AUTOMATING CREDIBILITY ASSESSMENT OF ARABIC NEWS

By

Mohamed Ibrahim Abdulla Hammad

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2013

AUTOMATING CREDIBILITY ASSESSMENT OF ARABIC NEWS

By

Mohamed Ibrahim Abdulla Hammad

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Engineering

Under the Supervision of

Assoc. Prof. Dr. Elsayed Eissa Hemayed
Associate Professor
Computer Engineering Department
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2013

AUTOMATING CREDIBILITY ASSESSMENT OF ARABIC NEWS

By

Mohamed Ibrahim Abdulla Hammad

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Engineering

Approved by the Examining Committee:

Prof. Dr. Mohamed Zaki Abdel-Megeed, External Examiner

Prof. Dr. Mohsen Abdel-Razek Rashwan, Internal Examiner

Assoc. Prof. Dr. Elsayed Eissa Hemayed, Thesis Main Advisor

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2013

Engineers Name:	Mohamed Ibrahim Abdulla Hammad
Date of Birth:	23/04/1988
Nationality:	Egyptian
E-mail:	ibra@eng.cu.edu.eg
Phone:	+2 0100 682 4347
Address:	29 El-Nasr Street, New Maadi, 11431 Cairo
Registration Date:	01/10/2010
Awarding Date:	-/-/-
Degree:	Master of Science
Department:	Computer Engineering

Supervisors:

Assoc. Prof. Dr. Elsayed Eissa Hemayed

Examiners:

Prof. Dr. Mohamed Zaki Abdel-Megeed	(External examiner)
Prof. Dr. Mohsen Abdel-Razek Rashwan	(Internal examiner)
Assoc. Prof. Dr. Elsayed Eissa Hemayed	(Thesis main advisor)

Title of Thesis:

Automating Credibility Assessment of Arabic News

Key Words:

Arabic language; credibility; information retrieval; machine learning; natural language processing; news; text analysis

Summary:

During the past few years internet has witnessed a massive increase of Arabic language users. Accompanied with this increase in the number of users is an increase in e-publishing. However, necessary laws and regulations are not yet available to control the credibility of e-published content. Furthermore, many political conflicts have risen after the Arab Spring. All of this led to an increasing demand for assessing the credibility of news in general and e-news in particular. Accordingly, some initiatives have been taken by media experts to measure the credibility of news, however, all these initiatives were completely manual and needed experts' efforts for evaluation. In this thesis, we present a system for automating credibility assessment of news articles based on two of the most important and most frequently violated criteria; (i) Does the news article indicate the source of the information? (ii) Does the news article indicate the time of occurrence of the reported event? For each of the chosen criteria, we build a text classification model to classify a news article as either violating the criteria or not. News articles previously evaluated by MCE Watch (a manual service for news credibility assessment) are used in building and evaluation of our model. Experimental evaluations show that our model can answer the first question with an accuracy that exceeds 82% and can answer the second question with an accuracy that exceeds 86%.

Table of Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

AI	Artificial Intelligence
AP	Associated Press
BOW	Bag-of-Words
CGP	Class General Perception
ESP	E-mail Specific Perception
ESPC	Egyptian Supreme Press Council
FAO	Food and Agriculture Organization
ham	not spam
KNN	K-Nearest Neighbor
MAE	Mean Absolute Error
ML	Machine Learning
MLE	Maximum Likelihood Estimator
MSA	Modern Standard Arabic
NBC	Naïve Bayes Classifer
NER	Named Entity Recognition
NN	Neural Network
RMS	Root Mean Square Error
SFS	Sequential Forward Selection
SMTP	Simple Mail Transfer Protocol
SNP	Saudi Newspapers
SVM	Support Vector Machines
SVM-NN	combination of SVM and KNN
TC	Text Categorization
TF	Term Frequency
TF	term frequency
TF-IDF	Term Frequency-Inverse Document Frequency

TF.RF	TF.Relevance Frequency
TIME-lexicon	lexicon of time-bearing word ngrams
TREC	Text REtrieval Conference

Acknowledgements

I would like to express my deepest gratitude to my supervisor Dr. Sayed Hemayed for his invaluable comments, support, guidance, and extreme patience. This research wouldn't have materialized without his encouragement and guidance through all its phases.

Dedication

To my Mother and Father, you always have been there for me; encouraging, pushing, and advising. To my sister, thank you for studying side by side with me. To my brothers, thank you all.

Abstract

During the past few years internet has witnessed a massive increase of Arabic language users. Accompanied with this increase in the number of users is an increase in e-publishing. However, necessary laws and regulations are not yet available to control the credibility of e-published content. Furthermore, many political conflicts have risen after the Arab Spring. All of this led to an increasing demand for assessing the credibility of news in general and e-news in particular. Accordingly, some initiatives have been taken by media experts to measure the credibility of news, however, all these initiatives were completely manual and needed experts' efforts for evaluation. In this thesis, we present a system for automating credibility assessment of news articles based on two of the most important and most frequently violated criteria; (i) Does the news article indicate the source of the information? (ii) Does the news article indicate the time of occurrence of the reported event? For each of the chosen criteria, we build a text classification model to classify a news article as either violating the criteria or not. News articles previously evaluated by MCE Watch (a manual service for news credibility assessment) are used in building and evaluation of our model. These news articles come from 17 different news sites and belong to different categories; Politics, Accidents, Economics, Arts, and Sports. Most of these articles are written in Modern Standard Arabic, however, a mix of both Modern Standard Arabic and the Egyptian dialect is also present. Experimental evaluations show that our model can answer the first question with an accuracy that exceeds 82% and can answer the second question with an accuracy that exceeds 86%.

Chapter 1: Introduction

Internet has become one of the main sources of our daily life news, hundreds or thousands of news articles, videos, blog posts, microblog posts, RSS or social network feeds are published daily. This big chunk of info disseminated by governments, organizations, or persons impacts heavily different aspects of our daily life including but not limited to aspects of health, education, society, economics, politics, and religion.

1.1 Motivation

During the past few years internet has witnessed a massive increase of Arabic language users with a total of 65,365,400 users as of 2010 and a growth rate of 2501.2% placing the Arabic language as the seventh most used language among all languages used on the internet and the first with respect to the growth rate [45].

Egypt - the largest Arab country in terms of population - is ranked as the 20th country over the world and the 1st among the Arabic speaking countries in terms of the number of internet users with an estimate of 21,691,776 users recorded in December, 2011 [45].

As can be seen, users in Arabic speaking countries in general and Egypt in particular are increasingly switching to producing/consuming electronic content from social networks, news sites, web blogs ... etc. Accordingly, publishers are marching in the same direction. Currently, many news publishers are providing their services online and others are providing online services beside their printed papers. However, this switch is not yet accompanied by necessary laws and regulations that control e-publishing without violating freedom of speech.

Nowadays and after two years from Egypt's January 2011 revolution, political conflicts are leaving no place in the media for an ordinary news consumer to evaluate the credibility of perceived information. This raises many questions about how technology can help an ordinary user evaluate the credibility of a piece of information before placing many (probably life changing) decisions on unreliable news.

1.2 Problem Statement

Information credibility is an important problem that has been considered previously by many researchers in media for newspapers, television and online news [1, 22, 39, 42, 46, 49, 59]. Researchers conducted surveys on what is perceived as credible. Nowadays in Egypt, a lot of political conflicts resulted in non-credible information being circulated in different news sites. This led media experts to take initiatives whose goal is to assess the credibility of news articles and their sources i.e. news sites. However, all these initiatives required manual experts' efforts in measuring the credibility. From there, we found an urgent need to automate such manual work. Thus, the problem can be stated as automating the credibility assessment of Arabic news.

1.3 Objectives

The main objective of this thesis is to provide a solution that automates the credibility assessment of Arabic news on the internet. This involves finding answers to the following questions:

1. Does the news article indicate the source of the information?
2. Does the news article indicate the time of occurrence of the reported event?

1.4 Publications

This work resulted in the following scientific outcome:

Mohamed Hammad and Elsayed Hemayed, "Automating Credibility Assessment of Arabic News", 5th International Conference on Social Informatics (SocInfo 2013), Kyoto Japan, 25-27 November 2013.

1.5 Organization of Thesis

The rest of this thesis is organized as follows. In Chapter 2 we present a background about the Arabic language and challenges facing researchers dealing with it. We also present in the same chapter a background about machine learning for text classification. In Chapter 3 we present our literature survey. It includes a review about information credibility, and text classification tasks. The architecture and details of our system are presented in Chapter 4. Chapter 5 presents an analysis of the dataset and the details and results of the conducted experiments. Finally, Chapter 6 presents our conclusions and future work.

Chapter 2: Background

In this chapter we present background information necessary to understand different issues in the upcoming chapters. We start by giving an overview of the Arabic language and challenges in automated processing of Arabic text. We follow this by the description of machine learning approaches to text classification.

2.1 Arabic Language

In this section we present an introduction about the Arabic language giving statistics about the language and its speakers. We also present challenges facing researchers who deal with Arabic text.

Arabic language is a semitic language (one of a set of languages spoken by people across Western Asia, North Africa, and the Horn of Africa) [69]. Arabic language is spoken - as of 2010 - by 293 million persons representing 4.23% of the world's population. It is ranked as the 5th most spoken language among all languages in the world; it is preceded by Mandarin, Spanish, English, and Hindi. Arabic language is mainly spoken in the Middle East and North Africa [66].

Modern Standard Arabic (MSA) is an updated version of Classical Arabic - the language of Quran. MSA is the form of Arabic currently in formal use by educated people in their writings. For example: daily newspapers, books, and governmental institutions all use MSA in their daily writings, schools and universities also teach MSA. On the other hand, native speakers use different forms of Arabic called dialects in their daily life conversations. Each Arabic speaking region has its own dialect e.g. Egyptians speak the Egyptian dialect [63] and Lebanese speak the Lebanese dialect [65]. Egyptian dialect is considered as the most spoken Arabic dialect with an estimate of 54 million speakers.

Arabic script has many interesting characteristics that pose challenges to researchers working on automated Arabic text processors. Arabic is considered as a morphologically rich language where each word can consist of multiple morphemes ("word parts that have independent meaning but may or may not be able to stand alone" [68]). An Arabic word can contain different kinds of affixes.

- يلعب (meaning "He plays") is an example of the verb لعب (meaning "He played") after attaching the prefix ي to indicate the present tense.

- اللعبة (meaning "The game") is an example of the noun لعبة (meaning "Game") after attaching the prefix ال (meaning the definite article "the")
- فهموا (meaning "They understood") is an example of the verb فهم (meaning "He understood") after attaching the suffix وا to indicate the verb is plural.
- ونباركهم (meaning "and we bless them") is an example of the verb بارك (meaning "He blessed") after attaching the prefix و (meaning "and"), the prefix ن (meaning the pronoun "we"), and the suffix هم (meaning the pronoun "them").

Arabic script does not have the notion of vowel letters instead vowels are represented by diacritics which are small marks that appear below or above the letters. For example:

- Fathah: Adding Fathah above the letter ب to become بَ changes its pronunciation to *ba*.
- Kasrah: Adding Kasrah below the letter ب to become بِ changes its pronunciation to *be*.
- Dammah: Adding Dammah above the letter ب to become بُ changes the pronunciation to *bu*.

These diacritics are rarely used in MSA writings nowadays which adds lots of challenges during reading and interpreting the written text. Writings rely on the reader to interpret the text correctly based on their deep knowledge of the language.

Arabic script does not have the notion of capitalization neither at the beginning of a sentence nor at the beginning of a named entity. Arabic script also does not have strict punctuation rules. Together the lack of capitalization and strict punctuation in Arabic leads to challenges in finding sentence boundaries as well as in Named Entity Recognition (NER) [20].

2.2 Machine Learning

In this section we give an overview of Machine Learning (ML) highlighting example applications, emphasizing classification problems, and giving description of some of the most used ML algorithms.

ML is the branch of computer science that deals with learning from data. ML is considered as a subbranch of Artificial Intelligence (AI) . Nowadays, ML is a core component of many applications including but not limited to Page Ranking, Machine Translation, Information Retrieval, Pattern Classification, Recommendation Systems, Advertising, Forecasting and many others.

Learning from data can be classified into 3 main approaches:

- Supervised Learning: given a data set of example inputs and their corresponding outputs, supervised learning tries to infer from these examples a mapping function that maps inputs to outputs. The learned function should be able to generalize to unseen examples and not just be limited only to correctly map the learned examples.
- Unsupervised Learning: given a data set of example inputs without output labels, unsupervised learning tries to partition data according to hidden properties in the data that may be similar between parts of the data.
- Semi-supervised Learning: is a combination of the above two learning methods. It utilizes the presence of a small amount of labeled examples and augment them with unlabeled data to learn better functions.

In this writing we focus on the first type – supervised learning. Supervised learning can learn mapping functions that map to discrete or continuous output labels. In case of continuous output labels, we call the problem a regression problem, however, in case of discrete output labels, we call the problem a classification problem. A special case of the classification problem is the Binary Classification problem which is one of the most frequently studied problems in machine learning.

2.2.1 Binary Classification Problem

The binary classification problem is the task of classifying items into one of two classes. One can come up with many example applications for this specific instance of ML problems such as the task of classifying tumors as either malignant or benign or the task of classifying emails as either spam or not.

Formally, a binary classification problem can be represented by giving a set of examples $X = \{x_1, x_2, ...x_n\}$ each with its corresponding output label $Y = \{y_1, y_2, ...y_n\}$ such that

$y_i \in \{0,1\}$. This labeled set of examples is called a training set. A learning algorithm is then applied to learn a mapping function from $X \rightarrow Y$ called $h_\theta(x)$. Given an unseen test case the learned mapping function should be able to predict the correct output label. Figure 2.1 presents an example of a binary classification problem. As can be seen from the figure, the learning algorithm did learn a linear function (the red straight line) that can perfectly separate the training examples of Class 1 from those of Class 2. Problems that are separable by a line in 2-D, plane in 3-D or hyperplane in 4 (or more)-D are called *linearly separable*. Each of the 2 dimensions in Figure 2.1 represents a feature or a signal of the training example. For example if the points represent tumor then the dimensions can represent the tumor size and cell shape. The data points are not always linearly separable as those in Figure 2.1 instead the learning algorithm may need to learn a more complex function to be able to separate the points from the classes efficiently. An example of this case is shown in Figure 2.2.

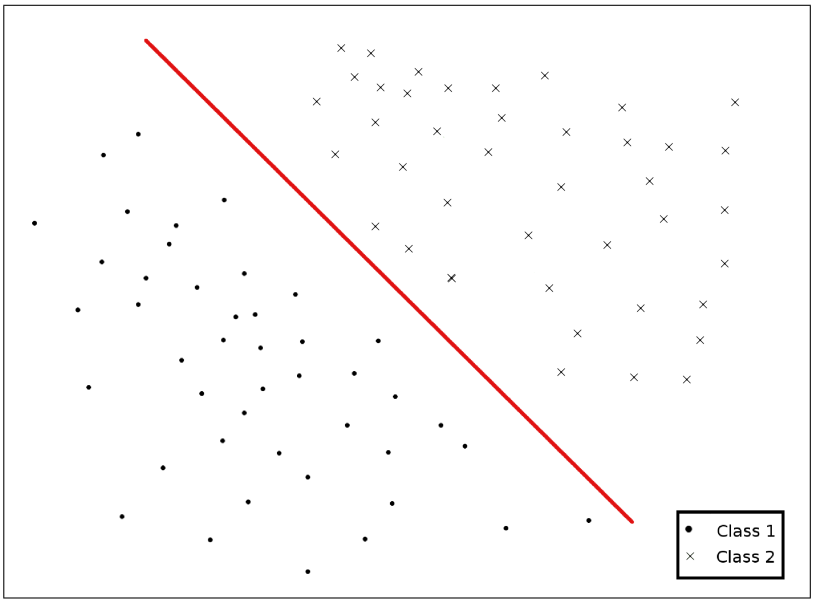


Figure 2.1: An example of a linear classifier classifying instances of two classes in a binary classification problem.

The data points are not always perfectly separable as those in Figure 2.1 and in Figure 2.2 instead the learned function may only be able to separate a portion of these points due to noise in the training set. In this case the task of the learning algorithm is to minimize the classification error. An example of such case is shown in Figure 2.3 with the best classifier indicated by the text “The Best One”.

There are multiple learning algorithms that have been used for classification problems in the literature. In the next subsection we describe some of them.

