only used**

### **3.2 Large Scale Annotation**

The positive findings from the small scale user study served as an indication for the plausibility of a larger full automated system for credibility prediction. Therefore, a larger corpus had to be collected and annotated for our credibility features, and then used as a train and test set for building an automated module for each of the 4 discovered credibility features. In other words, we aim to automate the evaluation of the 4 credibility features to be able to use them in a fully automated system for credibility prediction, and having a large annotated corpus was needed for testing such a system.

Next, we discuss the large scale annotation process that was launched by exploring its different steps including corpus collection, the required features for annotation, building the annotation interface, and finally the annotation experiment results and analysis.

### 3.2.1 Large Corpus Collection

Similar to the work done in the small corpus collection, we compiled a set of queries about topics hot at the time of the data collection, and used web search engines to manually collect a set of Arabic blog posts related to the suggested topics. In total, we collected 268 in two stages, mainly covering the following topics:

- المحكمة الدولية الخاصة بلبنان
- حكومة المصلحة الوطنية
- إنتخابات رئيس الجمهورية في لبنان
- كأس العالم
- الأزمة السورية
- الحرب في غزة
- الجيش اللبناني
- الدولة الإسلامية في العراق والشام
- أزمة السيسي

### 3.2.2 Features

As mentioned in section 3.1.6, our findings show that the four features: *reasonability*, *bias*, *sentiment*, and *objectivity*, had the highest correlation with credibility.

Therefore, in this annotation experiment, we asked the annotators to evaluate all those features along with the credibility of the Arabic blogs.

We also adapted the Merriam Webster [40] definition for each feature, and provided a scale as follows:

1) **Credibility:**

- Definition: the quality of being **believed or accepted** as true, real or honest
- Scale:
  - **Credible:** if you think the text is believable or accepted to be true .
  - **Not-Credible:** if you think the text is not believable and accepted to be true.
  - **Fairly Credible:** if you think the text is not fully believable, yet not completely not-credible.
  - Can't Decide, otherwise

2) **Reasonability:**

- Definition: Being in accordance with reason.
- Scale:
  - **Highly-Reasonable:** when the text is rich with reasonable arguments and judgments.
  - **Reasonable:** when the text contains reasonable arguments and judgments
  - **Acceptable:** when the text is mixed with reasonable arguments that are exaggerated and not that logical.

- **Exaggerated:** when the text only contains exaggerated arguments and lacks proper flow.
- Can't Decide, otherwise.
- Example: أساساً، ينطلق الخلاف الأميركي – الروسي / يبدو من الصعب تصديقه، لأسباب... جوهرية عدّة، أهمها: أولاً... ثانياً

### 3) **Bias:**

- Definition: A tendency to believe that some people, ideas, etc., are better than others that usually results in treating some people unfairly
- Scale:
  - **Biased:** if the text obviously over favors a certain group.
  - **Not-Biased:** if the text shows fairness between the different groups, without favoring of one group over the other.
  - **Fairly-Biased:** if the text shows some level of bias, yet not completely biased.
  - Can't Decide, otherwise.
- Example: وعلى وقع الموت و الخسائر .. حزب الشيطان اللبناني المجوسي الارهابي يناشد مواليه

### 4) **Objectivity:**

- Definition: Expressing or dealing with facts or conditions as perceived without distortion by personal feelings, prejudices, or interpretations. In our context, objectivity would reflect not using first-person words in the whole article, and referring to facts, links, laws, etc...
- Scale:



- **Highly-Objective:** if the text is rich with facts without any distortion by personal feelings and interpretations.
- **Objective:** if the text contains expressions revealing personal feelings, and contains facts.
- **Fairly-Objective:** if the text is poor facts without any distortion by personal feelings and interpretations.
- **Not-Objective:** if the text is loaded with personal feelings and interpretation.
- Can't Decide, otherwise

○ Examples:

- الم يظهر زيف تلك المحكمة المزعومة بعد ؟؟ انا اراها قد دفنت نفسها بنفسه  
(Not Objective)
- من أربع سنين، كتبت مقالاً بعنوان فخامة الفشل، وقتها زعلت وطلبتني مخابرات المنطقة وقتها واسألهم بنهاية ولايتك: عالتحقيق، اليوم عبالى ارجع شوف الشباب الي حققوا معي فشل ده ولا مش فشل يا مخابراتين يا بتوسع المخافر؟  
(Not Objective)
- في الواقع ما يحدث في سوريا الآن ليس بمعزل عن ما حدث بالعديد من الدول العربية من ثورات مناهضة للأنظمة الحاكمة ورافضة لسياسات التسلط والاحتكار. فمثل هذه السياسات من شأنها أن تنتهك حقوق الإنسان وتتعدى على الحريات سواء على المستوى الفردي أو الجماعي  
(Objective)

5) Sentiment:

- Definition: the opinion polarity that is expressed in the document. In other words, Sentiment refers to how positive or negative the text reflects.
- Scale: Very negative, fairly negative, neutral, fairly positive, very positive

### 3.2.3 Annotation

We built a custom made annotation interface as a web service to make the annotation process easier for the annotators. Annotators would register with their names, education level, and other information, and then login to be presented with an article to annotate for the pre-mentioned features. All annotators were native Arabic speakers familiar with the topics of the blogs in the corpus. Before they started the official annotation, all annotators were given a tutorial session followed by a mock-up hands-on annotation session to get them familiar with the problem and the annotation process. A golden set of annotated blogs were available also for annotators, as a sample annotation for some blogs. Annotator could also change an article if they feel that they can't annotate it properly, for example if they weren't very familiar with the specific topic discussed, or so. Additionally, annotators were given compensation for their annotations, and were given 2 weeks to complete their annotations, without having to be present in our labs. Finally, after the annotation were collected, all annotation which took unreasonably short time to annotate were omitted, as this means that the annotator didn't put enough effort in the annotation and most probably provided random scores for the features.

The web service created was made online, and can be accessed through this link [annotate.me-applications.com](http://annotate.me-applications.com).

Below are two snapshots from the annotation interface.

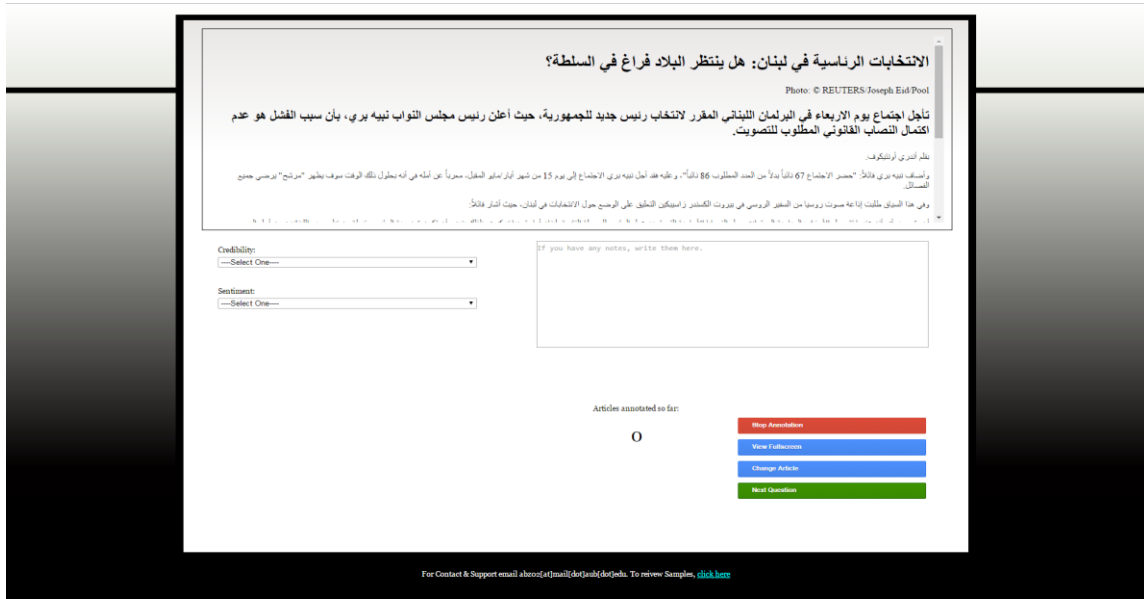


Figure 3-4 – Credibility and sentiment annotation.