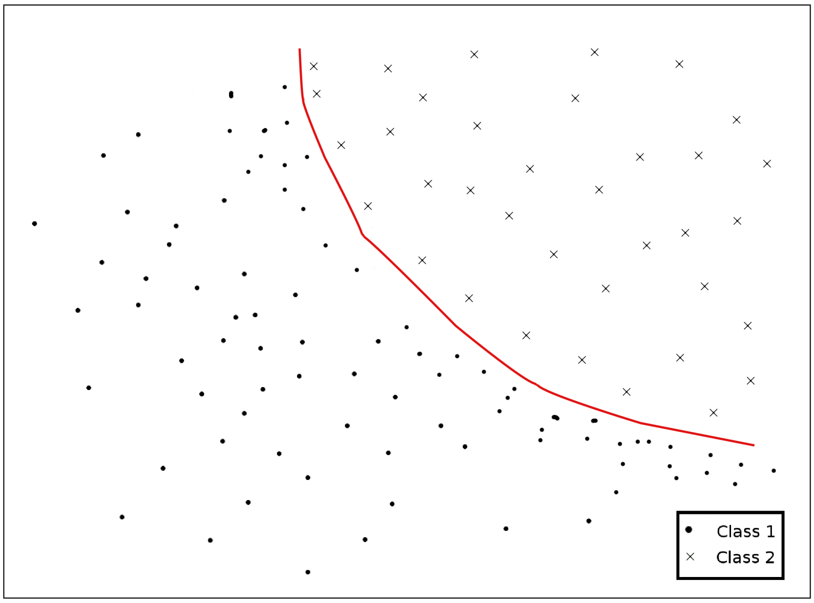


Figure 2.2: An example of a non-linear classifier classifying instances of two classes in a binary classification problem.

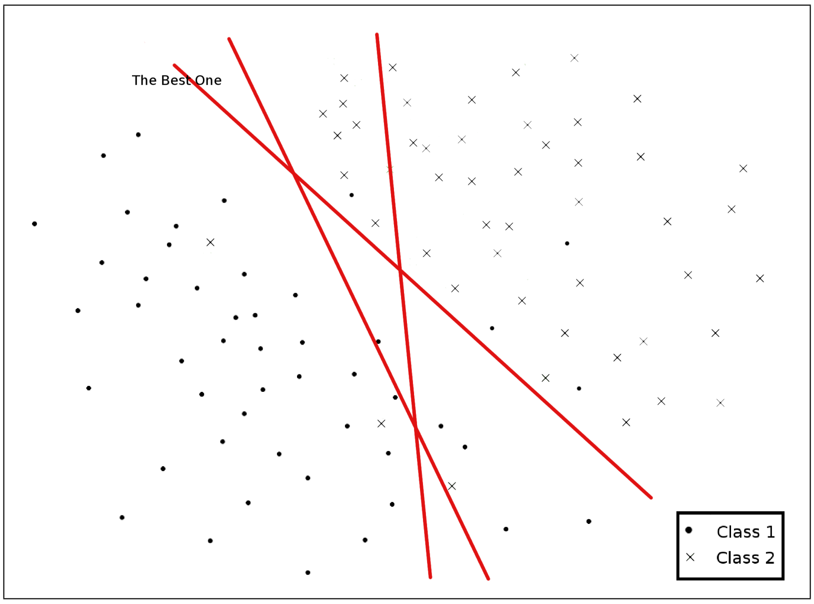


Figure 2.3: An example of a training set with noise. Multiple classifiers are shown that try to classify the data. The best one is marked and is the one that minimizes the classification error.

2.2.2 Learning Algorithms

In this subsection we describe multiple learning algorithms that have been used in binary classification problems. We present them in this order K-Nearest Neighbor (KNN) [43, 56, 64], Naïve Bayes Classifer (NBC) [43, 56, 67], and Support Vector Machines (SVM) [43, 56, 71].

2.2.2.1 K-Nearest Neighbor Classifier

KNN is one of the simplest classifiers used in the literature. For a test sample, KNN computes the distance between the test sample and all the training examples. The distance is computed using a distance function like Euclidean distance for example. Next, KNN finds the k nearest training examples to the test sample and performs a vote among these k examples based on the distance from the test sample and each of them. The test sample is finally classified as belonging to the class most common among the k nearest neighbor – k is a parameter that is supplied to the algorithm.

KNN performs the above classification procedure for each test sample, there is no need for a training phase, classification is performed directly as indicated above. An example of applying KNN is shown in Figure 2.4. As can be seen, the test sample will be classified as

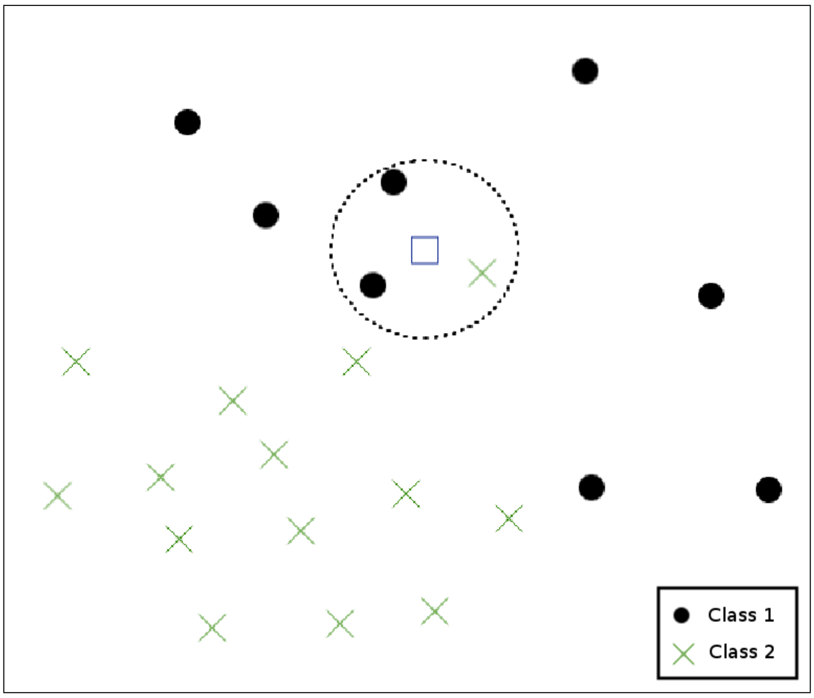


Figure 2.4: An example of applying KNN classifier in a binary classification problem. The blue square represents the test sample, class 1 (C1) is represented by green crosses, and class 2 (C2) is represented by filled black circles. The dotted circle shows the nearest neighbors for $k = 3$

belonging to C1 since two points of the three nearest neighbors are belonging to C1 versus only one point belonging to C2, thus the test sample shall be classified as belonging to C1.

Although being simple, KNN is sensitive to noise in the training data thus for noisy data KNN's performance degrades. KNN is also sensitive to the weighting of the features so the features should be weighted appropriately to reflect their relative importance. In addition to that KNN suffers from the need of intensive computations especially in the case of large training sets (or high dimensional feature spaces) since, for each test sample, it computes the distance to all the training samples.

2.2.2.2 Naive Bayes Classifier

NBC is a simple probabilistic classifier that assumes the value of one feature is independent of other features (independence assumption). For example, in case of classifying a document with NBC given that the features extracted from the document are the term frequencies, NBC assumes that the frequency of occurrence of any term is independent from all other terms in the document. Although, this oversimplifying assumption seems naive but NBC performs well in practice. Given a test sample X (a document for example) that should be classified into one of two classes $C = \{c_1, c_2\}$, NBC performs:

$$c_{map} = \arg\max_{c \in C} P(c|X) \tag{2.1}$$

Equation 2.1 means that we classify the document X to the most likely class or the class that has the maximum a posteriori probability (c_{map}). Bayes Probability Theorem can be written as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{2.2}$$

The test document X can be represented by multiple features $X = \{x_1, x_2, ...x_{n_X}\}$, these features can be, for example, the frequency of occurrence of each token from a vocabulary in the document X with n_X being the number of tokens in X. Substituting with Equation 2.2 in Equation 2.1 gives:

$$\begin{aligned} c_{map} &= \arg\max_{c \in C} \frac{P(c)P(X|c)}{P(X)} \\ &= \arg\max_{c \in C} \frac{P(c)P(x_1, x_2, ...x_n|c)}{P(X)} \end{aligned} \tag{2.3}$$

$P(c)$ is called the a priori probability since it can be estimated from the training data directly prior to looking at the test sample. The term $P(c)$ indicates that we put appropriate weight to each class according to its appearance in the training set thus if a test sample has features that makes it equally likely to be from either class then – using the apriori term – it will be classified to the most frequent class in the training set. The term $P(X)$ will have the same value for both classes so it can be neglected without affecting the classification. Thus c_{map} can be re-written as:

$$c_{map} = \arg\max_{c \in C} P(c)P(x_1, x_2, ...x_n|c) \tag{2.4}$$

If we further applied Bayes Theorem to the term $P(x_1, x_2, ...x_n|c)$, we get:

$$\begin{aligned} c_{map} &= \arg\max_{c \in C} P(c)P(x_1|c)P(x_2, x_3, ...x_n|c, x_1) \\ &= \arg\max_{c \in C} P(c)P(x_1|c)P(x_2|c, x_1)P(x_3, x_4, ...x_n|c, x_1, x_2) \\ &= \arg\max_{c \in C} P(c)P(x_1|c)P(x_2|c, x_1)...P(x_n|c, x_1, x_2, ...x_{n-1}) \end{aligned} \tag{2.5}$$

Using the independence assumption we get:

$$c_{map} = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P(x_i|c) \tag{2.6}$$

Equation 2.6 can be regarded as the final form of the NBC. Both terms $P(c)$ and $P(x_i|c)$ can be computed from the training data using, for example, the Maximum Likelihood Estimator (MLE) or any other density estimation procedure.

Theoretically, NBC is regarded as the optimal classifier given the exact values for the a priori and the conditional probability terms. However, in practice, NBC is not optimal that is because the value of each of these probability terms is estimated from the training data with an accompanied estimation error which degrades the classification performance.

2.2.2.3 Support Vector Machines

SVM is another example of supervised learning algorithms that has been used extensively in classification problems. Given training data of example inputs and outputs, SVM tries to find a hyperplane that separates the data such that the separation maximizes the distance between the closest point from each class and the hyperplane.

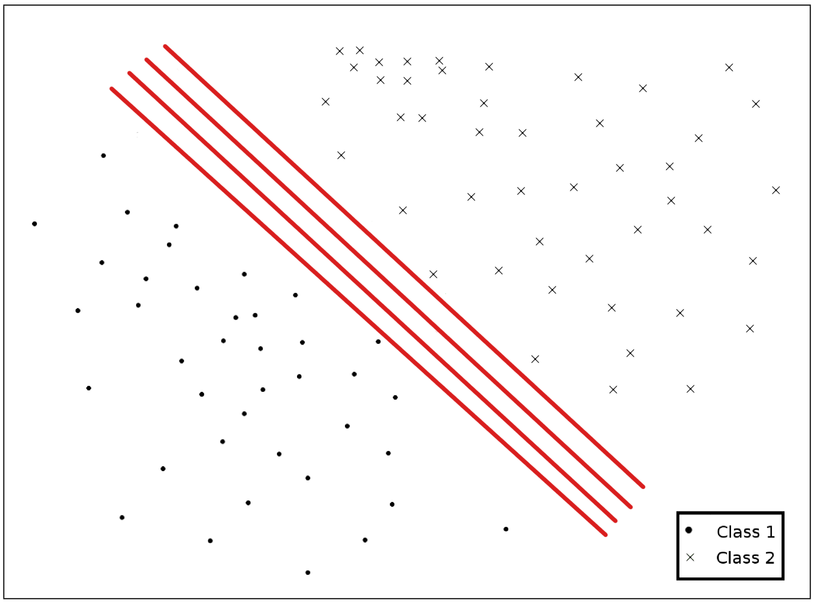


Figure 2.5: An example of multiple linear classifiers that correctly classify the examples from the 2 classes.

Figure 2.5 gives an example of multiple classifiers that can correctly classify the training data. Each of these classifiers separate the data such that the data belonging to one class are on one side of the classifier and the data belonging to the other class are on the other side of the classifier. Although, all of them correctly classify the training data but classifiers that are closer to one class are more prone to misclassifying further unseen data of this class that may lie very near to the decision boundary but on its other side. This decreases the ability of the classifier to generalize. An example of such case is shown in Figure 2.6, the classifier will classify the test sample as belonging to class 1 because it is on the same side as the training samples of class 1. On the other hand, the distance is

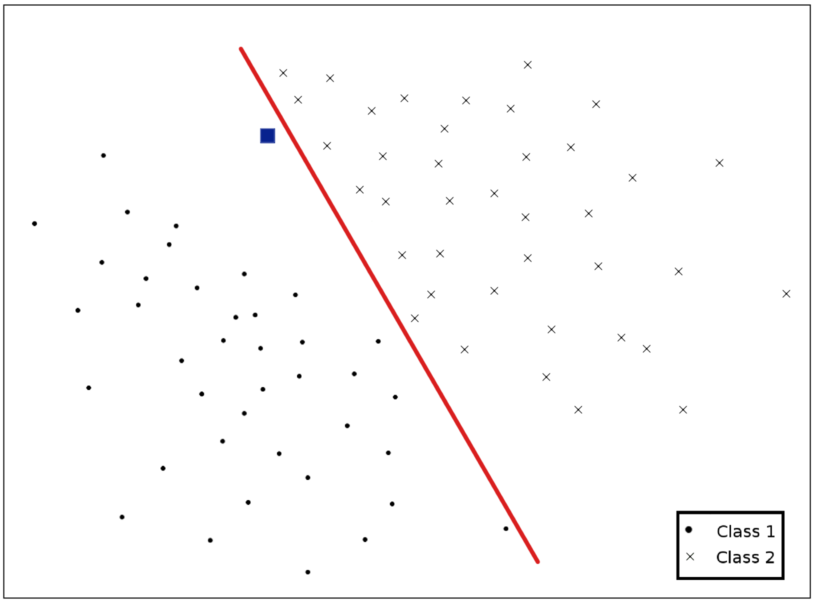


Figure 2.6: An example of a misclassified test sample (blue square) due to small margin between the classifier and the closest training point in class 2.

much smaller between the test sample and the training points of class 2 than that between the test sample and the training points of class 1. This error is caused by the small margin between the decision boundary and the nearest point in class 2.

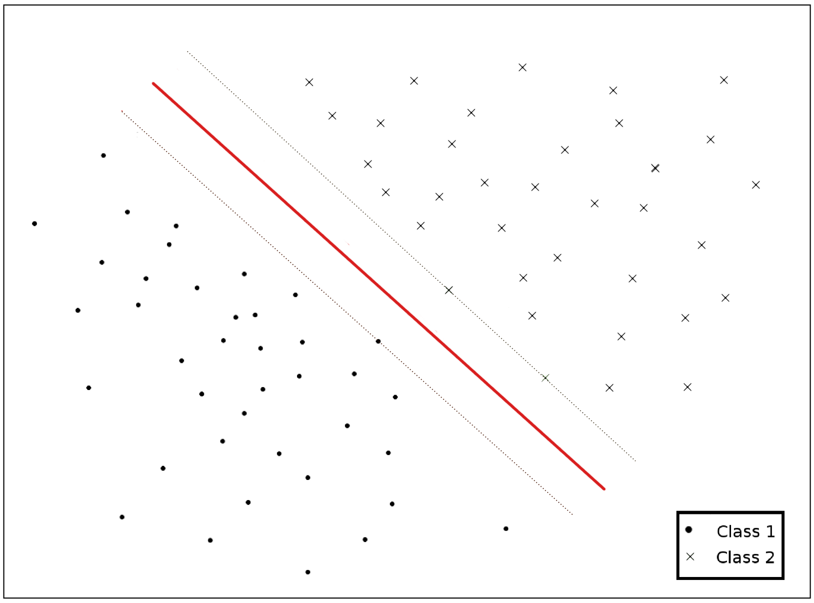


Figure 2.7: An example of SVM classifier shown in solid red color. The margins are shown in dotted lines.

Figure 2.7 shows an example of SVM classifier classifying the same data in Figure 2.5. As can be seen from the figure, the SVM classifier maximizes the distance from the nearest points on either sides of the classifier. The dotted lines in the figure act as the margins for this distance. The margins pass through the closest points to the decision boundary from both classes, these points are called the support vectors. The distance bordered by the margins is hence called the margin and it is maximum in this case.

2.2.3 ML for Text Classification

Text classification is a classification problem that involves classifying text into two or more classes. Examples of such text classification problems are:

- Text Categorization: the task of classifying text excerpts e.g. news articles into different categorizes e.g. Sports, Economics, and Politics
- Spam Filtering: the task of classifying emails into either being spam or not.
- Sentiment Analysis: the task of classifying subjective text as either expressing positive or negative sentiment.

Having presented ML problems in the previous section and specifically binary classification problems, text classification problems are regarded as an important example of such ML problems. The same learning algorithms presented in Section 2.2.2 are used in classification problems. However, in order to apply such learning algorithms one has to transform/represent the text in such a way to be able to input it to the learning algorithm. In the following sections we will present the typical steps followed in text classification problems and the most common approaches, from literature, to implement each step.

2.2.3.1 Text Preprocessing

The first step is to preprocess the input text whether it is in the form of articles, paragraphs, sentences...etc. Typical preprocessing involves one or more of the following steps:

- Tokenization: is the process of splitting the input text into tokens.
- Normalization: is the process of transforming one form of a letter into another form mostly because both forms are erroneously used interchangeably, examples include:
  - Replacing أ, آ and إ by ا

  - Replacing ىء by ئ
  - Replacing ى by ي
  - Replacing ة by ه
  - Removing بال, فال, وال, كال, ال and لل
  - Removing punctuation marks, diacritics, non-letters
  - Stopwords elimination

- Stemming: is the process of reducing words to their root form.

2.2.3.2 Text Representation

After the preprocessing step we end with a collection of words representing the preprocessed document. The next step is to represent this collection in a format suitable to input to learning algorithms. Of the most common representations:

- Bag-of-Words (BOW) model: a document is represented as an unordered collection/bag of words where each word is weighted using a weighting scheme. For example, the sentence "The quick brown fox jumped over the lazy dog" can be represented with a BOW model as follows:

the: w_1	quick: w_2
brown: w_3	fox: w_4
jumped: w_5	over: w_6
lazy: w_7	dog: w_8

- n-gram model: an n-gram is a contiguous sequence of n tokens. An n-gram model represents a document with of m tokens with $m-(n-1)$ n-grams where each one of them is weighted using a weighting scheme. For example, the sentence "The quick brown fox jumped over the lazy dog" can be represented with a 4-gram model as follows:

<the, quick, brown, fox>: w_1	<quick, brown, fox, jumped>: w_2
<brown, fox, jumped, over>: w_3	<fox, jumped, over, the>: w_4
<jumped, over, the, lazy>: w_5	<over, the, lazy, dog>: w_6

There are different weighting schemes among them are:

- Binary: the weight is 1 if the token is present one or more times in the document and 0 otherwise.
- Term Frequency (TF) : the weight corresponds to the number of times a token is found in the document.
- Term Frequency-Inverse Document Frequency (TF-IDF) : the TF is weighted by the inverse document frequency of the term. Thus terms appearing in a large number of documents have lower weight than those appearing in a small number of documents.

2.2.3.3 Classification

The last step is applying one of the learning algorithms described in Section 2.2.2 for training and then testing with unseen test set.

Chapter 3: Review of Literature

In this chapter we provide a review of the past work on information credibility in general and efforts towards automating credibility assessment for different information sources e.g. blogs, microblogs, and news. We also present a survey on approaches taken to solve different text analysis problems.

3.1 Information Credibility

In this section we present multiple efforts for credibility assessment. First, we present media researchers efforts to measure the perceived credibility of newspapers, TV, and internet. Next, we present multiple works that aim at automating credibility assessment of different content sources on the internet. Afterwards, we present a couple of initiatives taken by media researchers in Egypt during the past couple of years to address the increasingly important problem of news credibility assessment.

3.1.1 Manual Credibility Surveys for Newspapers, TV, and Internet

For long credibility of information disseminated via newspapers, television and internet has been a subject of research considered by media researchers. Gaziano and McGrath [22] conducted telephone surveys to construct an overall credibility score for newspapers and television news. A factor analysis termed credibility of 12 items showed whether newspapers and television news are fair, are unbiased, are accurate, are factual, are concerned about the community's well-being, are concerned about the public interest, tell the whole story, respect people's privacy, watch out after people's interests, separate fact and opinion, can be trusted, and have well-trained reporters. The survey showed that television and newspaper credibility scores were correlated moderately with each other suggesting that people's attitudes towards these media are similar to a certain degree. Other questions were asked to assess credibility of newspapers and television news during conflicting events, and with regard to different topics.

Abdulla et al. [1] extended the work of Gaziano and McGrath [22] to include online news in addition to newspapers and television news. The survey respondents rated online news highest in credibility. Newspaper readers indicate that newspapers to offer credibility must be balanced in storytelling, complete in providing information, objective and fair, accurate, and unbiased. Television viewers want news that is fair, balanced, trustworthy, accurate, objective, complete, believable, unbiased and honest. For online news to be credible, it also must be trustworthy, believable, accurate, complete, balanced and fair, and honest.

3.1.2 Automating Credibility Assessment

During the past few years, automating credibility evaluation of websites, web blogs, microblog feeds and news articles received much attention from computer science researchers.

3.1.2.1 Automating Credibility Assessment of Webpages

Multiple works considered automatic assessment of the credibility of webpages. Akamine et al. [2] used appearance information such as number of sentences, number of images, advertisements, presence of contact address, and privacy policy to assess the credibility of Japanese websites.

Xu et al. [73] relied on trust features of a website to indicate its credibility. Trust features like:

- Appearance information
- Owner of website: governments, universities and reputable organizations account for higher credibility
- Semantic contents of the website assessed by querying a search engine with the contents of the webpage; commonality of the content accounts for credibility.

Schwarz and Morris [53] used On-Page features, Off-Page features and aggregate features. On-Page features such as spelling errors, advertising and domain types, Off-Page features such as awards received by a website, PageRank, and information about how frequent a webpage was shared by other visitors, and aggregated features such as:

- General Popularity: the number of unique users visiting a webpage in a time period.
- Expert Popularity: a classification technique was used to identify experts in a topic and then calculate the number of visits of related pages in a time period.
- Geographic Reach: calculated based on the locations of site visitors using zip code information.

These features were visualized to the user as side-by-side with the webpage or as a chart with the search results to help users assess credibility. Users rated credibility of webpages and search results before and after visualization. The results indicated that the visualization increased the accuracy of the users' credibility ratings and their confidence in it. The results also show that this impact was much higher for search results than webpages; this was attributed to the page design that greatly influences a user's judgment.

Kawahara et al. [31] developed a system that grasps major statements and their contradictions to support the credibility analysis of web contents. Major statements are defined as linguistic expressions occurring with a high frequency in the set of Web pages on a given topic. Contradictions are the statements that contradict the major statements. The unit of statement used is predicate-argument structure (who does what). The following 3 steps conclude the approach:

- Extracting predicate-argument structures. This is done by extracting important sentences (those neighboring topic words), doing morphological analysis and then filter functional and meaningless predicate-argument structures.
- Merging synonymous and subsumed predicate-argument structures
- Detecting major and contradicting predicate argument structures. Major ones have high frequencies, their negations are the contradictions.

The work of Akamine et al. [2] and Kawahara et al. [31] was combined into WISDOM [3] which is a Japanese system that evaluates the credibility of information on the web from multiple viewpoints.

3.1.2.2 Automating Credibility Assessment of Blogs

Automating credibility assessment of blogs was also interesting for many researchers. Weerkamp and de Rijke [62] used post-level indicators, blog-level indicators or a combination of both to rank blog posts during retrieval. Post-level indicators are like text quality (capitalization, emoticons, shouting, spelling, and punctuation), post length, and timeliness with news. Blog-level indicators like spamminess, comments, topical consistency, and expertise. They use these indicators to perform ranking of the blog posts given a query. Two ranking algorithms were evaluated; credibility-inspired that takes only credibility into account by ranking the top *n* results of a baseline by credibility and a combined reranking that multiplies the credibility-inspired score of the top *n* results by their retrieval score. Both ranking algorithms led to larger improvements over the baseline with the credibility-inspired ranking giving higher precision scores.

Juffinger et al. [29] used similarity with verified content to indicate blog credibility. A news corpus is used as the verified content. For a query in a corpus of indexed blogs they do the following: (i) they observed that the temporal distribution of news articles and blog posts indicate a correlation between them; usually blog posts are done to comment on news. They utilized this correlation to filter out blogs with significantly different quantity structure compared to the news corpus (ii) the remaining blog posts are then compared to the news corpus using cosine similarity based on the nouns and verbs and one of three credibility levels is assigned to each one.

Al-Eidan et al. [5] proposed a technique that combines blog-level and post-level features from Weerkamp and de Rijke [62] together with the similarity with verified content from Juffinger et al. [29] to assess the credibility of Arabic blogs. They considered content from Aljazeera [8] and Saudi Press [51] as verified.

3.1.2.3 Automating Credibility Assessment of Microblogs

Microblogs e.g. Twitter credibility assessment has witnessed an increased attention from researchers during the past couple of years. Approaches use message based or content based features such as length of tweet, number of hashtags, lack or presence of inappropriate words, sentiment polarity in a tweet, similarity with verified content, presence of urls, retweets, special characters, emoticons, and @ mentions. Other approaches use source-based or user-based features such as number of followers, number of followings, ratio between followers and followings, if the account is verified by Twitter.com, age of user, and screen name. Al-Eidan et al. [4] combined message and source based features to classify Arabic tweets into three credibility classes, low, moderate or high. The main feature used is the similarity with verified content [29] using content from Aljazeera [8] and Saudi Press [51] as verified. Tweets together with verified content are entered to their system. First, tweets and news articles are preprocessed; preprocessing includes normalization, stop words removal, part of speech tagging and stemming. Next, all features are computed and a credibility score is calculated for tweets by weighing the value of each feature. They compared the output of their system with humanly annotated tweets for credibility and calculated precision and recall. The average precision reported is 0.52 and average recall is 0.56.

Gupta and Kumaraguru [24] used message-based features and user-based features to rank tweets by credibility during high impact events. They used a rank-based SVM to rank tweets by credibility. They compared their ranking versus Twitter's default inverse chronological rank on a set of 5,578 human annotated tweets. They concluded that extraction of credible information from Twitter can be automated with high confidence.

Castillo et al. [13] assessed the credibility of newsworthy tweets by first classifying tweets into two classes; either newsworthy or just a conversation and then classifying newsworthy tweets as credible or not. In addition to using message-based and source-based features, the authors adopted topic-based features and propagation-based features such as the degree of the root of the propagation tree and depth of a propagation tree. They tried multiple classifiers; J48 decision tree performed best with accuracy of 86% for credibility classification.

Kang et al. [30] used content-based, social-based features and a hybrid of both to assess topic specific credibility of tweets. J48 decision tree classifier was used to classify tweets into credible or not. Evaluation was done across a collection of tweets from 6 different topics and results were compared with manually assessed tweets. The best performance was 88.17% accuracy obtained by the social-based model.

3.1.2.4 Automating Credibility Assessment of News sites

Work done for automating credibility assessment of news sites didn't receive as high attention from researchers. Available works relied on transforming the problem into measuring the sentiment bias across multiple dimensions, visualizing that, and leaving credibility judgment to the users. Zhang et al. [75] developed a system that:

- Detects sentiment of a news article across four dimensions: $Joy \Leftrightarrow Sadness$, $Acceptance \Leftrightarrow Disgust$, $Anticipation \Leftrightarrow Surprise$, and $Fear \Leftrightarrow Anger$
- If a website is selected, a graph representing the sentiment bias along each of the 4-D is generated for each of the related sub-topics.
- If a subtopic is selected and no website, a graph representing the sentiment bias according to the selected subtopic is generated. It shows the bias according to the selected subtopic for all websites along each of the 4-D.
- If both a subtopic and a website are selected, a graph representing the sentiment bias according to the selected subtopic is generated. It shows the bias for each news article in the website along each of the 4-D

Kawai et al. [32] calculated the sentiment of each news site along another four sentiment dimensions: $Bright \Leftrightarrow Dark$, $Acceptance \Leftrightarrow Rejection$, $Relaxation \Leftrightarrow Strain$, and $Anger \Leftrightarrow Fear$. The sentiment of a news site is calculated as the average sentiments of the news articles from this site concerning a topic. The sentiment is then visualized on a Google map.

For multimedia news, Xu et al. [74] developed a material-opinion model to rank multimedia news by credibility. They analyzed video images (material) for the sequence of appearance of a target stakeholder (Obama for example) and measure the dissimilarity between the same news video presented by other stations. In addition they analyzed the surrounding text and captions of the video for opinion and again measure the dissimilarity from other stations. The combination of both material and opinion dissimilarity scores is used to assign a credibility score. The credibility scores were evaluated manually by users presented with the videos, captions and the calculated scores.

To sum up, most research in automating credibility measurement used domain specific features to indicate credibility but none tries to extract content features from the articles which can be used to signal different credibility aspects. In addition, the work done to automate credibility assessment of news content can be concluded in the works of Zhang et al. [75] and Kawai et al. [32] which analyzes the sentiment bias in news articles, visualizes that and leaves credibility judgment to the users. Furthermore, research in automating credibility measurement for Arabic content can be concluded in the work done by Al-Eidan [4, 6] for blogs and Twitter; both of them used news content from specific sources as verified content which doesn't fit our purpose of evaluating credibility for news content itself.

As can be seen, the presented findings until now doesn't signal any clues on how to come up with credibility judgments for news articles. However, recently multiple initiatives have been taken by media experts in Egypt to assess the credibility of news. The two upcoming subsections (3.1.3 and 3.1.4) will present examples of such initiatives.

3.1.3 Egyptian Supreme Press Council Initiative

The Egyptian Supreme Press Council (ESPC) released on May 14th, 2013 a report describing a study made to assess the credibility of 18 different Egyptian newspapers based on a total of 35 criteria extracted from the news articles. The criteria can be grouped into 4 different categories:

- Bias in presenting the information and lack of integrity.
- Misleading the public opinion and lack of precision.
- Inciting hatred and violence.
- Defamation and libel.

The study, which included professors from the faculty of Journalism and the faculty of Economics at Cairo University, manually analyzed articles published in the 18 newspapers in the period between January 20 and February 12, 2013 for violations of the criteria. Conclusions of the report came to that all newspapers violate the criteria but with different degrees with AlAhram being the least newspaper violating the criteria among all the considered newspapers with a score of 0.09 violation per edition, see Figure 3.1 for the score of all the newspapers.

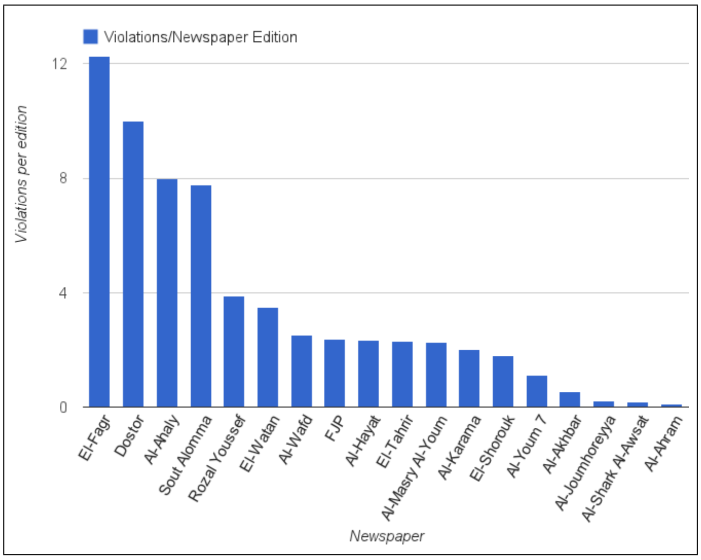


Figure 3.1: Results of ESPC study on the credibility of Egyptian newspapers.

3.1.4 Media Credibility in Egypt Initiative

In this section we describe in detail MCE Watch [44]; a manual service for credibility assessment of Arabic News. We start by giving a description of MCE Watch in 3.1.4.1. After that we analyze the pros and cons of such service in 3.1.4.2. Finally, we sum up our information about MCE watch in 3.1.4.3.

3.1.4.1 Description

MCE Watch is a service launched by a group of youth volunteers who don't belong to any political directions or parties. MCE Watch describes their service by:

> *"A project that has been established to develop media with trustworthiness and professionalism that promotes the journalism and media work in Egypt. This is achieved by accurate monitoring and control of various media by a team of experts in the world of journalism and media in cooperation with the browsers and followers of different news websites activating the concept of public monitoring and combining it with professional media monitoring."*[1]
> [44]

MCE Watch provides credibility monitoring services for news articles posted on 17 different Egyptian news sites; these sites are listed in table Table 3.1. A user of MCE Watch can submit a URL of any news article – published on any of the news sites supported by MCE Watch – for credibility judgment. A media expert from MCE Watch receives the request and analyzes the news article for potential credibility issues. After a time period the expert replies with his judgment and analysis that is posted on MCE Watch. The expert's reply indicates whether the news article is credible or not. Credibility of news articles is assessed based on preset criteria published by MCE Watch, these criteria are detailed in Table 3.2. If the article is not credible, reasoning is provided by the expert detailing his judgment and listing credibility criteria violated by the article.

Figure 3.2 shows an example from MCE Watch of a news article that was judged as non-credible. The article comes from *elwatan* news site. It violates three credibility criteria:

- *indicates_how_info_got*
- *answers_when*
- *has_source_info*

A reasoning about these violations is also provided by the expert.

[1] English translation of their Arabic description.

Table 3.1: News sites monitored by MCE Watch.