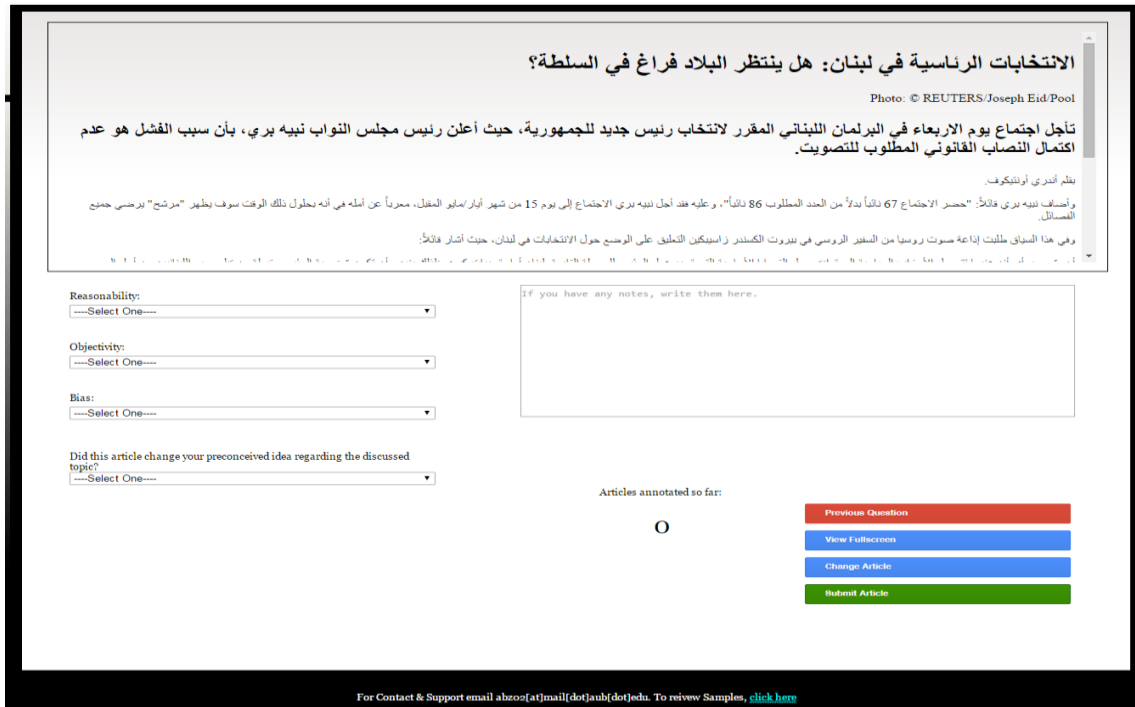


Figure 3-5 – The credibility features annotation

### 3.2.4 Annotation Results and Analysis

Each blog was annotated by 4 annotators, and a fifth annotation was provided by me, the thesis writer. The score for each blog was calculated based on majority vote on each feature, and a subset of the annotated corpus was extracted with full agreement.

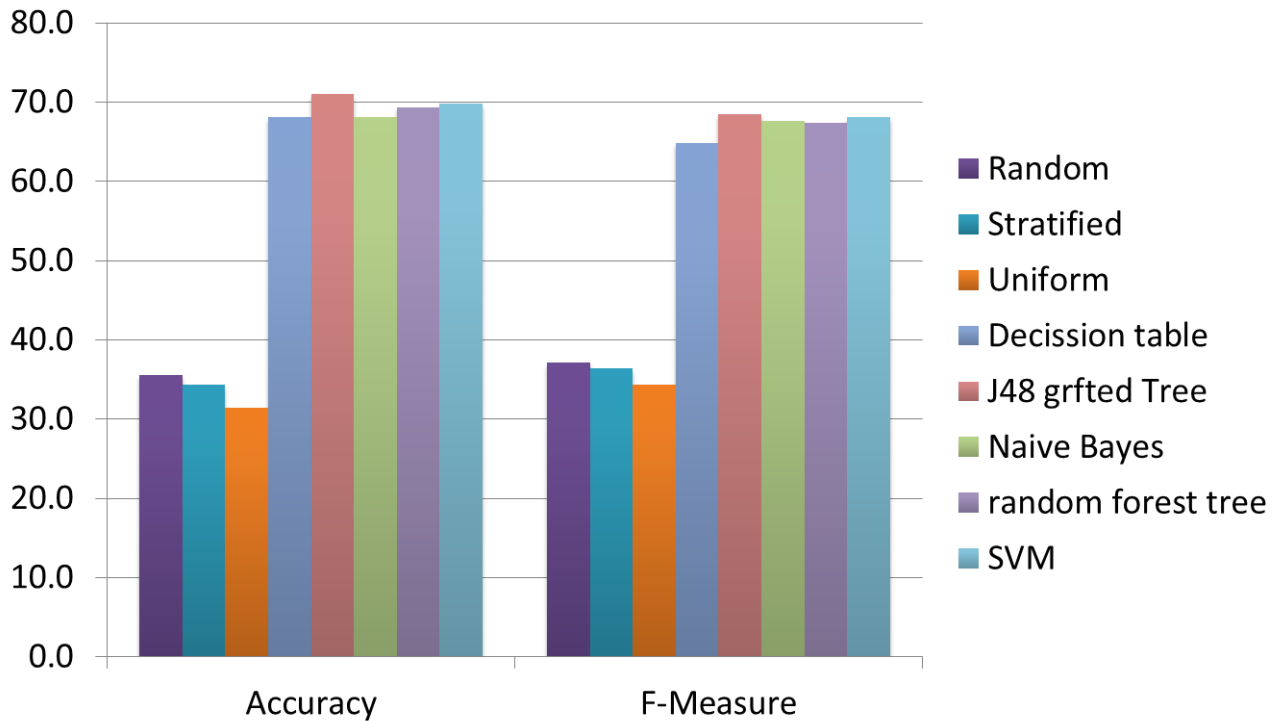
An inter-annotator score was also calculated using the Krippendorff's alpha measure. This measure was chosen as it is the most suitable for the setting of our annotation, given more than two annotators and features with nominal scores. The k-alpha score turned out to be 0.3, which is acceptable [41, 42], given the natural hardness of the task, and the tree-level nominal scale which drastically reduces the k-alpha value, as opposed of being numerical.

The corpus contained 125 credible articles, 79 fairly credible and 64 non-credible articles. The golden subset with full agreement contained 135 blogs.

A subset of this corpus and its annotations was published online, and details can be revised in [39].

As a proof of concept for the plausibility of building a fully automated system for credulity assessment of Arabic blogs, several credibility models were built on the data set generated with the given features.

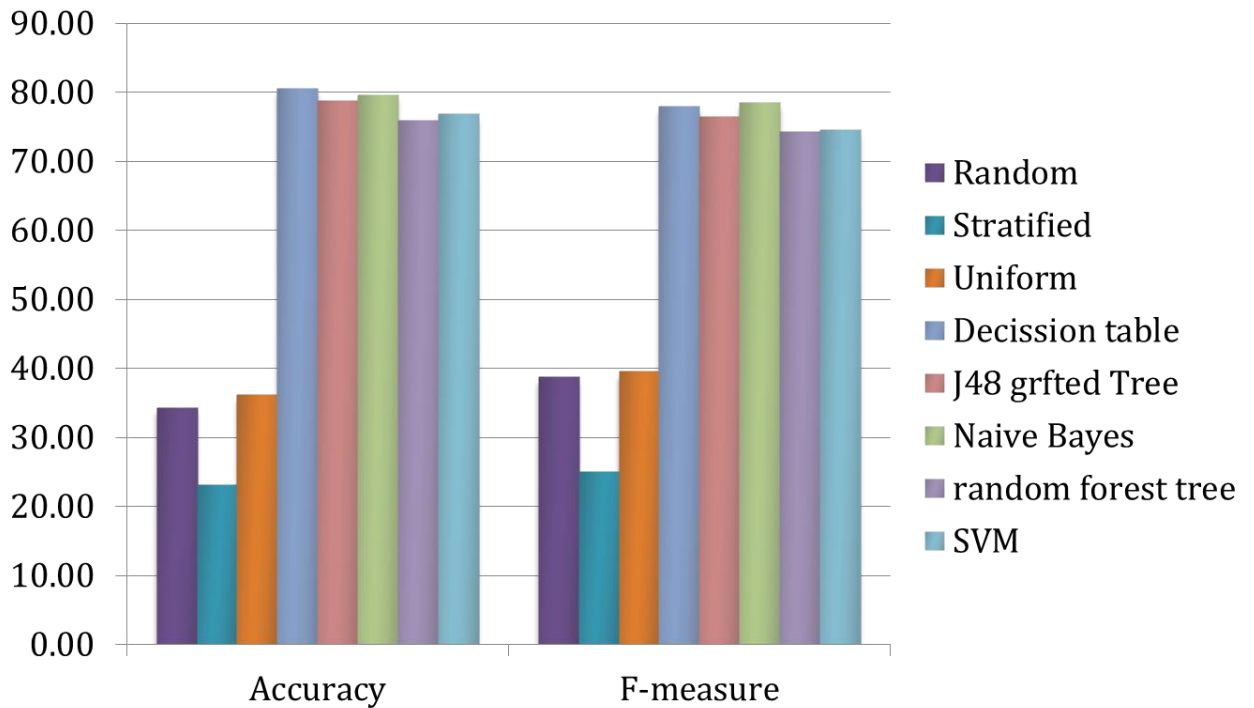
First, we built classification models using the full data set. The results were as follows:



**Figure 3-6 Classification results on full data set**

As we can see, the accuracy and f-measure of most classification algorithms is almost 70%, all higher than the baselines (Random, Stratified, and Uniform). Although we see a 10% drop from the small scale set, but this was normal, since in the previous experiment the set was small and therefore the values might not be as representative as the ones here. Still, 70% is considered acceptable and motivating for building a fully automated system on this data set. We also argue that if the data set was better balanced between the classes, we would get a better score. This can be achieved by collecting a yet larger set.

Next, we ran the same experiments using the golden subset, and got the results shown below.



**Figure 3-7 Classification results on golden data set**

As we can see in the above results, a 10 percent increase was almost achieved by all classification algorithms (except the baselines), which gives us a clear indication that credibility can be automatically assessed with high accuracy for the articles which humans have a common consensus on its credibility.

### 3.3 Feature Extraction

To this point, we don't have a fully automated system. Our current system still needs to be fed the feature scores in order to produce a credibility score.

To fully automate our system, we need to automatically extract the feature scores, and then feed them to the classification model and get a credibility score as a result.

To that end, we tried to explore each feature on its own, and the specific linguistic features related to it.

In the following subsections, we present the features explored for each feature.

### 3.3.1 Reasonability Exploration

It was noticed that articles with high reasonability score are rich with connectives like (بل)ولكن اتم وبعدما او بما أن), phrases indicating making conclusions ( وبالتالي وبالنتيجة او على ), numbering phrases, and some others. We also noticed a similar hypothesis in a credibility classifier built for English reviews [43], where they showed that credible reviews were rich with weak modals, and phrases similar to the ones we presented above, of course in English language.

### 3.3.2 Objectivity Exploration

For objectivity, it was noticed that the presence of past tense verbs, and 3<sup>rd</sup> person terms correlated with high objectivity. This makes sense since past tense usually represents a phrase stating facts (incidents already happened), and the use of 3<sup>rd</sup> person terms reflects non-subjective opinions and phrases. Additionally, objective articles were rich with quoted text.

### 3.3.3 Bias Exploration

Biased articles were rich with hate speech and negative description associated with named entities.

#### 3.3.4 Sentiment Exploration

Sentiment wasn't explored since its evaluation can be done by a heuristic on the total positivity and negativity score on each term in the document.

#### 3.3.5 Extraction Tools

To extract those features, we used several tools including Madamira [44], Morphological Analysis and Disambiguation tool for Arabic text, which helped us extract statistics on verb tenses, person tense, aspect, interrogatives, part-of-speech, and many other linguistic features.

Additionally, we used LIWC set (Linguistic Inquiry and Word Count) [45] to collect strong and weak modal terms, inference terms, connectives, necessity, numbering, and many others.

Finally, we used ArSenl [46] to evaluate the sentiment of the document, by applying a function over the positive/negative scores for each word.

We note that all the collected scores were normalized to properly handle different sized articles. Our feature space included 47 numeric features.



## CHAPTER 4

### PERFORMANCE EVALUATION

In this section, we present the performance of the fully automated system, when tested on different settings.

#### 4.1 Machine Learning Algorithms vs Baselines

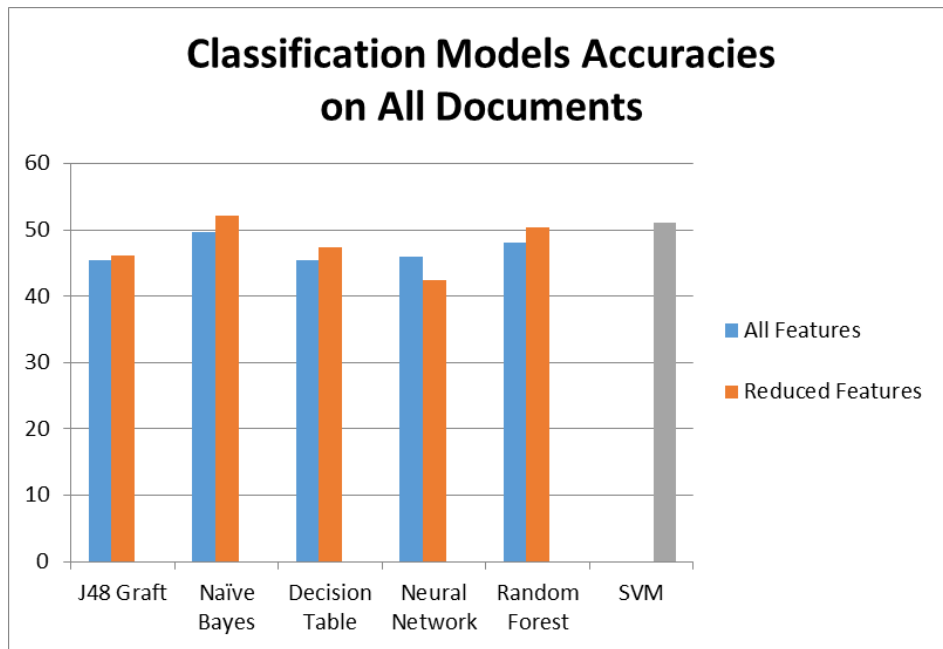
Our first approach was to feed the set of collected features above into classification algorithms, and explore the different results on different settings. We tried several settings including:

- Full data set vs Gold data set
- All features vs best subset of features

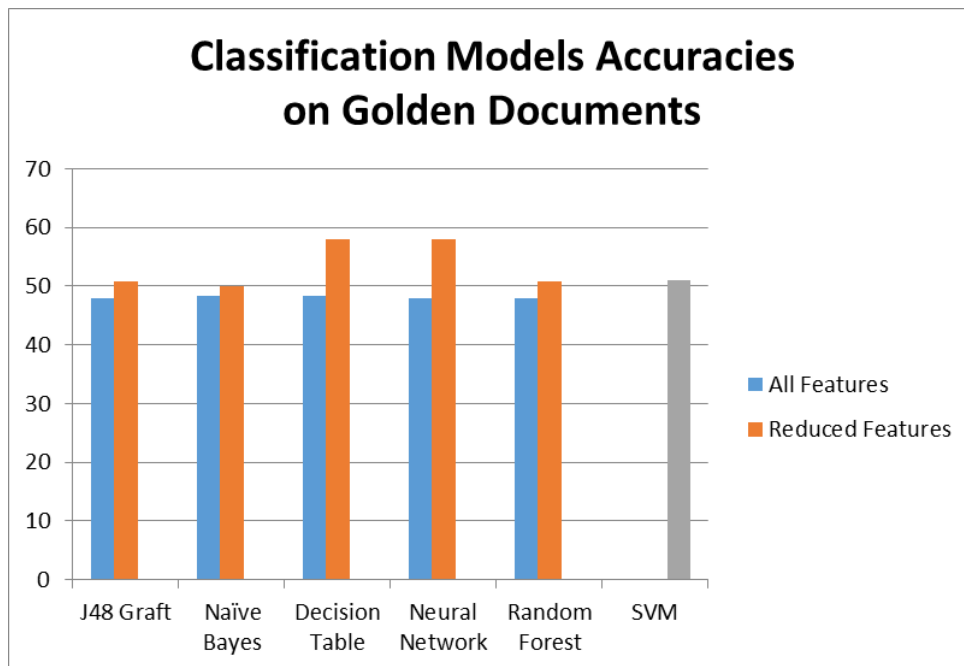
For the reduced features, we used feature reduction methods provided by WEKA to minimize our feature space from 47 features to the most significant 15 features.