Arabic Name	**URL**	**Alias**
أخبار مصر	egynews.net	egynews
المصري اليوم	almasryalyoum.com	almasryalyoum
شبكة رصد	rassd.com	rassd
بوابة الحرية والعدالة	fj-p.com	fj-p
الشروق	shorouknews.com	shorouk
مصراوي	masrawy.com	masrawy
الوفد	alwafd.org	alwafd
البديل	elbadil.com	elbadil
الدستور الأصلي	dostorasly.com	dostorasly
اليوم السابع	youm7.com	youm7
الأهرام	ahram.org.eg	ahram
التحرير	tahrirnews.com	tahrir
أخبار اليوم	akhbarelyom.org.eg	akhbar
المصريون	almesryoon.com	almesryoon
الدستور	dostor.org	dostor
الفجر	elfagr.org	elfagr
الوطن	elwatannews.com	elwatan

Figure 3.3 shows an example from MCE Watch of a news article that was judged as credible. The article comes from *youm7* news site. Being credible, it doesn't have any violations and it doesn't need any expert reasoning.

MCE Watch provides a credibility score for each news site. The score is calculated monthly based on the articles assessed during that month. All news sites are 100% credible at the beginning of a month. Each news site is assigned a pool of 500 points at the beginning of the month. Every article that is assessed as non-credible deducts points from this pool. The amount of deducted points corresponds to the violations done by the news article; Table 3.2 shows the amount of points assigned to each credibility criteria. The credibility of a news site at any point of time is calculated as a percentage of the remaining points, Equation 3.1.

$$score_{news\_site} = \frac{500 - \sum_{i=1}^{n} \sum_{j=1}^{m} v_{ij} * p_{ij}}{500} * 100\% \quad (3.1)$$

where:

- *news_site* is one of the 17 news sites supported by MCE Watch.

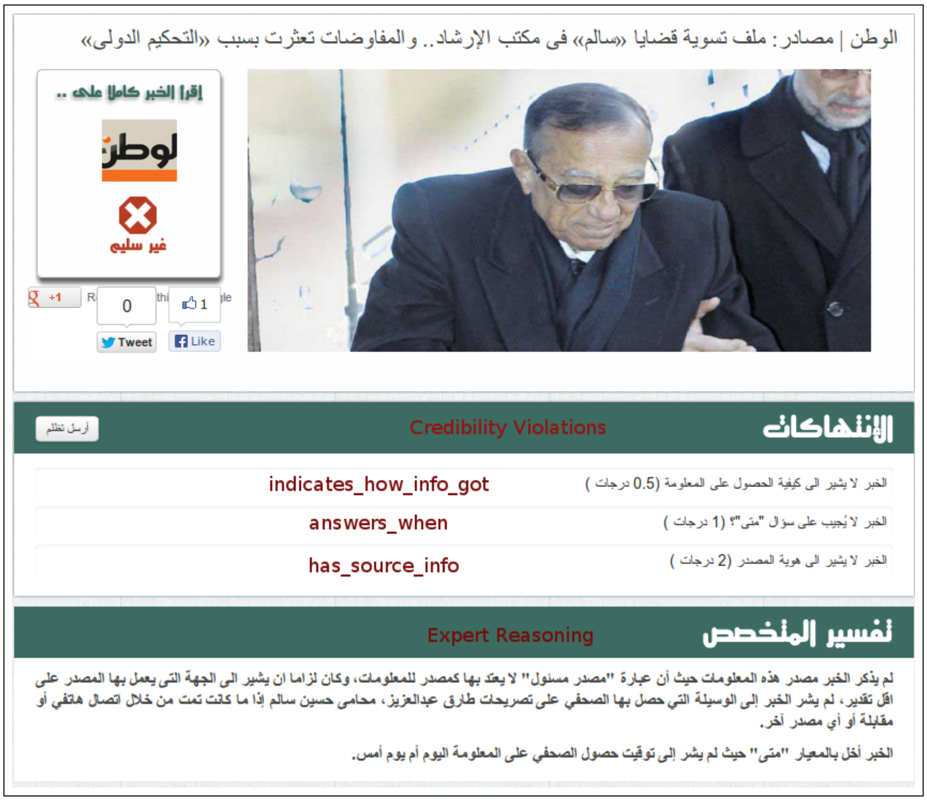
Table 3.2: Credibility evaluation criteria and points corresponding to each one

Alias	**Criterion**	**Points**
indicates_how_info_got	News article indicates how information was got	0.5
correct_info	News article doesn't contain incorrect or incomplete information	2
correct_news	News article doesn't report incorrect news	6
correct_photos	News article doesn't contain incorrect or manipulated photos	2
correct_order	Chronological order of information in the news article is correct	2
correct_video	Attached video is not conflicting with the article's text	2
correct_title	News article doesn't contain a misleading title	2
no_old_info	No old information posted as new	3
has_source	News article indicates the source's identity	2
correct_numbers	News article doesn't contain inaccurate numbers or statistics	2
unbiased	News article is unbiased	2
answers_how	News article answers 'how?'	0.5
answers_why	News article answers 'why?'	0.5
answers_where	News article answers 'where?'	1
answers_when	News article answers 'when?'	1
answers_who	News article answers 'who?'	4

- $score_{news\_site}$ is the credibility score assigned to *news_site*.
- n is the number of articles belonging to *news_site* that were judged by MCE Watch at the time of the score calculation.
- m is the number of credibility criteria checked by MCE Watch, from Table 3.2 $m = 16$.
- v_j is 1 if article i violates credibility criteria j. It is zero otherwise.
- p_j is the points corresponding to criteria j. The points are listed in the third column of Table 3.2

For the example shown in Figure 3.2, this article results in a total deduction of 3.5 points from *elwatan*'s pool of points. If we calculate $score_{elwatan}$ based on this article only, then $score_{elwatan} = \frac{500-(0.5+2+1)}{500} * 100\% = 99.3\%$.

The result of these score calculations is a ranking of the news sites by credibility. Figure 3.4 shows the most credible news sites for April 2013. *egynews* comes on top with 72.8% credibility followed by *almasryalyoum* and *masrawy* with 67.6% and 65%, respectively.



الوطن | مصادر: ملف تسوية قضايا «سالم» فى مكتب الإرشاد.. والمفاوضات تعثرت بسبب «التحكيم الدولى»

الإنتهاكات Credibility Violations

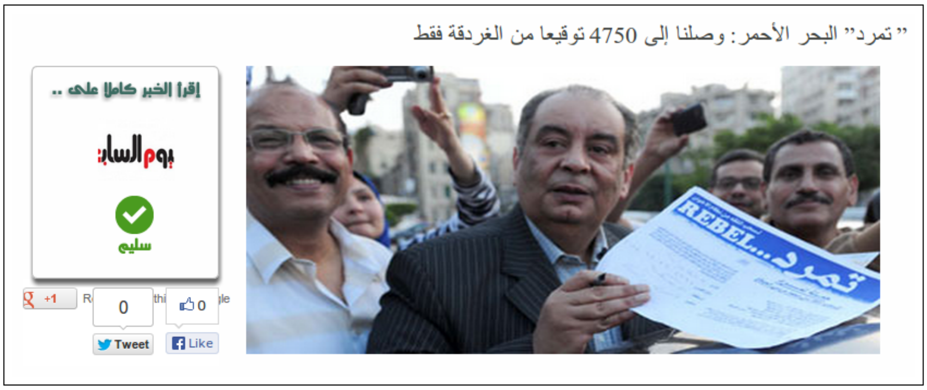
الخبر لا يشير الى كيفية الحصول على المعلومة (0.5 درجات)	indicates_how_info_got
الخبر لا يُجيب على سؤال "متى"؟ (1 درجات)	answers_when
الخبر لا يشير الى هوية المصدر (2 درجات)	has_source_info

تفسير المتخصص Expert Reasoning

لم يذكر الخبر مصدر هذه المعلومات حيث أن عبارة "مصدر مسئول" لا يعتد بها كمصدر للمعلومات، وكان لزاما ان يشير الى الجهة التى يعمل بها المصدر على اقل تقدير، لم يشر الخبر إلى الوسيلة التي حصل بها الصحفي على تصريحات طارق عبدالعزيز، محامى حسين سالم إذا ما كانت تمت من خلال اتصال هاتفي أو مقابلة أو أي مصدر آخر.

الخبر أخل بالمعيار "متى" حيث لم يشر إلى توقيت حصول الصحفي على المعلومة اليوم أم يوم أمس.

Figure 3.2: An example of a news article judged as non-credible by MCE Watch.



" تمرد" البحر الأحمر: وصلنا إلى 4750 توقيعا من الغردقة فقط

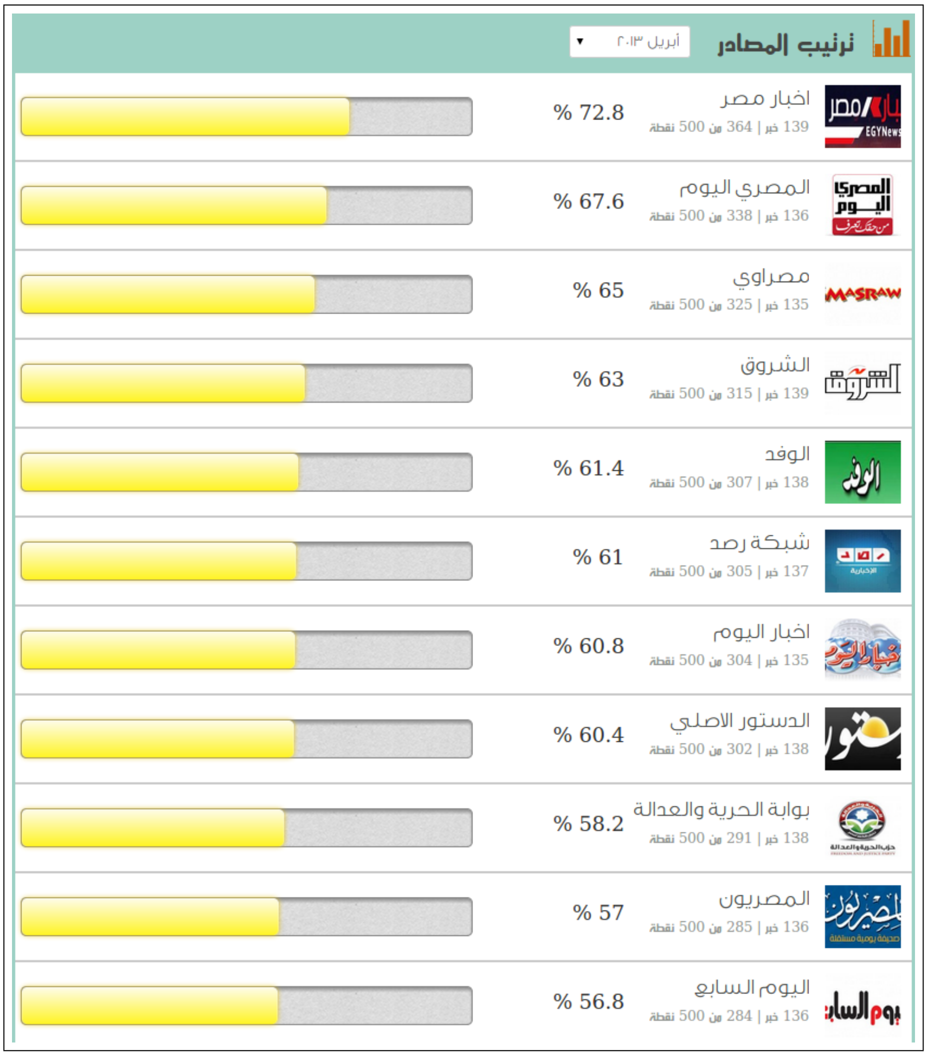
Figure 3.3: An example of a news article judged as credible by MCE Watch.

3.1.4.2 Pros and Cons

MCE Watch is like any system that has its advantages and disadvantages.

Advantages of MCE Watch

1. Allows a normal user to check the credibility of a news article before building any



ترتيب المصادر

أبريل ٢٠١٣ ▾

اخبار مصر	139 خبر | 364 من 500 نقطة	% 72.8
المصري اليوم	136 خبر | 338 من 500 نقطة	% 67.6
مصراوي	135 خبر | 325 من 500 نقطة	% 65
الشروق	139 خبر | 315 من 500 نقطة	% 63
الوفد	138 خبر | 307 من 500 نقطة	% 61.4
شبكة رصد	137 خبر | 305 من 500 نقطة	% 61
اخبار اليوم	135 خبر | 304 من 500 نقطة	% 60.8
الدستور الاصلي	138 خبر | 302 من 500 نقطة	% 60.4
بوابة الحرية والعدالة	138 خبر | 291 من 500 نقطة	% 58.2
المصريون	136 خبر | 285 من 500 نقطة	% 57
اليوم السابع	136 خبر | 284 من 500 نقطة	% 56.8

Figure 3.4: Most credible news sites for April 2013 according to MCE Watch.

decisions.

2. The reasoning provided by the expert allows a user to get insight into how news articles are assessed and how credible news articles should look like. This helps in building awareness among normal news consumers.

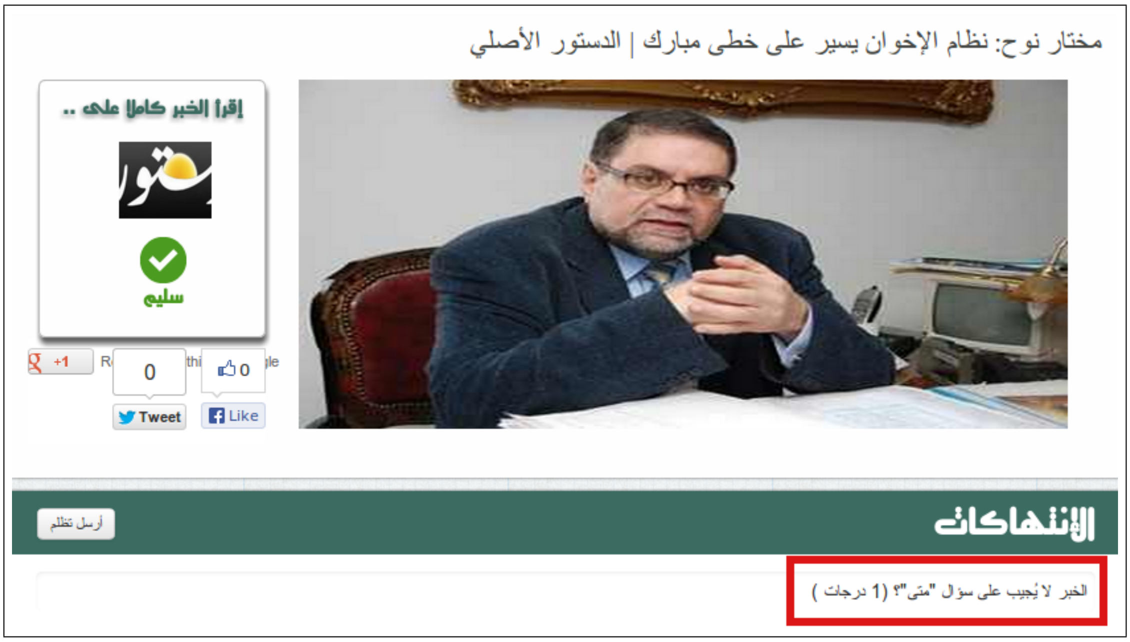
3. The reasoning provided by the expert allows news writers to take care in their future writings. This helps develop credibility awareness among news writers as well which in turn contributes towards developing more credible media.

4. Ranking of the news sites by credibility helps a user choose his source of information. Although the ranking changes every month but it gives a general indication of who is on top.

5. Ranking of the news sites by credibility attracts the attention of the chief editors of the news sites to promote credibility awareness among the writers since this impacts the reputation of the news site. This also contributes towards developing more credible media.

Disadvantages of MCE Watch
Given that MCE Watch is a manual service.

1. It is expensive to scale in terms of the number of requests served daily.
2. There is a delay accompanied with the response. This delay can be variable.
    - Requests done at late night are usually answered in the next morning.
    - More users requesting assessment at the same time implies more delay.
3. There can be a human error. An example of this is in Figure 3.5. The expert listed that the article violates *answers_when* although he did mark it as credible and indeed it doesn't violate this criteria since it explicitly stated:
وأضاف نوح خلال الندوة التى اقيمت فى نادى الرواد بمدينة العاشر من رمضان **مساءأمس** ...



مختار نوح: نظام الإخوان يسير على خطى مبارك | الدستور الأصلي

الخبر لا يُجيب على سؤال "متى"؟ (1 درجات)

Figure 3.5: An example of an article with human error.

3.1.4.3 Summary of MCE Watch

To sum this initiative up, we presented a detailed description of a manual service for credibility assessment of news articles. Users submit URLs of the articles to the service, a media expert analyzes the articles and replies with his judgment. Such a service has the advantages of helping users check the credibility of news when in doubt and directs them to credible news sites. It also helps news reporters pay attention at possible credibility pitfalls which would help in developing credible media. Since the service is manual it has the disadvantages that there is always a delay before getting a response, it is expensive to scale and exposed to human error.