We used WEKA classification platform to classify our documents to credible, fairly credible, or non-credible scores, using the feature vectors as inputs. We explored several classification models on the original and golden subsets, using the full feature space and the reduced one.

As a baseline, we used an SVM with the tf-idf scores as input.

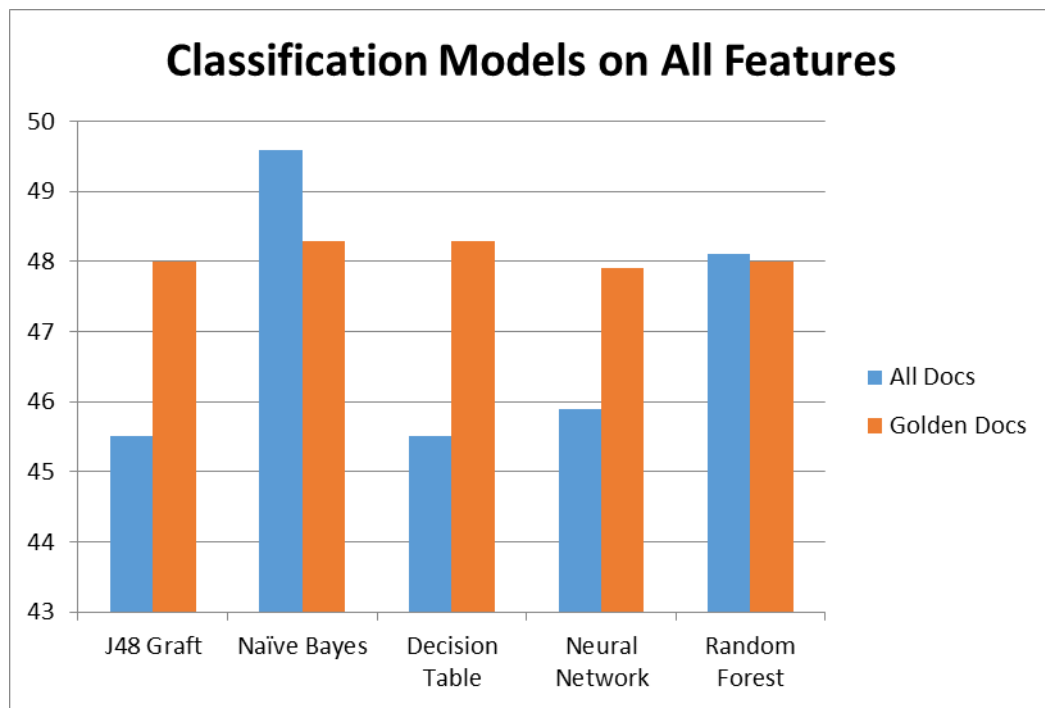


**Figure 4-1 Classification Models accuracies on All documents as input, comparing all features and reduced features inputs**

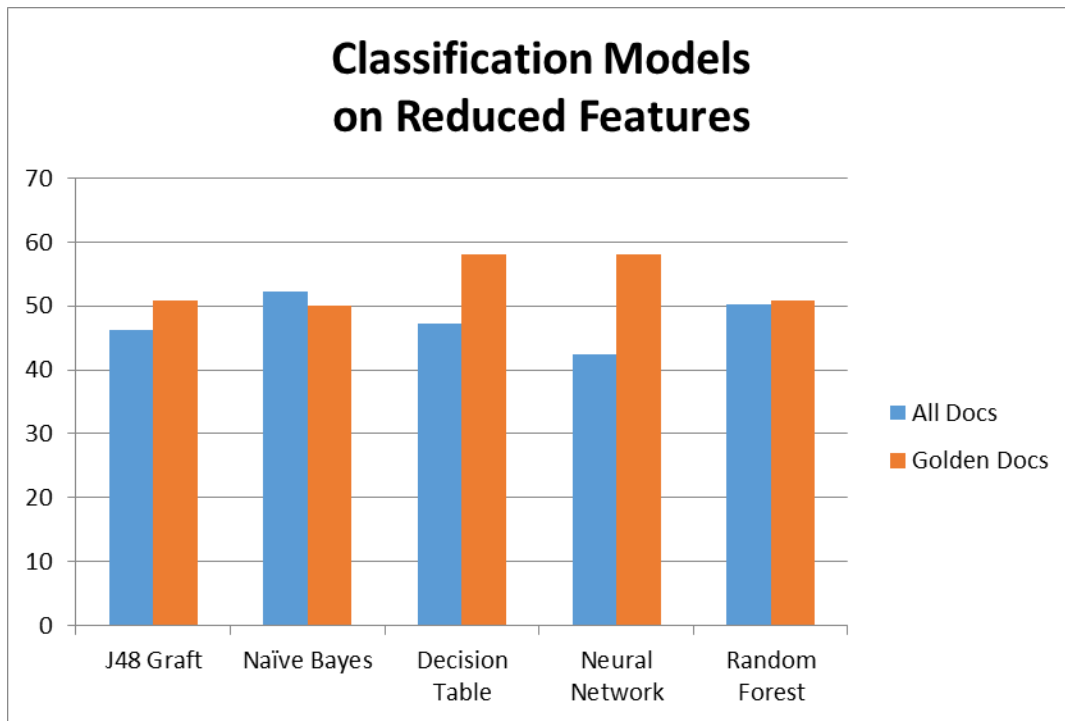


**Figure 4-2 Classification Models accuracies on Golden documents as input, comparing all features and reduced features inputs**

As we can see, in both settings, when we used the reduced features space, we got higher values in almost all the classifiers. However, the results were close to the SVM baseline except for Neural networks and Decision table on the Golden subset, using the reduced features space.



**Figure 4-3 Classification Models Accuracies on full feature space**



**Figure 4-4 Classification Models Accuracies on reduced feature space**

Now, we compare the results between all documents as input, and the golden subset. We can observe that almost always the golden subset is giving us better results.

## 4.2 Deep Learning Models

As a different approach, we try to explore deep learning models to assess the credibility of the Arabic blogs, and the 4 related features. The use of deep learning models can be helpful since deep learning models “use dense hidden layers for automatic feature combinations, which can capture complex global semantic information that is difficult to express using traditional discrete manual features” [47]. We can also use deep learning models to represent sentences and documents with dense vectors, helping in evaluating the credibility of the blog post.

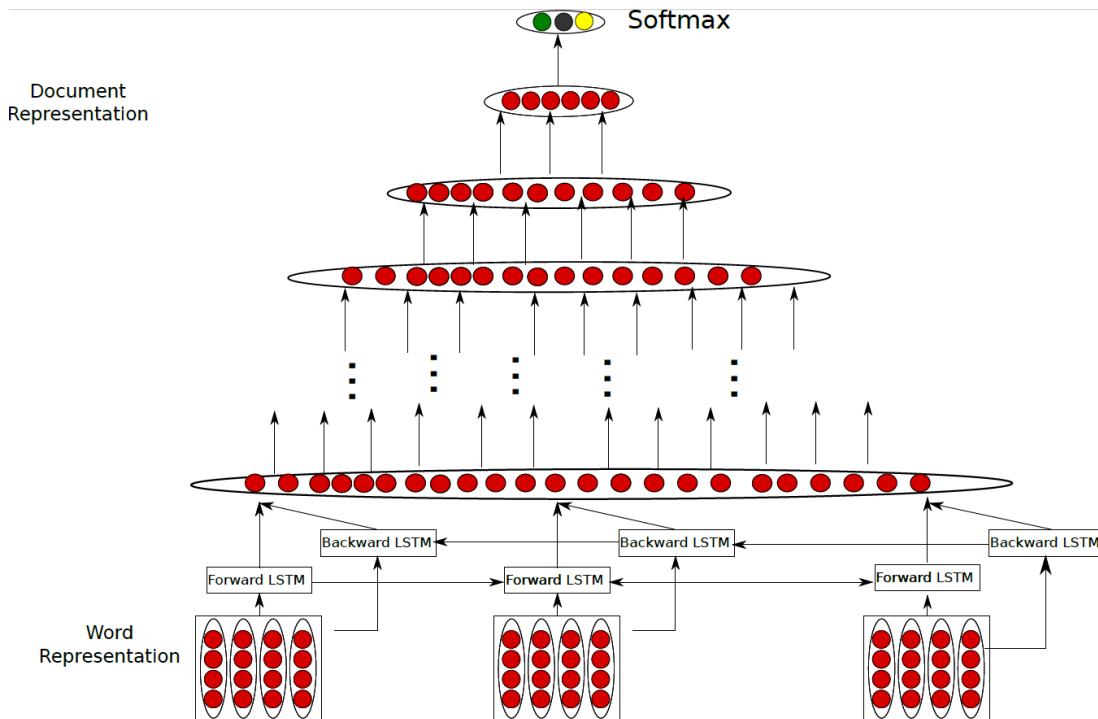
We note that deep learning models are very successful on image recognition tasks, but are still limited in performance on NLP related tasks, especially when the task is hard for a human to complete like credibility assessment, rather than for a machine.

As we did with the classification models in section 4.1, we explore different deep learning models on different settings. Specifically, we try the LSTM (Long Short Term Memory) model, which will help us link previous information in text with later information. We also incorporated a CNN (Convolutional Neural Network) to help us build sentence representations before building document representations.

Those two settings can take raw text blogs posts as input, and build document representations to finally produce credibility scores, with the difference that one of them produces and intermediate sentence representations on the way. We adapt this strategy from [47].

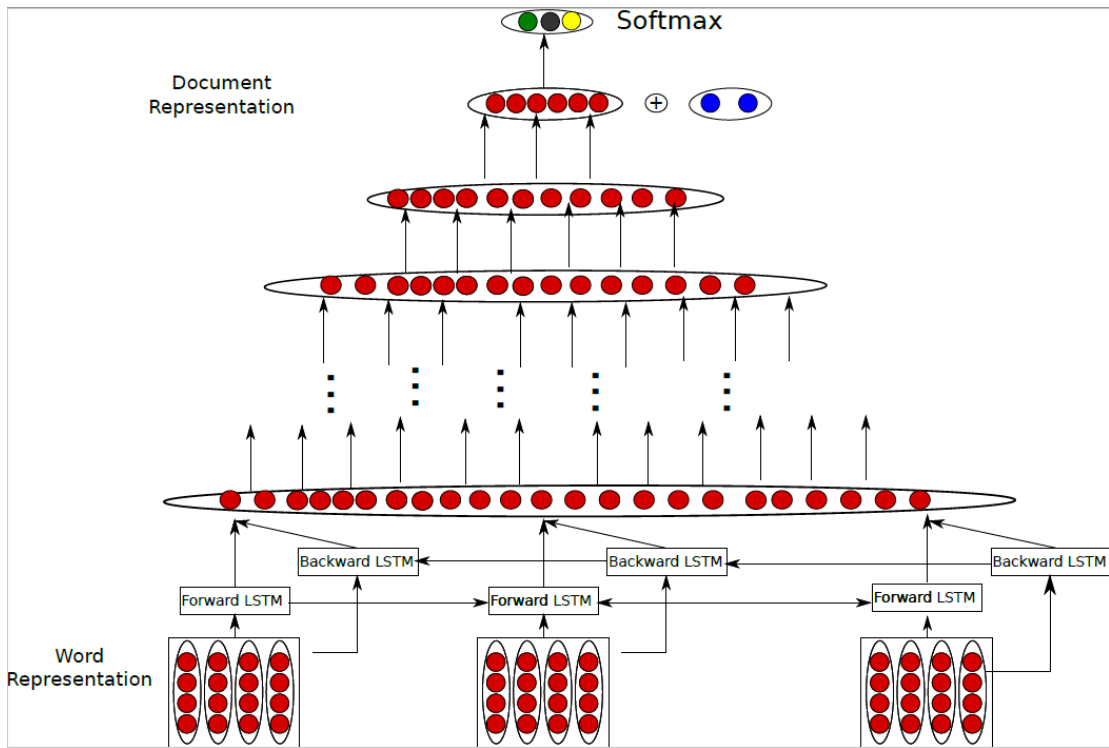
Additionally, we merged the feature vectors we produced in section 3.3.5 to the document vectors before predicating credibility, and compared the results with previous settings.

Finally, we tried to predict the 4 credibility features first, and then use them to predict credibility, as a final setting, all having the training set split 80/10/10 for training/tuning/testing.



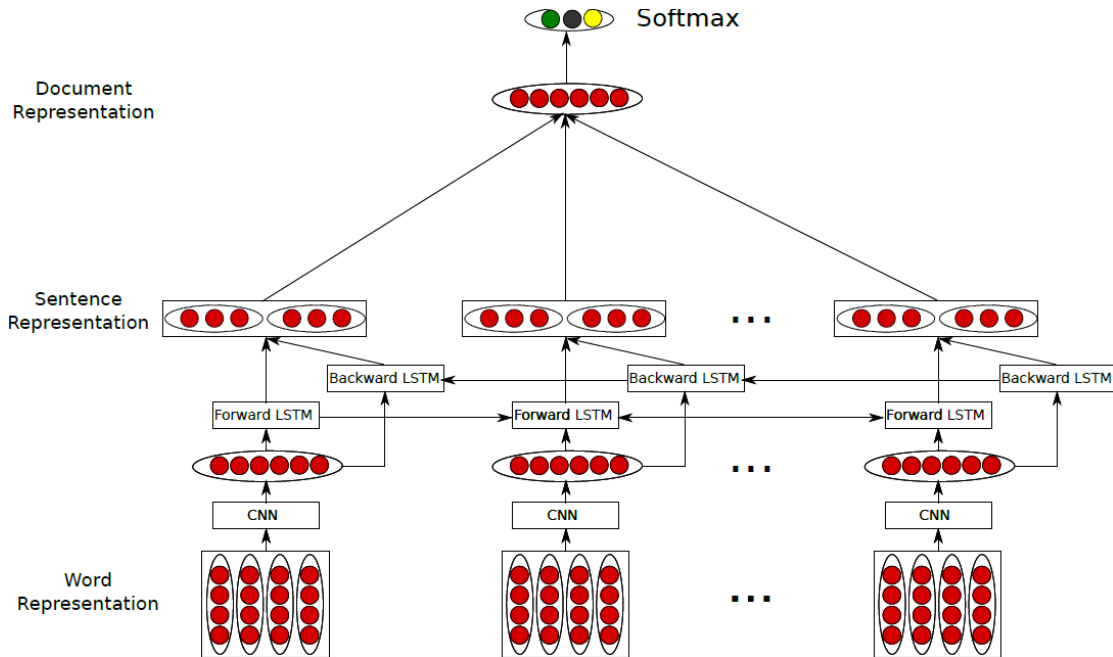
**Figure 4-5 LSTM Model**

This model directly produces document vectors from word embeddings, without making use of the feature vector space or sentence representations.