From a computer science perspective, the news articles previously judged by MCE watch for credibility can be utilized to build a model to automate such credibility assessment process; this basically outlines our approach. Consequently, we are going to deal with text classification problems which we survey in the next section.

3.2 Text Classification

In this section we present a brief review of the work done on multiple problems involving text processing.

3.2.1 Text Categorization

Text Categorization (TC) is the task of assigning a category or more from a set of categories $C = \{c_1, c_2, ..., c_{|C|}\}$ to a document d from a set of documents D. Sebastiani [55] provided a formal definition of this task as follows:

> *"Text categorization is the task of assigning a Boolean value to each pair $\langle d_j, c_i \rangle \in D \times C$, where D is a domain of documents and $C = \{c_1, c_2, ..., c_{|C|}\}$ is a set of predefined categories. A value of T assigned to $\langle d_j, c_i \rangle$ indicates a decision to file d_j under c_i, while a value of F indicates a decision not to file d_j under c_i . More formally, the task is to approximate the unknown target function $\breve{\phi} : D \times C \rightarrow \{T, F\}$ (that describes how documents ought to be classified) by means of a function $\phi : D \times C \rightarrow \{T, F\}$ called the classifier (aka rule, or hypothesis, or model) such that $\breve{\phi}$ and ϕ coincide as much as possible"*

Machine learning algorithms have been considered for long to perform text categorization tasks. NBC classifiers have been considered for this task in multiple works [34, 35, 38, 40, 41]. In 1994, Lewis and Gale [41] used NBC classifier to assign keywords/tags to the titles of news articles collected from Associated Press (AP) [10]. They investigated the effect of multiple data sampling techniques on the effectiveness of the model and showed that a substantial decrease in the need of manually labeled data can be achieved using uncertainty sampling. Lewis [40] reviewed some of the variations of NBC used in information retrieval. Larkey and Croft [38] combined NBC with other classifiers for text categorization in the medical domain. They concluded that using multiple classifiers improves the classification performance. Kim et al. [35] proposed a Poisson NBC text classification model with a weight-enhancing method and per-document text normalization. Experimental results showed that their model is much better than the traditional NBC especially when there is a small number of training documents, however, it fails to outperform the state-of-the-art SVM classifiers.

SVM classifiers have been considered for text categorization in multiple works [28]. In 1998, Joachims [28] showed that SVMs achieve better classification performance than other classification techniques namely NBC, KNN, Rocchio and Decision Trees. He conducted his experiments on two different datasets; one contains news articles from Reuters [60] and another with disease information. Text was represented using BOW model with TF-IDF weighting, stemming, and stopwords removal. Experimental results showed that SVM consistently achieve good performance on categorization tasks. Lan et al. [37] studied the different term weighting schemes (e.g. binary, TF, IDF, TF-IDF, and TF.CHI) for text categorization using SVM. They proposed a new weighting scheme TF.Relevance Frequency (TF.RF) and showed that TF.RF gives significantly better performance over other weighting schemes using two widely-used datasets when compared based on the micros-averaged precision/recall break-even point.

For Arabic, Al-Shalabi et al. [6] performed news categorization of articles from Aljazeera [8], Al-Nahar, Al-Hayat, Al-Ahram, Al-Dostor and the Arabic version of the Food and Agriculture Organization (FAO) into politics, economy, sports, health-medicine, health-cancer, or agriculture-plants. They represented the articles as a BOW model with lightly stemmed words and TF-IDF weighting. Classification was done using KNN with cosine similarity as the distance function. Al-Shalabi and Obeidat [7] suggested using a combination of unigrams and bigrams for news categorization. They considered news from 4 categories (Computer, Economics, Education, and Engineering) collected from Aljazeera [8], Al-Nahar, Al-Hayat, and Al-Dostor. First they processed the news articles as follows:

- Removed non-Arabic letters, digits, single Arabic letters, punctuation, special symbols and diacritics.
- Removed stopwords.
- Replaced أ and إ by ا
- Replaced ة by ه
- Replaced ىء by ئ
- Removed بال, فال, وال, and كال

- Removed ال and لل except from الله and لله

Next BOW model with TF-IDF weighting was used twice; once with word unigrams and another time with word unigrams and bigrams. Document frequency was then used to eliminate features that occur in 3 or less documents. Finally KNN was then employed to classify test samples with the distance measure being the cosine similarity. Experimental results showed that the combination of unigrams and bigrams peformed better than unigrams alone when compared based on the average F-Measure. Gharib et al. [23] performed news categorization of articles from 3 Egyptian news sites (Al-Ahram, Al-Akhbar, and Al-Gomhoria) into 6 categories (Arts, Economics, Politics, Sports, Woman, or Information Technology). First, articles were tokenized to extract words. Next, stop words were removed and the remaining words were stemmed using a hybrid of light stemmer and statistical stemmer. After that, feature selection was employed; document frequency was first applied to remove rare terms followed by information gain to select the most informative features. Articles were then laid out using a BOW model with normalized TF-IDF weighting and finally the articles were classified using a number of different classifiers (KNN, Rocchio, NBC, and SVM). Experimental results showed that SVM outperforms the other classification algorithms in higher dimensional feature spaces. Khreisat [33] developed a solution to categorize articles from Egyptian news sites into one of four categories namely sports, economy, technology, and weather. She first normalized all text by:

- Removing punctuation marks, diacritics, non-letters, and stopwords
- Replacing أ, آ and إ by ا
- Replacing ىء by ئ

News articles were represented by their character trigram profile. Next, she used a nearest neighbor classifier with the distance measure being either Manhattan distance or Dice similarity measure. Experimental results showed that classification performance based on PR was better using Dice similarity measure. Alsaleem [9] compared the results of using SVM and NBC to classify news articles from Saudi Newspapers (SNP) into 8 categories culture, economics, general, information, technology, politics, social, and sports.

3.2.2 Spam Filtering

Email spam is one of the most important problems in computer science that has been given much attention by researchers during the last decade. Spam is considered as unwanted or junk email, it has multiple definitions among the shortest of them is "Unsolicited Email" [18,57,58]. Another definition is "Spam is the use of electronic messaging systems to send unsolicited bulk messages, especially advertising, indiscriminately." [70].

There are many solutions to the spam problem presented in the literature; Blanzieri and Bryl [12] provided a survey on the learning-based techniques. They also briefly outline other efforts such as efforts relying on enhancing or updating the existing email transmission protocols e.g. Simple Mail Transfer Protocol (SMTP) to be less vulnerable to spam by checking the ID of the sender [21, 52, 72] or placing obstacles in front of the email users that are small in case of sending few emails and massive in case of sending a large number of emails [36, 61].

Learning methods for spam received greater attention from researchers, these methods typically perform classification of email messages into spam or not spam (ham) and filter out the identified spam. We are more interested in these learning methods because many of them operate on the text of the email message. These methods involve using one or more machine learning algorithm like k-NN, Naive Bayes, SVM...etc. Before running any of these algorithms, data has to be extracted from an email message and formatted in such a way to be understood by the learning algorithms. Guzella and Caminhas [25] outlined in their survey that researchers use one or more of the following preprocessing steps:

1. tokenization, which extracts words from the message subject and body.
2. lemmatization, which reduces words to their root forms.
3. stop-word elimination, which remove some words that commonly appear in both spam and ham messages like "the", "and", "or", "for"...etc.
4. representation, which converts the words from the message to a format understood by the learning algorithm. They also outlined that the commonly used representation is the BOW model with terms being represented as:

    - words
    - character n-grams
    - word n-grams

    These representations are then encoded either in a binary fashion indicating occurrences or with a weighting scheme e.g. term frequency (TF) or TF-IDF.

NBC was one of the first learning algorithms to be used for spam filtering. In 1998, Sahami et al. [50] and Pantel and Lin [47] investigated the use of Naive Bayes classifier for spam filtering. Pantel and Lin used stemmed words or character n-grams (n = 3) in their BOW model and they employed frequency based techniques to remove words that are irrelevant to the classification problem. Sahami et al. investigated adding domain specific features (e.g. presence of phrases like “FREE!” or “only $” in the text of the email, the type of the domain used in sending the message (.edu, .org, .com...etc.), and the percentage of non-alphanumeric characters) to the BOW model on its classification performance. They showed by experimental results that domain specific features have high impact on increasing the classification performance. Çıltık and Güngör [15] used Naive Bayes classifier based on word n-grams for email spam filtering on both English and Turkish. They only considered the first few words of the email and ignored the rest. They created two models (i) Class General Perception (CGP) model that classifies emails into either spam or ham and (ii) E-mail Specific Perception (ESP) model that uses similarity between the test email and every article in the training set to determine if the test email is spam or not (similar to KNN). They combined both models to achieve an increase in the performance of CGP without much time overhead.

A number of researchers used SVM for spam filtering [14, 19, 27, 54]. In 1999, Drucker et al. [19] investigated using SVM for spam filtering; they examined the use of BOW model with binary, TF and TF-IDF weighting schemes. They compared their results with other learning algorithms and concluded that SVM with binary features and boosting with decision trees with TF features are the best candidate models. They also concluded that SVM performance doesn’t degrade with too many features. Islam et al. [27] proposed using SVM with linear kernel and BOW model augmented with domain dependent features for spam filtering. Sculley and Wachman [54] used SVM for online spam filtering. Emails were represented as BOW model with either words or character n-grams weighted in a binary fashion. They concluded that online SVM give state-of-the-art classification performance on large benchmark datasets from Text REtrieval Conference (TREC) 2005 and 2006. Chhabra et al. [14] investigated the effect of multiple kernels on the performance of SVM for spam filtering. Again they represented their data using a BOW model and concluded that the classification performance decreases as the degree of the polynomial kernel increases. Blanzieri and Bryl [11] suggested a combination of SVM and KNN (SVM-NN) for spam filtering. Once more they represented their data with a BOW model with features encoded in a binary fashion. After that they selected a number of the most frequent features/words in the training data to take into account and ignore the rest. To classify an unknown test email, KNN is first employed to determine the closest K samples to the test sample, next an SVM model is trained on the the closet K samples and classification is performed according to this model. Experimental results showed that SVM-NN outperformed k-NN and SVM significantly for low dimensional feature spaces but with less clear advantage over SVM in higher dimensional spaces due to SVM-NN’s higher sensitivity to irrelevant features.

Other machine learning algorithms were also used for spam filtering for example Clark et al. [16] used Neural Network (NN) for email categorization and spam filtering. They showed that NN outperforms NBC, KNN and several other learning algorithms.

Chapter 4: Methodology

In this chapter we describe our approach to automate credibility assessment of Arabic news.

The description of MCE Watch given in Chapter 3 indicates that media researchers have already done the work necessary to evaluate the credibility of Arabic news. However, this work lacks automation which is the job to be done by computer science researchers.

Our approach relies heavily on the work done by MCE Watch. In other words, our work is a step towards automating their work. Figure 4.1 shows a high level block diagram of our system architecture. Each of the upcoming subsections presents details for a block in this diagram.

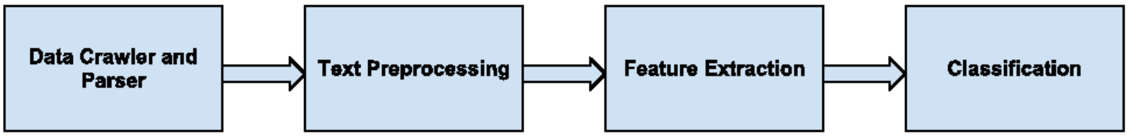


Figure 4.1: High level system architecture.

The system starts by downloading and parsing data from MCE Watch – the next subsection presents details of this stage. The result of this stage is a labeled data set of news articles; each article is labeled as either credible or not and if not credible a list of the violations of the criteria in table Table 3.2 is compiled.

From the crawled data, the distribution of non-credible articles among different credibility criteria is presented in table Table 4.1. As can be seen, *indicates_how_info_got*, *has_source*, and *answers_when* are the most frequently violated criteria. Together, *has_source* and *answers_when* are violated in 80.64% of the non-credible articles. If we added *indicates_how_info_got* to them we get a coverage of 91.5%. We choose to automate *has_source* and *answers_when* only leaving out *indicates_how_info_got* because

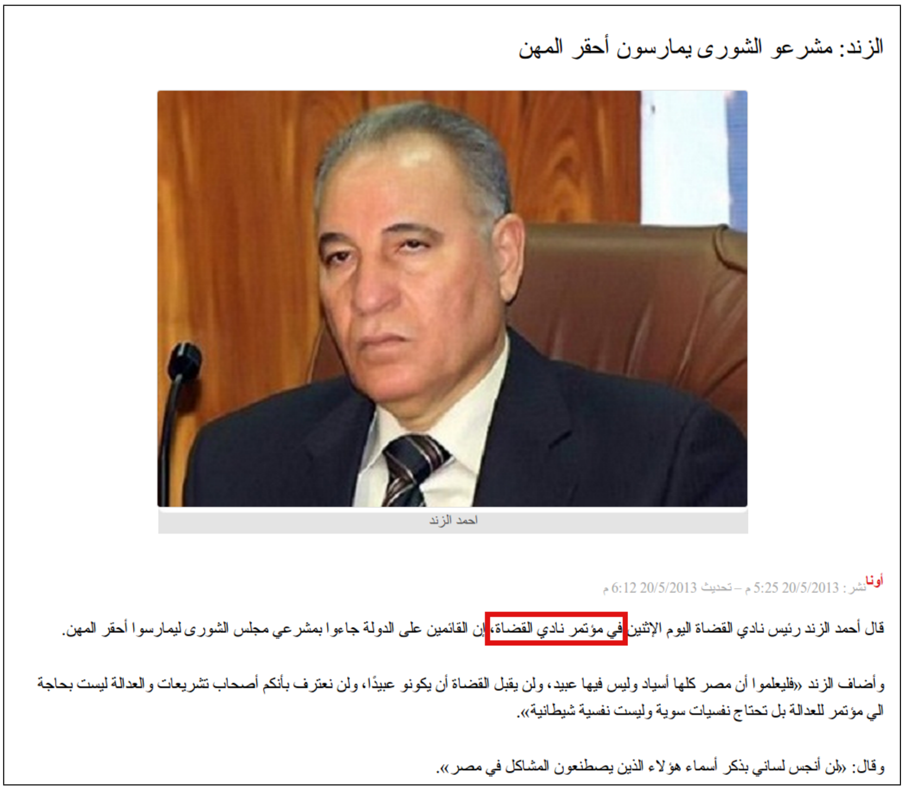
1. *indicates_how_info_got* is assigned low weight of 0.5 points vs 1 and 2 for *has_source* and *answers_when*, respectively.
2. *indicates_how_info_got* and *has_source* are highly correlated, since both are tightly coupled to the source, the later indicates if the source was explicitly mentioned in the article while the former indicates how did the source provide the writer of the article by information. In terms of numbers, 90.28% of the articles that violate *has_source* also violate *indicates_how_info_got*.
3. Our manual analysis of the collected data show that *indicates_how_info_got* has very high degree of noise. We present in Figure 4.2 an example of an article that doesn't

[2] Percentage doesn't sum to 100% because an article can violate more than one criteria at the same time.