**Figure 4-6 LSTM + Discrete Features Model**

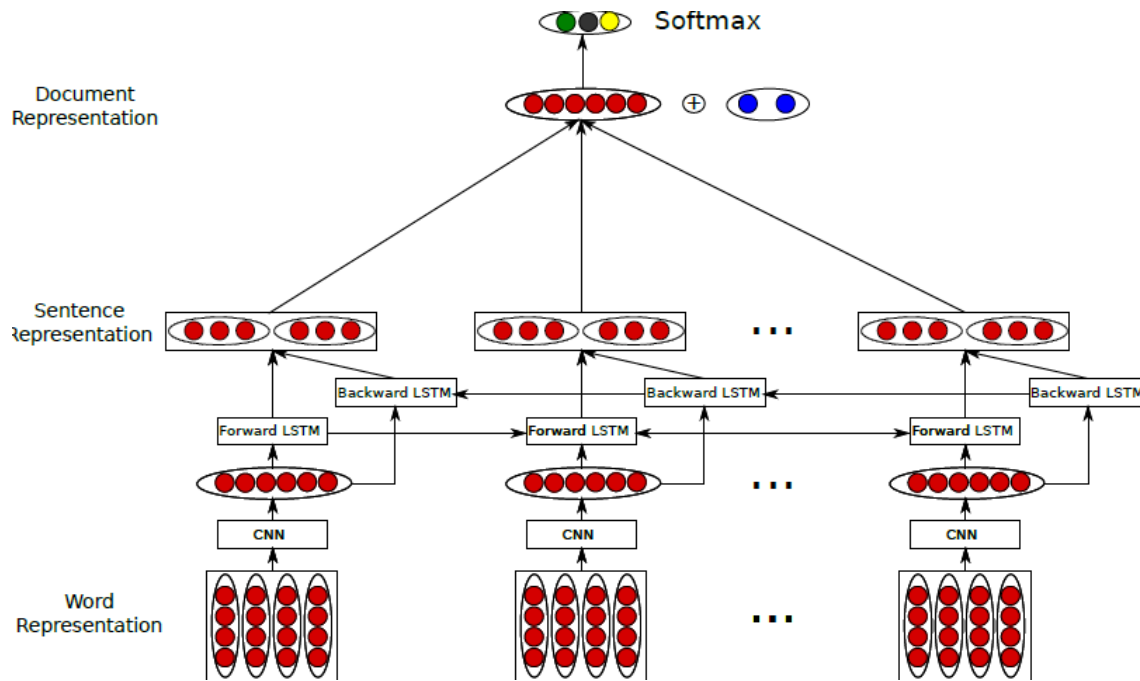
This model is similar to the previous one, but incorporates the discrete features vector on the document level, before predicting credibility.



**Figure 4-7 CNN + LSTM Model**

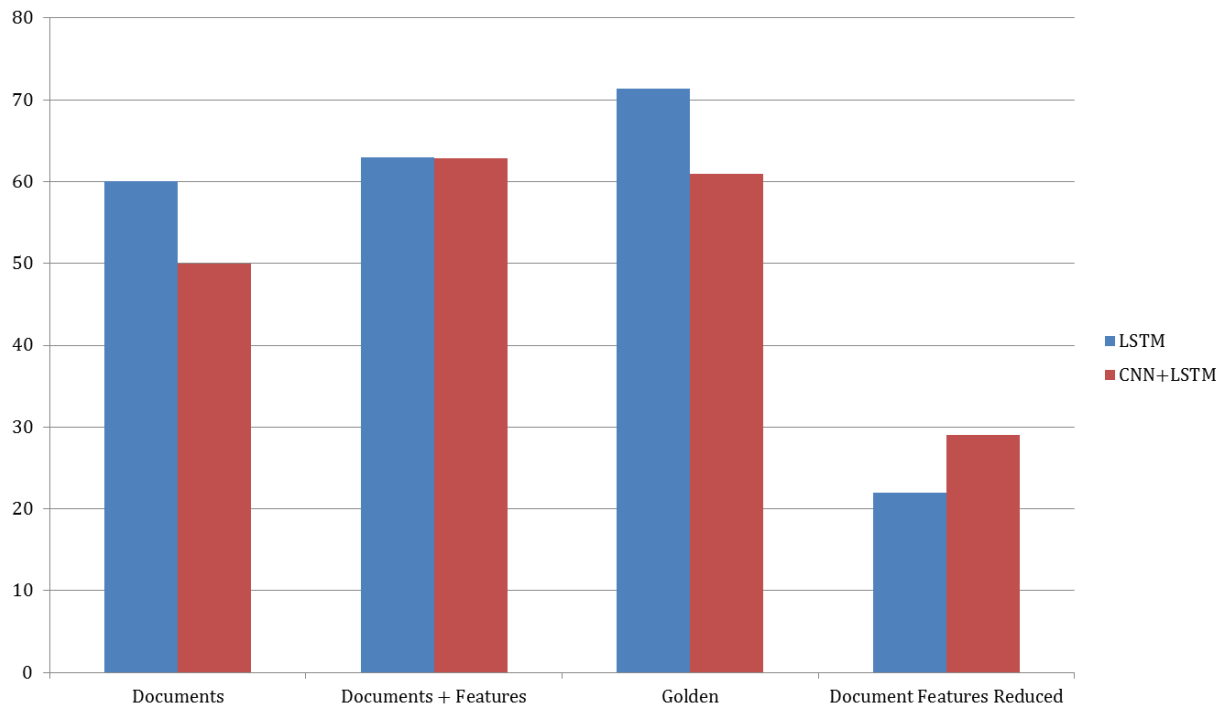
This model produces sentence level vectors using CNN networks with filters 1, 2, and 3, producing uni-grams, bi-grams and tri-grams. After that, it builds up the sentence representations vectors with dense layers to produce document representations and finally predict credibility.





**Figure 4-8 CNN + LSTM + Discrete Features Model**

This model is similar to the previous one, but incorporates the discrete features vectors with the document vectors before predicting credibility.



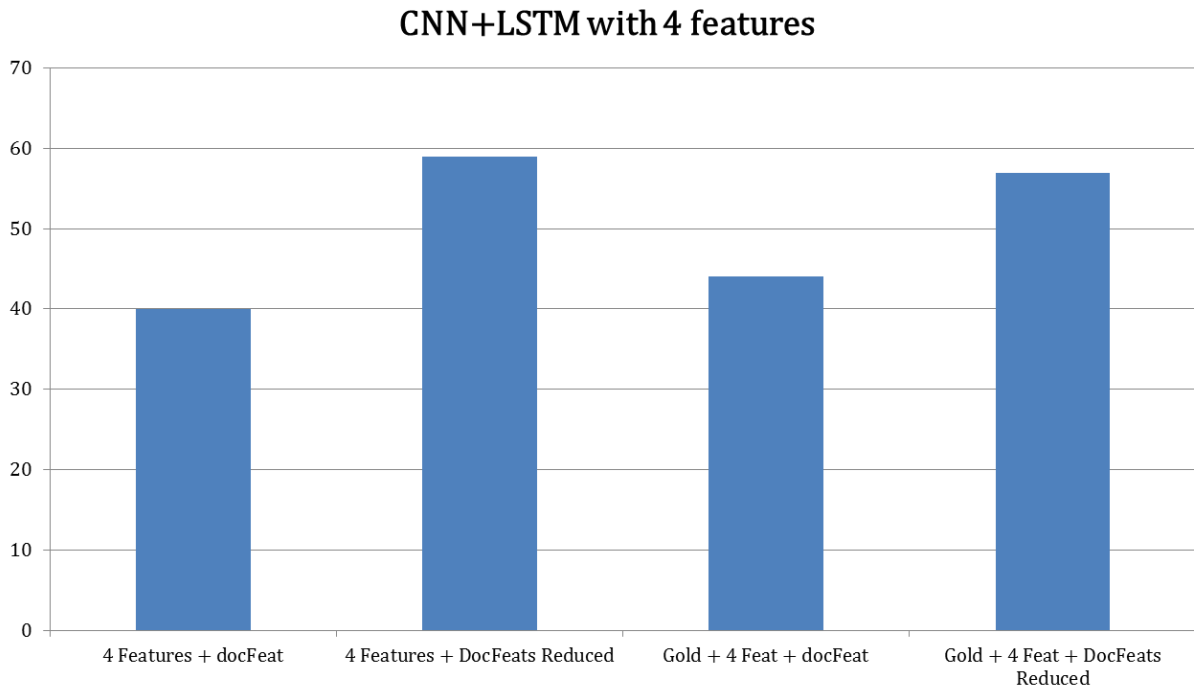
**Figure 4-9 LSTM vs CNN + LSTM Accuracies**

In this figure, we compare the results when having the documents only as input, the documents incorporated with the features, the golden subset, and the documents with the features reduced as before.

It is obvious that the CNN + LSTM model is not giving better results in most cases. We also notice that incorporating document features is pushing up the accuracies by 3%, from 60% to 63%. However this results is not yet satisfying.

We can observe a satisfying result when we use the Golden subset, getting a 11% jump to 71.4%.

We next try to predict the 4 feature score on an intermediate level before predicting the credibility of the document.

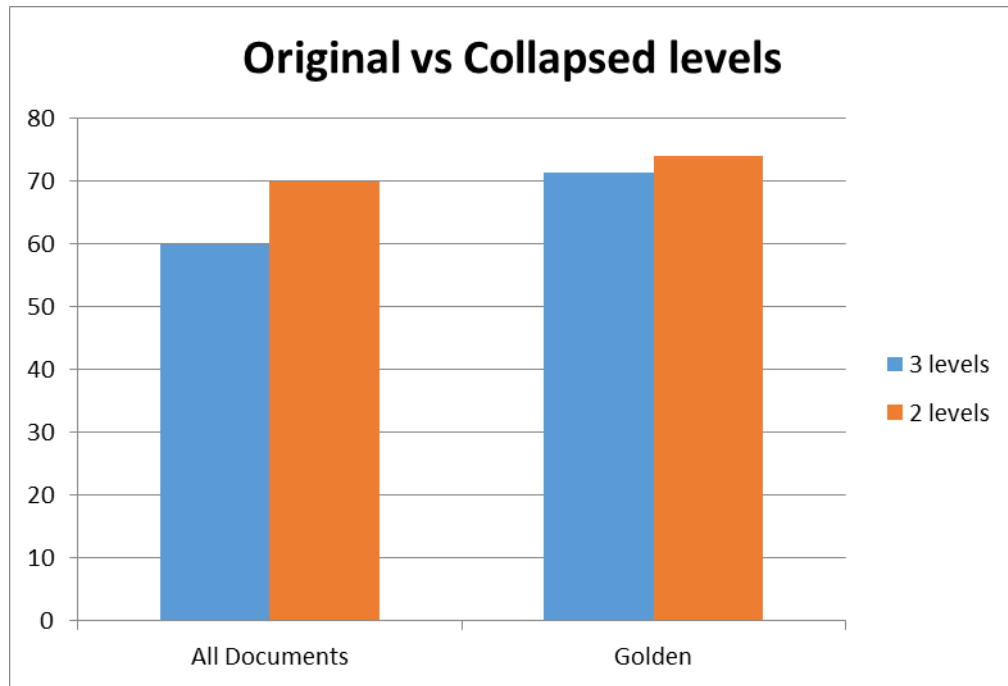


**Figure 4-10 The 4 credibility features prediction in intermediate level**

In the above diagram, we present the score of different model that predict the 4 features in the intermediate level before predicting the credibility. We can notice that in all settings, the prediction of the 4 features led to a decrease in accuracy, except in the second one where we observed a 30% rise to 59%. However, this score is still lower than what we achieved without predicting the 4 features above, which was 71.4 % accuracy. Therefore, it seems that predicting the 4 features isn't a productive step towards credibility assessment, and predicting credibility directly is better.

In literature, we observed that they [47] had a two level classification, instead of 3. Additionally, they experimented on separate topics and on all topics combined. In our next

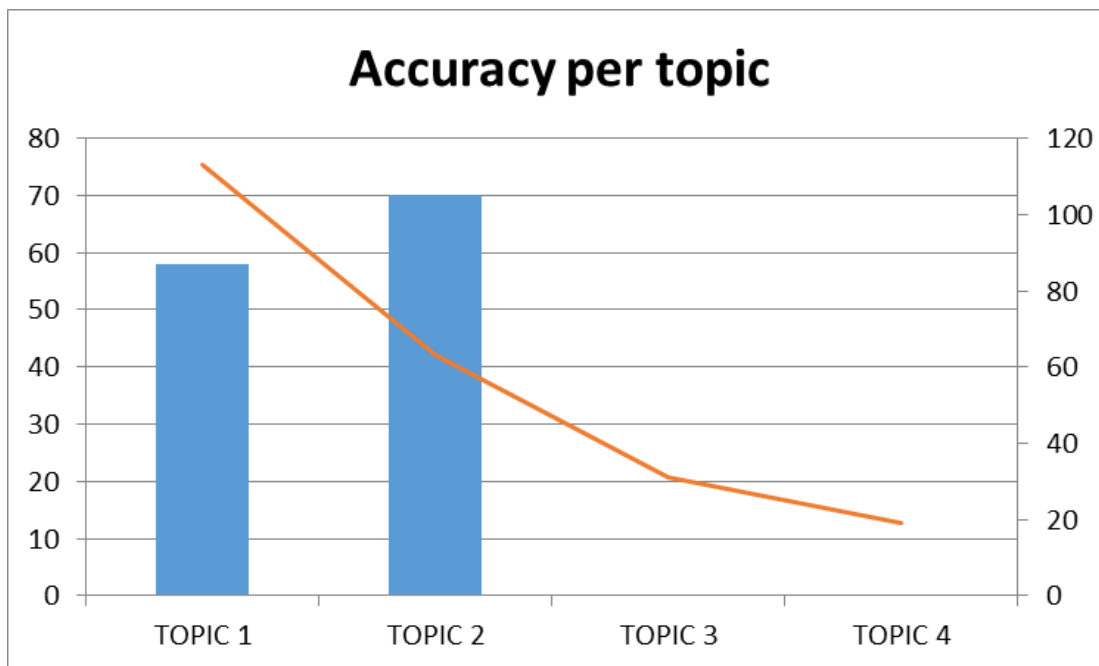
step, we collapse our credibility scores in one setting into two levels: Credible and non-credible; and in another setting separate the blogs based on topic.



**Figure 4-11 Accuracies on original credibility levels vs on collapsed levels**

The figure above shows how the accuracies increase when the credibility levels are collapsed into 2 levels instead of 3 for both full data set input, and golden subset inputs. We also notice a very satisfactory result of 74% accuracy in the golden subset.

Finally, we show the results when subsets were created based on topic, rather than joining all topics together in one data set.



**Figure 4-12 Accuracy measure per topic, with corpus size differences**

The above graph shows how different topics can produce different results. We can also notice that topic 2 alone gave a higher results by 10% than one all topics were combined together, indicating that the low score of 60% may be due to the heterogeneous training set. We also note that topics 3 and 4 gave no results since their corpus size was too small for the model to be trained on.

### **4.3 Discussion**

We first notice that the use of deep learning models produce better results in general than the classification algorithms. This is definitely makes sense and is expected since the deep learning models extracts complex linguistic features that the classification algorithms can't extract, in addition to the feature vectors that can be incorporated with the model.

However, the slight increase of 2% between the deep learning models with the feature vectors incorporated compared to the models relying solely on the raw document text wasn't expected. One would expect a much higher increase since the model is being fed fundamental data with high correlation with credibility. We argue that this might be due to the inaccuracy of the tools used (Madamira, LIWC, and ArSnel) as some obvious mistakes were found while testing. It might also be the case that we missed some key features and didn't include them in the feature vector space, causing the classification to be less accurate.

Additionally, Deep learning models are known to be hungry for large data sets including an order of thousands of blogs to give good results on text analysis, and in our case, we only had 268 blogs which is too little for a blog to be saturated on training.

Moreover, our data set is composed of several topics, which makes the content vary a lot and therefore making it even harder for the deep learning model to give good results.

Finally, the nature of credibility assessment is hard and has ambiguities if done by humans, since credibility has many dimensions and human subjectivity definitely is a major factor in credibility assessment; therefore, the corpus annotation might be a bit inaccurate, causing the deep learning model to be chancy in its decisions and assessments.

Yet, with all the mentioned limitations, our deep learning model achieved a very satisfying accuracy of 74%. In a similar work [47] they had a data set of 2600 reviews and

got 91% accuracy when test on a single topic, however, when they tested there method on a cross-topic setting, they achieved 57.3% accuracy, making our work better.

## CHAPTER 5

### FUTURE WORK

As for the future work, our aim is to overcome the major limitation which is our relatively small data set for deep learning. So our major effort should be spent on expanding our data set in an unsupervised way. We strongly argue that expanding the data set will push up the accuracy score significantly.