Table 4.1: Distribution of non-credible articles among different credibility criteria

Criterion	**% of Violations**[2]
indicates_how_info_got	**49.67**
correct_info	2.22
correct_news	0.45
correct_photos	16.73
correct_order	0.01
correct_video	0.42
correct_title	3.51
no_old_info	0.04
has_source	**24.83**
correct_numbers	0.00
unbiased	2.28
answers_how	1.15
answers_why	2.32
answers_where	0.79
answers_when	**69.69**
answers_who	3.49

violate *indicates_how_info_got* and we highlight "في مؤتمر نادي القضاة" which indicates how the writer got the information. The noise in the labels for this criteria



الزند: مشرعو الشورى يمارسون أحقر المهن

احمد الزند

أونانشر : 20/5/2013 5:25 م – تحديث 20/5/2013 6:12 م

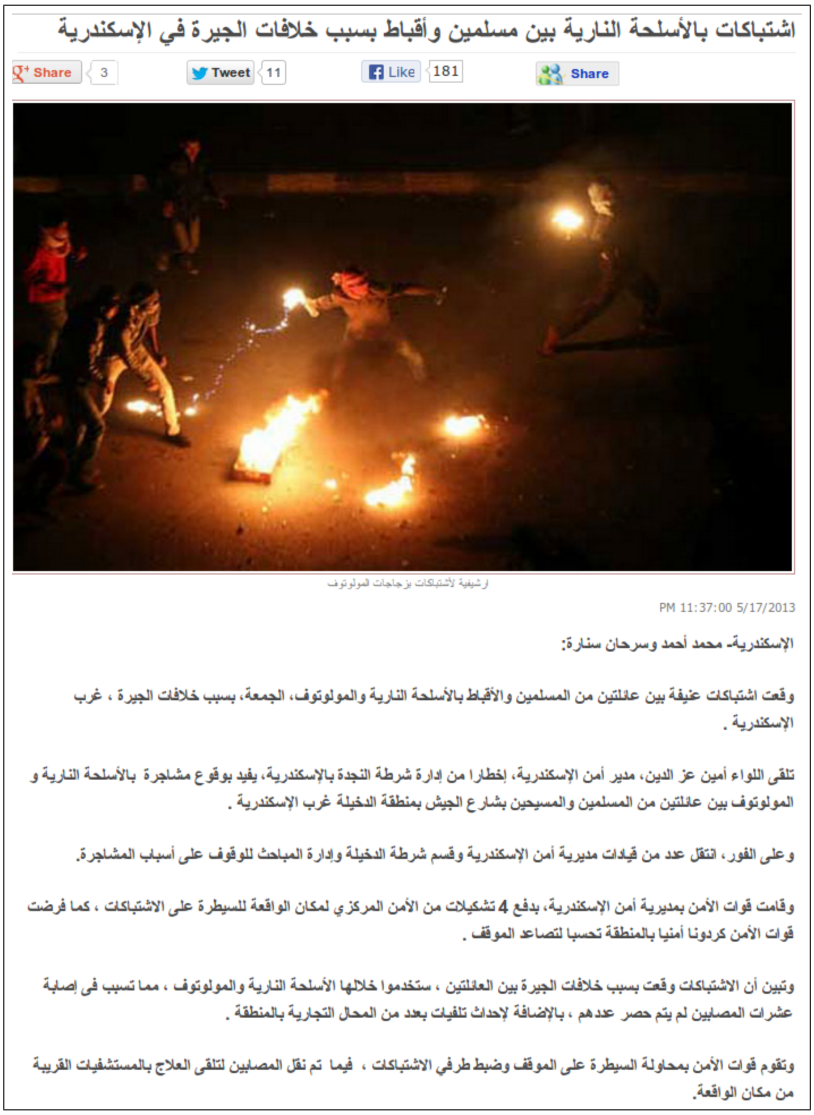
قال أحمد الزند رئيس نادي القضاة اليوم الإثنين في مؤتمر نادي القضاة، إن القائمين على الدولة جاءوا بمشرعي مجلس الشورى ليمارسوا أحقر المهن.

وأضاف الزند «فليعلموا أن مصر كلها أسياد وليس فيها عبيد، ولن يقبل القضاة أن يكونو عبيدًا، ولن نعترف بأنكم أصحاب تشريعات والعدالة ليست بحاجة الي مؤتمر للعدالة بل تحتاج نفسيات سوية وليست نفسية شيطانية».

وقال: «لن أنجس لساني بذكر أسماء هؤلاء الذين يصطنعون المشاكل في مصر».

Figure 4.2: Example of an article that doesn't violate *indicates_how_info_got* and labeled correctly.

can be seen by looking at the examples given in Figure 4.3 and Figure 4.4. Figure 4.3 shows an example article that was marked as not violating *indicates_how_info_got*. Another very similar article that was marked as violating *indicates_how_info_got* is shown in Figure 4.4. Both articles doesn't indicate how info was got but they were labeled differently.



اشتباكات بالأسلحة النارية بين مسلمين وأقباط بسبب خلافات الجيرة في الإسكندرية

ارشيفية لأشتباكات بزجاجات المولوتوف

PM 11:37:00 5/17/2013

الإسكندرية- محمد أحمد وسرحان سنارة:

وقعت اشتباكات عنيفة بين عائلتين من المسلمين والأقباط بالأسلحة النارية والمولوتوف، الجمعة، بسبب خلافات الجيرة ، غرب الإسكندرية .

تلقى اللواء أمين عز الدين، مدير أمن الإسكندرية، إخطارا من إدارة شرطة النجدة بالإسكندرية، يفيد بوقوع مشاجرة بالأسلحة النارية و المولوتوف بين عائلتين من المسلمين والمسيحين بشارع الجيش بمنطقة الدخيلة غرب الإسكندرية .

وعلى الفور، انتقل عدد من قيادات مديرية أمن الإسكندرية وقسم شرطة الدخيلة وإدارة المباحث للوقوف على أسباب المشاجرة.

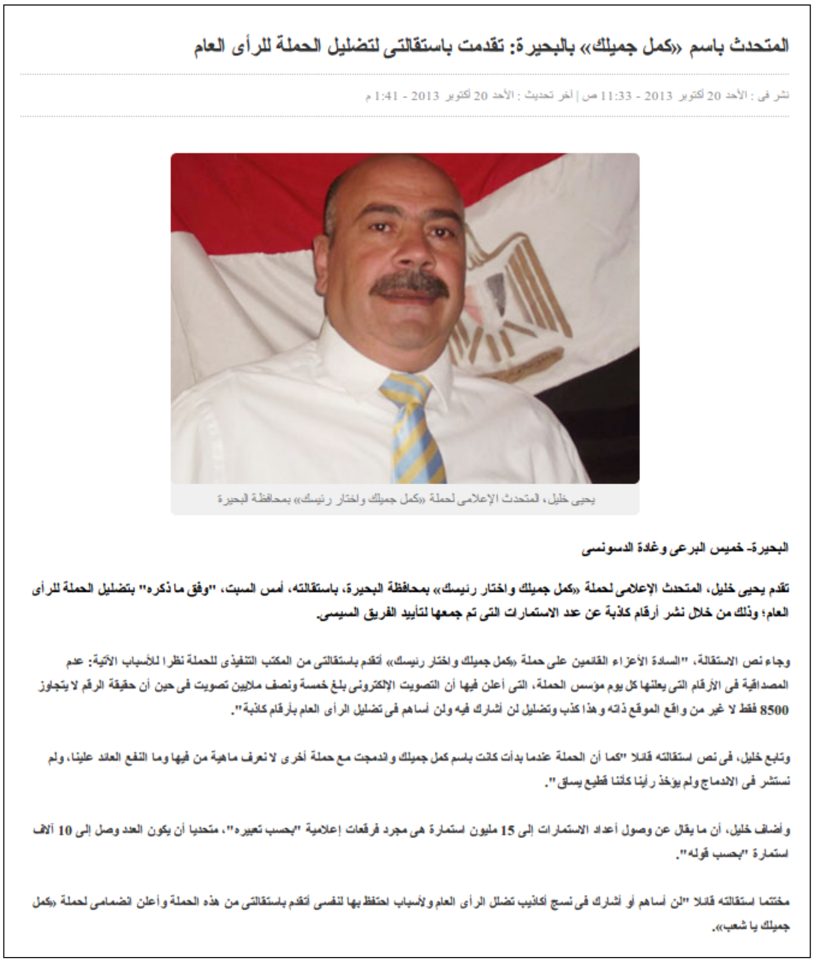
وقامت قوات الأمن بمديرية أمن الإسكندرية، بدفع 4 تشكيلات من الأمن المركزي لمكان الواقعة للسيطرة على الاشتباكات ، كما فرضت قوات الأمن كردونا أمنيا بالمنطقة تحسبا لتصاعد الموقف .

وتبين أن الاشتباكات وقعت بسبب خلافات الجيرة بين العائلتين ، ستخدموا خلالها الأسلحة النارية والمولوتوف ، مما تسبب فى إصابة عشرات المصابين لم يتم حصر عددهم ، بالإضافة لإحداث تلفيات بعدد من المحال التجارية بالمنطقة .

وتقوم قوات الأمن بمحاولة السيطرة على الموقف وضبط طرفي الاشتباكات ، فيما تم نقل المصابين لتلقى العلاج بالمستشفيات القريبة من مكان الواقعة.

Figure 4.3: Example of an article that violates *indicates_how_info_got* and labeled incorrectly.

For each of the two criteria we build a binary text classifier. Typically, we perform preprocessing on the text, feature extraction, training and classification. In sections Section 4.2-Section 4.4 we discuss the details of each of these steps. We discuss details of preprocessing in the section following that for feature extraction since we base some of our preprocessing on the features.



المتحدث باسم «كمل جميلك» بالبحيرة: تقدمت باستقالتى لتضليل الحملة للرأى العام

نشر فى : الأحد 20 أكتوبر 2013 - 11:33 ص | آخر تحديث : الأحد 20 أكتوبر 2013 - 1:41 م

يحيى خليل، المتحدث الإعلامى لحملة «كمل جميلك واختار رئيسك» بمحافظة البحيرة

البحيرة- خميس البرعى وغادة الدسوقى

تقدم يحيى خليل، المتحدث الإعلامى لحملة «كمل جميلك واختار رئيسك» بمحافظة البحيرة، باستقالته، أمس السبت، "وفق ما ذكره" بتضليل الحملة للرأى العام؛ وذلك من خلال نشر أرقام كاذبة عن عدد الاستمارات التى تم جمعها لتأييد الفريق السيسى.

وجاء نص الاستقالة، "السادة الأعزاء القائمين على حملة «كمل جميلك واختار رئيسك» أتقدم باستقالتى من المكتب التنفيذى للحملة نظرا للأسباب الآتية: عدم المصداقية فى الأرقام التى يعلنها كل يوم مؤسس الحملة، التى أعلن فيها أن التصويت الإلكترونى بلغ خمسة ونصف ملايين تصويت فى حين أن حقيقة الرقم لا يتجاوز 8500 فقط لا غير من واقع الموقع ذاته وهذا كذب وتضليل لن أشارك فيه ولن أساهم فى تضليل الرأى العام بأرقام كاذبة".

وتابع خليل، فى نص استقالته قائلا "كما أن الحملة عندما بدأت كانت باسم كمل جميلك واندمجت مع حملة أخرى لا نعرف ماهية من فيها وما النفع العائد علينا، ولم نستشر فى الاندماج ولم يؤخذ رأينا كأننا قطيع يساق".

وأضاف خليل، أن ما يقال عن وصول أعداد الاستمارات إلى 15 مليون استمارة هى مجرد فرقعات إعلامية "بحسب تعبيره"، متحديا أن يكون العدد وصل إلى 10 آلاف استمارة "بحسب قوله".

مختتما استقالته قائلا "لن أساهم أو أشارك فى نسج أكاذيب تضلل الرأى العام ولأسباب احتفظ بها لنفسى أتقدم باستقالتى من هذه الحملة وأعلن انضمامى لحملة «كمل جميلك يا شعب».

Figure 4.4: Example of an article that violates *indicates_how_info_got* and labeled correctly.

4.1 Data Crawler and Parser

Figure 4.5 presents a detailed architecture of this stage. First, we crawl MCE Watch for news articles judged between December 2012 and April 2013. As shown in Figure 4.5, we use HTTP GET requests to download the HTML pages of MCE Watch. Next, we parse the HTML from MCE Watch to populate a database with URLs of news articles, the name of the news site for each article, judgment of MCE Watch; credible or not and the list of violations in case of a non-credible article. After that we download the news articles and parse each one extracting the title and text. All parsers use XPath expressions to extract the required text from the downloaded HTML pages. The structure of HTML pages is different across multiple news sites and accordingly a different XPath expression has to be designed for each.

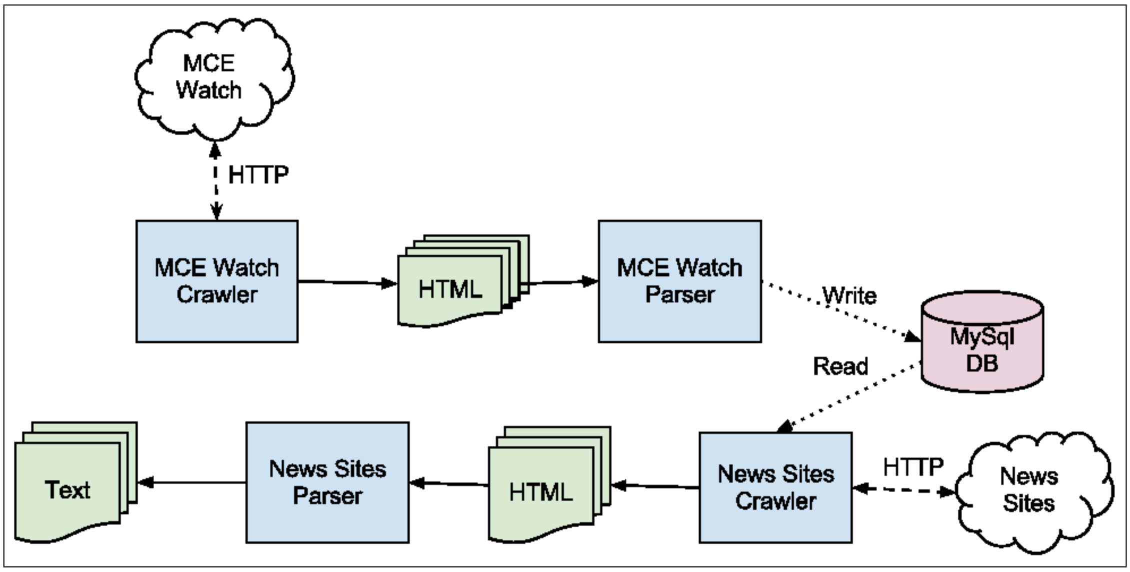


Figure 4.5: Architecture of Data Crawler and Parser.

4.2 Feature Extraction

In this subsection we present the features for each of the chosen criteria. We start by features for *answers_when* followed by features for *has_source_info*.

4.2.1 Does the Article Answer 'When'?

We model the answer to this question as a binary text classification problem. We classify news articles into one of two classes; **HasWhen** or **NoWhen**. Articles of **HasWhen** class doesn't violate *answers_when* criteria while articles of **NoWhen** do violate it. We use a combination of features to represent each article:

- *N-GRAMS(NG)*: we run experiments with combinations of word n-grams with $n \leq 3$. We use term frequencies without inverse document frequency weighting (IDF).
- *TIME(T)*: this is a binary feature indicating whether any word from our manually created lexicon of time-bearing word ngrams (TIME-lexicon) exists in the article. The lexicon consists of all weekdays, today, yesterday, short ago and a few time ago. Table 4.2 shows the lexicon.
- *TIME-DISTANCE(TD)*: this is a real-valued feature. It is calculated as the distance of the time-bearing word found by *TIME* relative to the beginning of the article. It is transformed by an exponential function to the range[1, e^1], Equation 4.1. Our analysis indicates that time-bearing words that refer to the time of the reported event

Table 4.2: Lexicon of time-bearing word ngrams (*TIME*-lexicon).

Arabic Word	**English Translation**
اليوم	Today
امس	Yesterday
السبت	Saturday
منذ قليل	Since shortly
الاربعاء	Wednesday
الاثنين	Monday
الثلاثاء	Tuesday
الخميس	Thursday
الجمعة	Friday
الاحد	Sunday
قبل قليل	Shortly before

tend to occur early on in the article.

$$TD = \begin{cases} e^{(1-\frac{T_i+1}{|D|})} & \text{if } T = 1 \\ 1 & \text{if } T = 0 \end{cases} \tag{4.1}$$

where T_i is the index of time-bearing word in a vector of the article words and $|D|$ is the number of words in the article.

- *TIME-CONTEXT(TC)*: this is a real-valued feature. It is calculated as the minimum absolute distance between the time-bearing word found by *TIME* and any of the words in a manually created lexicon of contexts (*LC*), Equation 4.2. Contexts include conferences, tweets, Facebook updates, meetings, TV shows, telephone calls, and statements. We only search for context words in a frame of $\pm\epsilon$.

$$TC = \begin{cases} \min\limits_{T_i-\epsilon<j<T_i+\epsilon} \frac{|T_i-j|}{\epsilon} & \text{if } T = 1 \text{ and } j \text{ in } LC \\ 1 & \text{otherwise} \end{cases} \tag{4.2}$$

where T_i is the index of time-bearing word in a vector of the article words. In our experiments, we choose $\epsilon = 18$.

4.2.2 Does the Article Has ‘Source Info’?

We again model the answer to this question as a binary text classification problem. We classify news articles into one of two classes; **HasSourceInfo** or **NoSourceInfo**. Articles of **HasSourceInfo** class doesn’t violate *has_source* criteria while articles of **NoSourceInfo** violate it. We use a combination of features to represent each article:

- *N-GRAMS*: we run experiments with combinations of word n-grams with $n \leq 3$. We use term frequencies without IDF weighting.
- *SOURCE*: this is a binary feature indicating whether the news article is referring to one of the popular news agencies as a source. Table 4.3 lists the agencies we consider here.

Table 4.3: A list of agencies that news articles refer to as a source of their information.

English Name	**Arabic Name**
Middle East News Agency	ا ش ا
Reuters	رويترز
Fox News	فوكس نيوز
Associated Press	اسوشيتدبرس
Anadolu Agency	الاناضول

- *VERB*: this is a binary feature indicating whether a verb from a manually created lexicon of verbs exists. All verbs indicate that someone said something. Table 4.4 lists these verbs.

Table 4.4: Lexicon of verbs which indicate that someone said something (*VERB*-lexicon).