Additionally having the system complete, we aim to publish it online and make it available for public use.

Finally, we aim to explore more model and features that might have been missed in this work.



## CHAPTER 6

### CONCLUSION

In this work, we presented the first fully automated credibility prediction system for Arabic blog posts. We also built the first credibility annotated Arabic blog corpus, and made it online for research use.

While building the system, we explored many classification models, a large set of features. Our classifier makes use of an underlying deep learning model that relies on document representation and a set of extracted features to predict credibility. Our classifier achieves 74% accuracy, and competes with state of the art model built for similar tasks.

In the end, credibility is a major feature in any document one reads, and therefore, it is very important to be very aware how it can be assessed.

## CHAPTER 7

### REFERENCES

- [1] B. A. Nardi, D. J. Schiano, M. Gumbrecht and L. Swartz, "**Why we blog**," *Commun ACM*, vol. 47, pp. 41-46, 2004.
- [2] P. Golding and P. Elliott, *Making the News*. Longman London, 1979.
- [3] (17/12/2007). *After 10 Years of Blogs, the Future's Brighter Than Ever*. Available: [http://archive.wired.com/entertainment/theweb/news/2007/12/blog\\_anniversary](http://archive.wired.com/entertainment/theweb/news/2007/12/blog_anniversary).
- [4] (1999). *Peter Merholz Webpage*. Available: <https://web.archive.org/web/19991013021124/http://peterme.com/index.html>.
- [5] (2000). *Weblogs: A History and Perspective*. Available: [http://www.rebeccablood.net/essays/weblog\\_history.html](http://www.rebeccablood.net/essays/weblog_history.html).
- [6] (). *Types of Blogs*. Available: <https://wordpress.com/types-of-blogs/>.
- [7] (2014). *Which type of blog gets the most Google searches?*. Available: <http://blogambitions.com/blogs-with-highest-searches/>.
- [8] (2015). *2 Million Blog Posts Are Written Every Day, Here's How You Can Stand Out*. Available: <http://www.marketingprofs.com/articles/2015/27698/2-million-blog-posts-are-written-every-day-heres-how-you-can-stand-out>.
- [9] (2012). *Infographic: 24 Hours on the Internet*. Available: <http://www.digitalbuzzblog.com/infographic-24-hours-on-the-internet/>.
- [10] (2015). *Cumulative total of Tumblr blogs 2011-2015*. Available: <http://www.statista.com/statistics/256235/total-cumulative-number-of-tumblr-blogs/>.
- [11] (2011). *Number of blogs worldwide from 2006 to 2011*. Available: <http://www.statista.com/statistics/278527/number-of-blogs-worldwide/>.
- [12] (2013). *How Many Blogs Are There?*. Available: <http://snitchim.com/how-many-blogs-are-there/>.
- [13] C. Parker and S. Pfeiffer, "**Video blogging: Content to the max**," *MultiMedia, IEEE*, vol. 12, pp. 4-8, 2005.

- [14] A. Java, X. Song, T. Finin and B. Tseng, "**Why we twitter: Understanding microblogging usage and communities**," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, 2007, pp. 56-65.
- [15] (). *Credibility Definition by Merriam Webster*. Available: <http://www.merriam-webster.com/dictionary/credibility>.
- [16] B. Fogg and H. Tseng, "The elements of computer credibility," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1999, pp. 80-87.
- [17] B. J. Fogg, "**Prominence-interpretation theory: Explaining how people assess credibility online**," in *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, 2003, pp. 722-723.
- [18] B. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford and E. R. Tauber, "**How do users evaluate the credibility of web sites?: A study with over 2,500 participants**," in *Proceedings of the 2003 Conference on Designing for User Experiences*, 2003, pp. 1-15.
- [19] C. N. Wathen and J. Burkell, "Believe it or not: Factors influencing credibility on the Web," *J. Am. Soc. Inf. Sci. Technol.*, vol. 53, pp. 134-144, 2002.
- [20] M. J. Metzger, "Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, pp. 2078-2091, 2007.
- [21] J. Schwarz and M. Morris, "Augmenting web pages and search results to support credibility assessment," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 1245-1254.
- [22] A. Olteanu, S. Peshterliev, X. Liu and K. Aberer, "Web credibility: Features exploration and credibility prediction," in *Advances in Information Retrieval* Anonymous Springer, 2013, pp. 557-568.
- [23] P. Anderson, "Consumer health web site evaluation checklist," *Ann Arbor, MI: University of Michigan*, 2001.
- [24] J. W. Fritch and R. L. Cromwell, "Evaluating Internet resources: Identity, affiliation, and cognitive authority in a networked world," *J. Am. Soc. Inf. Sci. Technol.*, vol. 52, pp. 499-507, 2001.
- [25] A. Scholz-Crane, "Evaluating the future: A preliminary study of the process of how undergraduate students evaluate Web sources," *Reference Services Review*, vol. 26, pp. 53-60, 1998.

- [26] A. J. Flanagin and M. J. Metzger, "Perceptions of Internet information credibility," *Journalism & Mass Communication Quarterly*, vol. 77, pp. 515-540, 2000.
- [27] M. Meola, "Chucking the checklist: A contextual approach to teaching undergraduates Web-site evaluation," *Portal: Libraries and the Academy*, vol. 4, pp. 331-344, 2004.
- [28] B. Ulicny, K. Baclawski and A. Magnus, "New metrics for blog mining," in *Defense and Security Symposium*, 2007, pp. 65700I-65700I-12.
- [29] B. Ulicny and K. Baclawski, "New metrics for newsblog credibility." in *Icwsn*, 2007, .
- [30] V. L. Rubin and E. D. Liddy, "Assessing credibility of weblogs." in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 187-190.
- [31] A. Juffinger, M. Granitzer and E. Lex, "Blog credibility ranking by exploiting verified content," in *Proceedings of the 3rd Workshop on Information Credibility on the Web*, 2009, pp. 51-58.
- [32] R. M. B. Al-Eidan, H. S. Al-Khalifa and A. S. Al-Salman, "Towards the measurement of arabic weblogs credibility automatically," in *Proceedings of the 11th International Conference on Information Integration and Web-Based Applications & Services*, 2009, pp. 618-622.
- [33] 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) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, 2011 IEEE Third International Conference On, 2011, pp. 1-8.
- [34] Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, Harald Schoen and Peter Gloor, Daniel, C. Castillo, M. Mendoza and B. Poblete, "Predicting information credibility in time-sensitive social media," *Internet Research*, vol. 23, pp. 560-588, 2013.
- [35] A. Nakamura, Y. Suzuki and Y. Ishikawa, "Clustering editors of wikipedia by editor's biases," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2013 IEEE/WIC/ACM International Joint Conferences On, 2013, pp. 351-358.
- [36] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychol. Bull.*, vol. 76, pp. 378, 1971.
- [37] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, pp. 159-174, 1977.

- [38] 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.
- [39] Ayman AL Zaatari, Rim El Ballouli, Shady Elbassuoni, Wassim El-Hajj, Hazem Hajj, Khaled Shaban, Nizar Habash, Emad Yehya, "Arabic Corpora for Credibility Analysis," LREC 2016, 23-28 May 2016, Portorož (Slovenia)
- [40] Merriam Webster. Available: <http://www.merriam-webster.com/>
- [41] Thelwall, Mike, et al. "Sentiment strength detection in short informal text." *Journal of the American Society for Information Science and Technology* 61.12 (2010): 2544-2558.
- [42] Stoyanov, Veselin, and Claire Cardie. "Annotating Topics of Opinions." LREC. 2008.
- [43] Mukherjee, Subhabrata, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. "People on drugs: credibility of user statements in health communities." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.
- [44] Pasha, Arfath, et al. "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic." LREC. Vol. 14. 2014.
- [45] Francis, James W. Pennebaker Martha E., and Roger J. Booth. *Linguistic Inquiry and Word Count*. Technical Report, Dallas, TX: Southern Methodist University, 1993.
- [46] Badaro, Gilbert, et al. "A large scale Arabic sentiment lexicon for Arabic opinion mining." *ANLP 2014* 165 (2014).
- [47] Ren, Yafeng, and Yue Zhang. "Deceptive Opinion Spam Detection Using Neural Network