Arabic Verb	**English Translation**
قال/قالت	He/She said
اوضح /اوضحت	He/She explained
اشار/اشارت	He/She pointed
انتقد	He criticized
اضاف	He added
تابعت	She continued
اكد	He confirmed
ناقشت	She discussed
نفى	He denied

4.3 Preprocessing

For both tasks, we perform basic normalization [17] of Arabic text. This normalization involves transformation of one form of a letter to another form mostly because both forms are erroneously used interchangeably:

- We transform all forms of *alef*: “أ, إ, آ” to “ا”.
- We transform all forms of *alef layyena*: “ى to “ي”
- We remove all Arabic diacritics. We also remove all punctuation and digits.

We tokenize each news article into a vector of words and discard tokens of length < 3. In case of *has_source*, we add two processing steps on the tokens:

- If token starts with “ال” – meaning “the”, we strip it. Like, “الأحزاب” – meaning “the parties” → “أحزاب” – meaning “parties”.
- If token starts with “و” – meaning “and”, it is commonly written as the first letter of the following word as in “وقال” – meaning “and he said”. We remove it giving “ قال” – meaning “he said”.

We don’t do the above for *“answers_when”* because all Arabic weekdays are commonly preceded by “*The*” and when stripped can mean a different thing leading to ambiguity. For example, “الأحد” – meaning “Sunday”, when stripped gives “أحد” – meaning “Someone”. “الإثنين” – meaning “Monday” gives “إثنين” – meaning the number“Two”. Similarly, for “ الخميس” and “الجمعة” – meaning “Thursday” and “Friday” respectively, when stripped gives “خميس” and “جمعة” that correspond to common Egyptian male names.

4.4 Classification

We use SVM with linear kernel from Python’s Scikit-learn package [48]. Literature suggests that it does well on text classification tasks. We set c to 1. All feature vectors are L2 normalized.

Chapter 5: Results and Evaluation

In this chapter we describe the details of our experiments and discuss their results. We conducted different categories of experiments; the first one presents classification results for *answers_when* criteria. It presents results for the different combinations of features presented in Section 4.2.1. The second one presents classification results for *has_source* criteria. It presents results for the different combinations of features presented in Section 4.2.2. The third and last one compares the credibility scores we assign to different news sites to the ones assigned by MCE Watch.

In the next subsection we present statistics and details of our data set. After that we go into the details of our experiments.

5.1 Data Description

A total of 9358 articles were crawled; 2606 articles judged as credible and 6752 articles judged as not credible by violating at least one of the credibility criteria listed in Table 3.2. The dataset is obviously unbalanced with 72.15% of the articles being not credible and 27.84% being credible. The main reason behind this is the way the service operates; users submit articles for credibility assessment when in doubt. However, our assumption is that this imbalance is not a good representative of the general case, we try to overcome this in our experiments.

The articles belong to different news categories; *Politics*, *Accidents*, *Economics*, *Arts*, and *Sports*. This diversity in news categories poses a challenge to our analysis but in the mean time it tests the ability of our model to generalize. One more challenge introduced by the data set is that some articles are written in Modern Standard Arabic(MSA) and others are in the Egyptian dialect.

Since labeled Data was crawled from MCE Watch, the labeling is done by media experts but this didn't involve annotating every article by multiple annotators and filtering conflicting annotations. Given that multiple experts provide the service it is common to find very similar cases where reasoning differs. It is also common to find cases that were judged and reasoned based on well known incidents at the time of judgment like public speeches, TV shows ... etc. without having a reference to the incident in the news article.

Length of articles in our news corpus is not uniform, shortest article has 158 characters (the article has no body just a title) and longest one has 45423 characters with average length of 1980.2 characters.

We divide our data into 80% training set and 20% test set. During development (*DEV*), we perform 5-fold cross validation where we train with 4 folds and test on the 5th, we report the average accuracy of all 5 folds. During testing (*TEST*), we use the best performing settings on DEV to train with all the training set and report accuracy on the test set.

For all experiments we compare our results against the baseline which is the majority class.

5.2 "When?" Experimental Results

For this problem, we have a total of 9358 articles divided among the two classes. For **HasWhen** class, we have a total of 4652 articles; 2606 credible articles and 2046 non-credible articles but doesn't violate *answers_when* criteria – we augmented the credible articles by this set of non-credible articles to overcome class imbalance (35.64% → 49.71%). For **NoWhen** class, we have a total of 4706 articles that violate *answers_when* criteria.

5.2.1 Experiment 1: Comparing Multiple Classifiers

In this experiment we use all combinations of n-grams for $n \leq 3$ to compare the efficiency of [26] KNN, NBC, and SVM classifiers and determine the best performing one. From Figure 5.1 (Table 5.1 shows the exact figures), all classifiers perform their best when unigrams are included in the model. In such case, NBC and KNN performance is very close for appropriately large values of k since the classification decision is based on a larger number of the training examples. SVM performs best with large difference (9.5% - 11.2%) on all combinations including unigrams. SVM's good performance conforms with the results reported in literature on other text classification problems.

Table 5.1: Exact figures for comparison between KNN (k = 25, k = 101, k = 1001), NBC, and SVM classifiers using all combinations of n-grams for $n \leq 3$

	25-NN	**101-NN**	**1001-NN**	**NBC**	**SVM**
n = 1	59.18	61.82	68.62	67.39	79.16
n = 1, 2	59.39	61.78	68.07	67.97	78.99
n = 1, 2, 3	59.46	62.22	67.67	67.74	78.98
n = 2	59.06	61.23	63.44	66.89	71.71
n = 2, 3	59.68	61.01	63.84	66.02	71.3
n = 3	56.72	58.91	56.76	61.82	61.45

From now on, we will carry out our experiments using SVM.

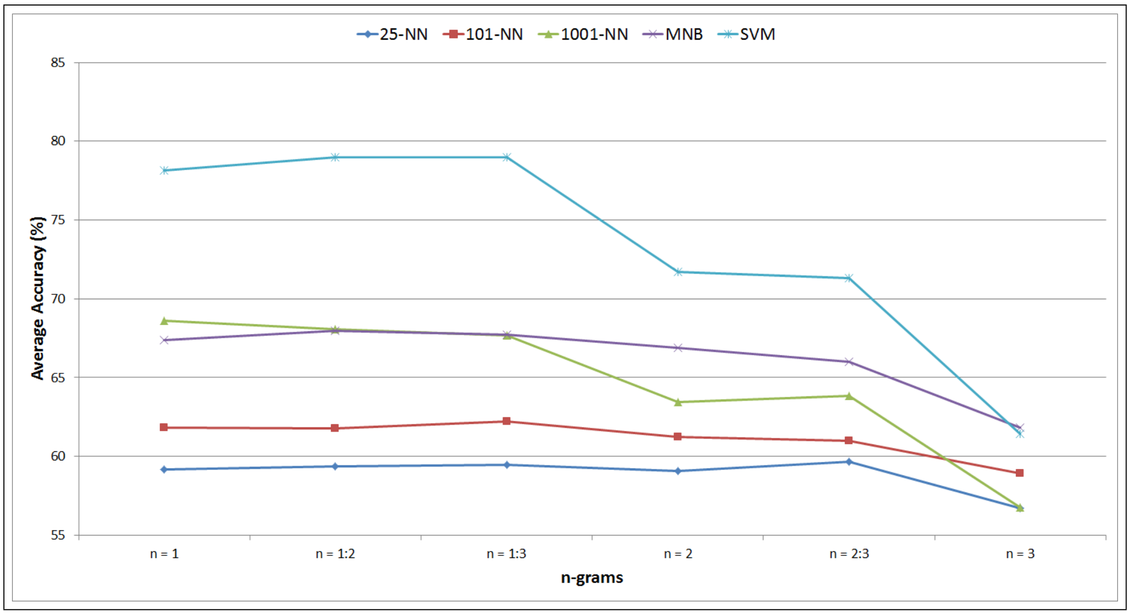


Figure 5.1: Comparison between KNN (k = 25, k = 101, k = 1001), NBC, and SVM classifiers using all combinations of n-grams for $n \leq 3$

5.2.2 Experiment 2: N-GRAMS Selection

This experiment determines the best setting of N-GRAMS on DEV. First we try all combinations of N-GRAMS for $n \leq 5$; we do this once using TF-IDF weighting and another time using TF without IDF weighting. Figure 5.2 shows a bar chart comparing both settings. As can be seen, most combinations of N-GRAMS that don't use IDF weighting give higher accuracies than those with IDF weighting.

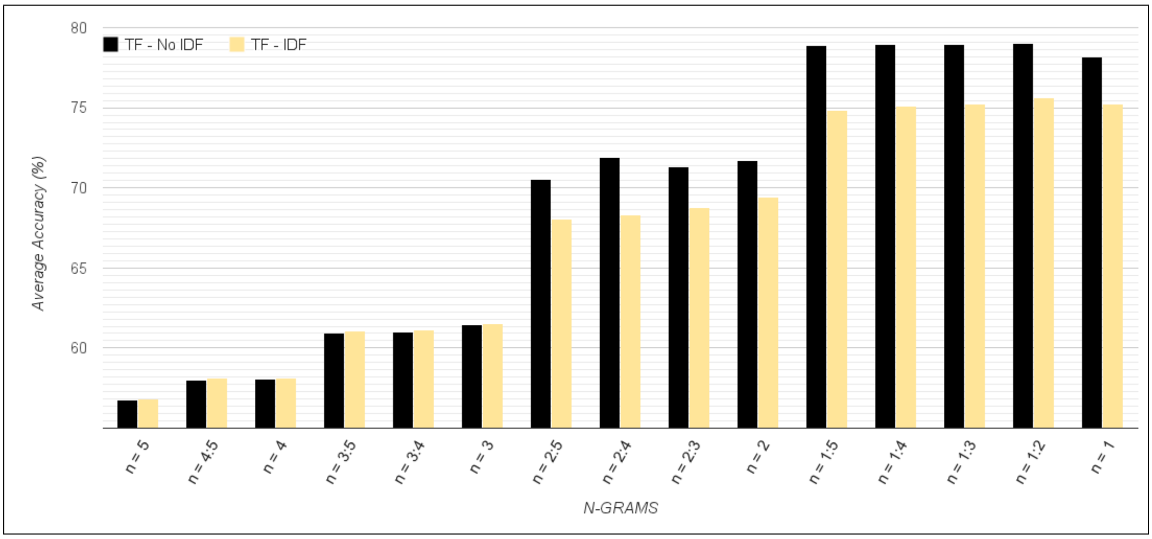


Figure 5.2: Comparing results for N-GRAMS when $n \leq 5$ using TF-IDF weighting versus TF only.

Table 5.2 shows the actual accuracy figures for N-GRAMS using TF only on both DEV and TEST. Unigrams + bigrams together give the highest accuracy of 78.99% on DEV, which is 28.7% higher than the baseline. The same setting also gives the highest accuracy of 77.12% on TEST which is 26.83% higher than the baseline. The accuracy on DEV is only 0.01% higher than that of unigrams + bigrams + trigrams and that of N-GRAMS with n = 1, 2, 3, and 4, however, we choose unigrams + bigrams as our best N-GRAMS setting for its smaller dimensionality regardless of the insignificant difference in accuracy.

Table 5.2: *answers_when* accuracy(%) on both DEV and TEST using all combinations of *N-GRAMS* $n \leq 5$.

	DEV	**TEST**	**Baseline**
n = 1	78.16	76.27	50.29
n = 1, 2	**78.99**	**77.12**	50.29
n = 1, 2, 3	78.98	76.91	50.29
n = 1, 2, 3, 4	78.98	76.43	50.29
n = 1, 2, 3, 4, 5	78.89	75.37	50.29
n = 2	71.71	68.00	50.29
n = 2, 3	71.29	66.67	50.29
n = 2, 3, 4	71.92	66.88	50.29
n = 2, 3, 4, 5	70.54	66.99	50.29
n = 3	61.44	57.60	50.29
n = 3, 4	60.98	56.69	50.29
n = 3, 4, 5	60.93	56.91	50.29
n = 4	58.08	54.19	50.29
n = 4, 5	58.00	54.13	50.29
n = 5	56.74	53.07	50.29

5.2.3 Experiment 3: Effect of domain dependent feature (TIME)

Next, we take a look at the feature *TIME*. This feature uses a manually crafted lexicon of time-bearing word ngrams. The lexicon was created using Sequential Forward Selection (SFS) algorithm. Table 5.3 shows the execution of this algorithm on our lexicon. At each iteration we choose to add a new word from the candidate words to our lexicon only if it increases the average accuracy on DEV. The algorithm stops running when none of the remaining words does a performance increase. The final set of words are those in Table 4.2.

Table 5.3: Using SFS to develop the *TIME*-lexicon.

الماضي	القادم	غدا	غد	قبل قليل	منذ قليل	امس	اليوم	الخميس	الاربعاء	الثلاثاء	الاثنين	الاحد	السبت	الجمعة
51.12	46.69	51.58	50.82	50.53	52.29	55.25	**73.9**	53.36	55.24	54.66	55.44	55.14	55.5	56.64
72.71	73.23	74.36	73.91	74.06	75.48	**76**	-	75.13	75.38	75.24	75.36	75.16	75.9	75.17
74.73	75.1	76.4	75.96	76.15	77.51	-	-	77.08	77.31	77.16	77.21	77.11	**77.63**	77.07
76.21	76.69	77.84	77.5	77.76	**79.13**	-	-	78.63	78.9	78.79	78.77	78.52	-	78.52
77.66	78.15	79.33	78.94	79.26	-	-	-	80.13	**80.37**	80.26	80.25	80.01	-	80.01
78.78	79.38	80.5	80.13	80.5	-	-	-	81.24	-	81.39	**81.48**	81.21	-	81.21
79.73	80.48	81.59	81.22	81.61	-	-	-	82.35	-	**82.43**	-	82.16	-	82.25
80.76	81.47	82.49	82.11	82.56	-	-	-	**83.3**	-	-	-	83.1	-	83.18
81.61	82.36	83.31	82.95	83.43	-	-	-	-	-	-	-	83.92	-	**83.98**
82.35	83.04	83.98	83.62	84.11	-	-	-	-	-	-	-	**84.55**	-	-
82.96	83.56	84.55	84.19	**84.69**	-	-	-	-	-	-	-	-	-	-
83.1	83.7	84.67	84.33	-	-	-	-	-	-	-	-	-	-	-

From Table 5.3, the best accuracy we get on DEV using the *TIME*-lexicon is 84.69%. However, avoiding some common pitfalls can lead to slight accuracy increase. Precisely, the following bigrams contain the word "اليوم" although they don't refer to a time:

- اخبار اليوم
- اليوم السابع
- المصري اليوم
- الحياة اليوم

The first three of these bigrams refer to names of news sites and the last one refers to a popular Egyptian talk show. Table 5.4 shows the increase in the average accuracy achieved by ignoring each of these bigrams. The histogram of the resulting feature *TIME* is shown in Figure 5.3. As can be seen, *TIME* does separate the majority of the training samples for the two classes, however, there is an error in this separation that is larger for **NoWhen** class. The accuracy achieved using TIME is 85.52% on DEV that is 35.23% higher than the baseline and 6.53% higher than N-GRAMS. It performs even better on TEST with a slight accuracy increase of 0.03%.

Table 5.4: Effect of ignoring non time-bearing bigrams during calculation of *TIME* on accuracy of DEV.

Bigram	**Accuracy (%)**	**Accuracy Increase**
اخبار اليوم	*84.85*	*+0.16*
اليوم السابع	*85.02*	*+0.17*
المصري اليوم	*85.18*	*+0.16*
الحياة اليوم	*85.52*	*+0.34*

Figure 5.3: Distribution of training data on the values of feature *TIME* for both NoWhen and HasWhen classes.

TIME = 0		**TIME = 1**	
NoWhen	HasWhen	NoWhen	HasWhen
3002	322	761	3398

5.2.4 Experiment 4: Combining N-GRAMS, TIME-DISTANCE and TIME-CONTEXT

TIME-DISTANCE is based on TIME; it utilizes an observation from the dataset that time-bearing words referring to the time of the reported event tend to occur early on in the article. From Table 5.5, the average accuracy achieved using TIME-DISTANCE is 85.26% on DEV which is 34.97% higher than the baseline but it is less than TIME by 0.26% which questions the correctness of our observation.

Table 5.5: *answers_when* accuracy(%) on both DEV and TEST.

	NG (n = 1, 2)	*T*	*TD*	*TD* + *TC*	*NG* (n = 1, 2) + *TD* + *TC*
DEV	78.99	85.52	85.26	86.25	**86.65**
TEST	77.12	85.55	86.19	**87.68**	87.36
Baseline	50.29	50.29	50.29	50.29	50.29

TIME-DISTANCE + TIME-CONTEXT achieves an average accuracy of 86.25% on DEV which is slightly higher than both TIME-DISTANCE and TIME. This combination captures cases like this one:

خلال تغريدته على حسابه الشخصى على **تويتر اليوم الاثنين**

Our best setting on DEV is achieved by combining TIME-DISTANCE + TIME-CONTEXT + unigrams + bigrams which gives an average accuracy of 86.65%. Though this is the best setting, it has the bold disadvantage of having much higher dimensionality and hence more expensive computations with insignificant accuracy increase (only 0.52% increase over TIME-DISTANCE + TIME-CONTEXT)

5.2.5 Experiment 5: Effect of Preprocessing

This experiment demonstrates the effect of preprocessing on the accuracy of the model. The average accuracy of the model on DEV is computed once with preprocessing (in this case only normalization) and another time without preprocessing. From Table 5.6, we can see that normalization have a positive effect on the accuracy of the model with 2.5% higher accuracy than with raw text. This conforms with the recommendations of the literature on other text classification problems.

Table 5.6: Effect of preprocessing on accuracy of "When" model.

Without Preprocessing	**With Preprocessing**
83.67	86.31

5.2.6 Error Analysis

In this subsection we provide an analysis about cases that our model predicts incorrectly. We start by providing analysis for the bigger portion of the error; articles erroneously classified as **HasWhen**. We follow that by an analysis for the smaller portion of the error; articles erroneously classified as **NoWhen**. Our analysis makes use of examples to show common error cases.

5.2.6.1 Incorrect Classification of “NoWhen”

This part of the classification error, erroneously classifying an article as **HasWhen**, is the largest part of the error (≊ 56%). It is caused by two main reasons:

Noise in placing “NoWhen” labels This source of error is due to placing incorrect **NoWhen** labels by MCE Watch, as we mentioned previously each article is annotated by a single human without any mechanism to filter conflicting judgment in addition to the presence of human error. We provide multiple examples of this in Table 5.7.

Table 5.7: Examples of noisy “**NoWhen**” labels. (Text in bold in the left cell indicate the incorrectness of the expert reasoning in the right cell.)

Article Text	Expert Reasoning
نفى المهندس خيرت الشاطر نائب المرشد العام لجماعة الاخوان المسلمين اليوم **الاثنين** صحة الشائعات التى ترددت بشأن زواجه من عارضة أزياء سورية تبلغ من العمر ٢١ عاما مقابل مهر يبلغ مليون دولار، واصفا تلك الأنباء بالأكاذيب المغرضة التى تستهدف صرف الناس عن قضايا الوطن الحقيقية.	الخبر حدد التاريخ باليوم ولم يحدد اسم اليوم لذلك فانه لم يعطي للقارئ موعدا محددا للخبر
اكد مصدر دبلوماسي رفيع المستوي بوزارة الخارجية ان الخارجية لم تتلق حتى الآن اي اعتذار من المستشار محمود مكي بشأن منصب سفير مصر لدى الفاتيكان الذي تم ترشيحه له سابقا. وقال المصدر ، لوكالة أنباء الشرق الأوسط **اليوم الجمعة** ، ان وزارة الخارجية مستمرة في إجراءات سفر المستشار مكي.	لم يشر الخبر إلى توقيت تصريح المصدر بدقة
أجبر موظفو مجلس الدولة الدائرة الثالثة برفع جلستها المنعقدة **اليوم** بشكل مفاجئ، وذلك بعد قيام الموظفين باقتحام قاعة الدائرة الثالثة بالمجلس بمكبرات الصوت.	لم يشر إلى توقيت اجبار الموظفين للدائرة برفع الجلسة
صرح اللواء مهندس وائل المعداوى وزير الطيران المدني بأن أول القرارات التي إتخذها بعد توليه المنصب هو تشكيل لجنة من الماليين المحترفين لاعادة الهيكلة المالية لشركات قطاع الطيران وذلك لتجنب الخسائر التي تحدث فى القطاع نظرا للظروف السياسية التي تمر بها مصر. جاء ذلك خلال المؤتمر الصحفي الذى عقده الوزير مع عدد من قيادات الوزارة وشركات الطيران ورجال الصحافة والإعلام عقب إجراء تجربة الطوارئ بمطار الغردقة **الاثنين.**	لم يذكر الخبر تاريخ الحدث (تصريحات وزير الطيران) حيث ذكر انه اليومْ ولم يذكر التاريخ بدقة، ويجب التنويه الي ان تاريخ كتابة الخبر علي الموقع المكتوب اعلي التقرير الصحفي لا يعني انه تاريخ الحدث.

Erroneously interpreting time-bearing words This source of error is due to erroneously interpreting the time words. There are several examples of such case depicted in Table 5.8. Most of these erroneous interpretations are due to: (i) the article talks about different topics but doesn't mention the time for all of them so we erroneously capture the time for any of them and generalize it for the whole article. In other words, we capture only one time per article but a news article can report more than one event and needs to mention time for each of them (ii) time is mentioned but doesn't refer to the time of the event reported by the article instead it refers to the time of some other detail inside the reported event.

Table 5.8: Examples of erroneous interpretations for time-bearing words.

Article Text	Expert Reasoning	Explanation
درجة الحرارة تنخفض فجراً لـ ٥ درجات.. هيئة الأرصاد: أمطار غزيرة على القاهرة **الخميس** أكد الدكتور على قطب، مدير عام التحاليل والتنبؤات بهيئة الأرصاد الجوية، على أن درجة الحرارة فجراً	الخبر لم يذكر متى قال على قطب هذه التصريحات	Thursday refers to the expected time of heavy rain but not the time of Ali Kotb's speech.
أكد الدكتور حاتم عبد اللطيف وزير النقل أن أولوياته فى الفترة القادمة تختص بثلاث أولويات ضرورية منها الاهتمام بعنصر السلامة والأمان بكافة قطاعات الوزارة. وأضاف عبد اللطيف في تصريح صحفي اليوم أن الحكومة... **من ناحية أخرى قال محمد الشحات المستشار الإعلامي والمتحدث الرسمي لوزارة النقل إن الوزير أبدي اهتماما خاصا بالقطاعات والهيئات الخدمية في الوزارة وطلب علي الفور لقاءات عاجلة مع رؤساء هذه القطاعات والهيئات لبحث المشاكل التي تهم المواطن.**	لم يشر إلى توقيت إطلاق هذه التصريحات وذلك علي الرغم من أنه التزم بالمهنية وذكر الكيفية التى حصل بها على تصريحات الوزير وتوقيتها إلا أن الجزء الاخير من الخبر لم يراع ذلك	Article reports two speeches. The time for the first one was mentioned as "today" but the second one was ignored. We judge the article as a whole.
وطالب السائقون أثناء وقفتهم الاحتجاجية الشربيني بتقديم اعتذار لهم ونشره فى الصحف والفضائيات عما بدر منه تجاههم خلال مؤتمره الصحفى **أمس الثلاثاء** ، مهددين بالتصعيد تجاهه، ورفع دعوى قضائية ضده بتهمه إهاناتهم وسبهم وقذفهم ومحاولة تحريض الرأى العام ضدهم.	الخبر أخل بالمعيار آمتىْ حيث أنه لم يذكر متى اعتصم سائقى مترو الأنفاق اليوم أم أمس مما أخل بمصداقية الخبر.	The time we capture is that of a conference which may be different from the time of the vigil which is the topic of the article. The vigil's time is not reported.
صرح مصدر مسئول فى حزب التجمع بأن مشروع قانون التظاهر والاعتصام الذى أعدته اللجنة التشريعية بمجلس الشورى الاخوانى السلفى بالاشتراك مع لجنة حقوق الإنسان بالمجلس ونشرت ملامحه صحيفة الحرية والعدالة الناطقة باسم حزب الحرية والعدالة **يوم الأحد** هو قانون استبدادى...	لم يشر إلى توقيت الحصول عليه	The time we capture doesn't refer to the time of the speech which is the topic of the article. Instead it refers to the time of some other event described in the speech.

5.2.6.2 Incorrect Classification of “HasWhen”

This part of the error, erroneously classifying as **NoWhen**, is the smallest part of the error (≊ 44%). It is caused by two main reasons:

Noise in Placing “HasWhen” labels This source of error is again due to placing incorrect **HasWhen** labels by MCE Watch. We show multiple examples of such incorrect labeling in Table 5.9.

Cases not captured by our features There are multiple cases in which our features don’t capture the time indication from the article’s text. In Table 5.10 and Table 5.11 we show the two most common cases. Table 5.10 presents examples of cases where there is an explicit mention of a time-bearing word in the article’s text, however, our TIME-DISTANCE feature incorrectly ignores these words because they appear somewhat far to the end of the article. Actually, the heuristic that motivated us to develop TIME-DISTANCE was that time-bearing words appearing towards the end of the article are most likely not referring to the time of the reported event but probably to the time of some other detail in the reported event. However, as can be seen from the examples, this heuristic doesn’t work for all cases i.e. there are some cases in which time-bearing words appear at the end of the article but still refer to the time of the reported news. Table 5.11 presents examples of cases where there is an implicit mention of the time of the reported event but our model fails to capture it. Typically such implicit mentions refer to an event like a conference, a ceremony or some other event that happens annually or once in a life time and the reported news occurred during it. However, cases like when the article refers to a daily show is not considered enough, instead the the day of the episode that contained the reported news is necessary to eliminate any ambiguity.

5.2.7 Summary of “When?” Experimental Results

Table 5.5 summarizes our experiments, first we try all combinations of N-GRAMS for $n \leq 5$ and determine the best setting on DEV, this was found to be the combination of unigrams + bigrams that gives 78.99% average accuracy on DEV which is 28.7% higher than the baseline and gives 77.12% accuracy on TEST which is 26.83% higher than the baseline. The domain dependent feature *TIME* gives an accuracy of 85.52% on DEV that is 35.23% higher than the baseline and 6.53% higher than *N-GRAMS*. It performs even better on TEST with a slight accuracy increase of 0.03%. The right part of Table 5.5 displays results of using the combination of *TIME-DISTANCE* and *TIME-CONTEXT* features which performs 1.0% and 1.5% higher than *TIME* for DEV and TEST, respectively. Combining these two features with *N-GRAMS* – by appending them to the L2 normalized N-GRAMS feature vectors and re-normalize – gives our best setting on DEV that exceeds the baseline by 36.3%. Although this setting is 37% higher than the baseline for TEST, it is -0.32% less than *TIME-DISTANCE* + *TIME-CONTEXT* i.e. it doesn’t generalize as

Table 5.9: Examples of noisy “**HasWhen**” labels.

Article Text	**Explanation**
٢ إبريل الحكم في دعوى عزل هشام قنديل لامتناعه عن تنفيذ حكم قررت محكمة جنح الدقي تحديد جلسة ٢ إبريل المقبل للنطق بالحكم في دعوى حبس وعزل رئيس الوزراء الدكتور هشام قنديل، لامتناعه عن تنفيذ حكم محكمة القضاء الاداري المقامة من عمال طنطا للكتان. وكأنت المحكمة قد أصدرت قرارها فى الجلسة الماضية، بإعادة فتح باب المرافعة فى القضية. يذكر أن أحد المحامين قد أقام دعوى بصفته وكيلاً عن عمال شركة طنطا للكتان، رقم ١٢٠١١ لسنة ٢٠١٢ ضد رئيس الوزراء بشخصه، مطالبًا فيها بحبسه وعزله، لامتناعه عن تنفيذ حكم محكمة القضاء الإداري الصادر فى الدعوى رقم ٣٤٢٤٨ ببطلان بيع مصنع طنطا للكتان للسعودي عبدالإله الكعكي.	The article doesn’t mention the time of the court’s decision. In other words, it doesn’t mention the time of the court session in which the decision was made. A correct example would be:“قررت محكمة جنح القاهرة في جلستها المنعقدة اليوم...”
علاء مبارك: ههاجر لاستراليا قال جمال مبارك نجل الرئيس المخلوع محمد حسنى مبارك أثناء تجوله داخل قفص الاتهام فى قضية التلاعب بالبورصة : أنا مش هسيب بلدي وهعيش فى مصر ومش هسافر أعيش برة. وذكر محرر مصراوى أن علاء مبارك دخل فى الحوار ساخراً: أنا ههاجر لاستراليا، وكان رد جمال على هذه الجملة أن شقيقه الأكبر يقول ذلك على سبيل المزاح.	The article again doesn’t mention when the quoted speeches were said. It mentioned that the speeches were said in the court’s cage during the trial of Gamal and Alaa Mubarak but it didn’t indicate the time of the court session in which the reporter acquired the speeches from them.
الطب الشرعي: التقرير النهائي عن الجندي الأسبوع المقبل أكد دكتور إحسان كميل جورجي، رئيس مصلحة الطب الشرعي علي أنه تم الكشف علي الناشط الراحل محمد الجندي من خلال لجنة من الطب الشرعي وأن هناك لجنة أخرى كشفت عليه، لكن لم تصدر أي من اللجنتين التقرير النهائي عن سبب الوفاة، مؤكدا علي عدم تسلم النيابة أي تقارير خاصة بالجندي، مشيرا إلي تسلم التقرير النهائي مطلع الأسبوع المقبل، وذلك فى مداخلة هاتفية مع الإعلامي جابر القرموطي في برنامج مانشيت.	The article doesn’t mention the airing time of the TV show that Dr. Ihsan called. Even if the show has fixed known times it is still expected to mention the date of the episode that included these statements.

good as TIME-DISTANCE + TIME-CONTEXT. This setting also has the disadvantage of very high dimensionality with an insignificant increase in accuracy when compared to TIME-DISTANCE + TIME-CONTEXT.

Table 5.10: Examples of classification errors of **HasWhen** due to mistakes of TIME-DISTANCE.

Article Text	**Explanation**
حزب الحرية والعدالة : نحترم أحكام القضاء وسننتظر قرارات العليا للانتخابات أعرب حزب الحرية والعدالة، الذراع السياسي لجماعة الإخوان المسلمين، عن احترامه لأحكام القضاء، في إشارة إلى قرار محكمة القضاء الاداري بوقف انتخابات مجلس النواب. وقال الدكتور مراد علي المستشار الإعلامي لحزب الحرية والعدالة، في تصريح مقتضب له **الأربعاء** : نحترم أحكام القضاء، وسننتظر ما ستتخذه اللجنة العليا للانتخابات.	The article does mention the time of the statement (Wednesday), however, our model misses it since it is far to the end of the article. In other words, TIME-DISTANCE ≊ 1.16 which erroneously implies that the time-bearing word "Wednesday" does not refer to the time of the statement.
رئاسة الوزراء : بعض الأفراد يجبرون الموظفين على مغادرة المصالح الحكومية للترويج للعصيان المدنى صرح المتحدث الرسمي باسم رئاسة مجلس الوزراء بأن ما يتردد فى وسائل الاعلام وبعض المواقع الإلكترونية عن العصيان المدنى عار تماماً من الصحة ، وأن الذى يحدث فعلياً أن بعض الأفراد يحاولون إجبار الموظفين ببعض المصالح الحكومية وخاصة دواوين بعض المحافظات مثل محافظة الدقهلية ـ على الخروج ويعتدون عليهم فى حالة عدم الاستجابة لهم. وقال المتحدث فى بيان له **اليوم** إن الحكومة تحى وتقدر وعى المواطنين فى كافة مواقع العمل والذين لم يستجيبوا لهذه الدعوات التى لا تريد بمصر خيراً وتعطل عجلة الإنتاج وتعطل مصالح المواطنين، وأنها تهيب بالمواطنين العاملين فى مواقع العمل والانتاج أن يعاونوا جهاز الشرطة حرصاً على الصالح العام.	Again the article mentions the time (today) almost near the end of the article so it gets a value of 1.45 for TIME-DISTANCE and our model erroneously ignores it.

5.3 "Source Info" Experimental Results

For this problem we have a total of 4283 articles distributed among two classes. Class **HasSourceInfo** has a total of 2606 articles and class **NoSourceInfo** has a total of 1677 articles.

Table 5.11: Examples of classification errors of **HasWhen** due to unhandled time indicators.

Article Text	Explanation
محافظ كفر الشيخ يشارك الإخوة المسيحيين الاحتفال بعيد الميلاد **قدم المهندس سعد الحسينى محافظ كفر الشيخ التهنئة للاخوة الأقباط بمناسبة عيد الميلاد المجيد بكنيسة مارى جرجس** بمدينة كفر الشيخ بحضور الدكتور ماجد القمري رئيس جامعة كفر الشيخ ، واللواء محمد الشاذلى مدير امن كفر الشيخ والمهندس حافظ عيوى السكرتير العام والدكتور محمد فؤاد عبد المجيد القيادي بجماعة الإخوان المسلمين ، ومدير الامن الوطنى ومدير المخابرات العا ووكيل وزارة الأوقاف والقيادات السياسية والشعبية والتنفيذية ورجال الدين المسيحي والإسلامى. حيث كان فى استقبال المحافظ القمص بطرس بطرس بسطوروس وكيل مطرانيه كفر الشيخ ورجال الكنيسة. وقد تحدث القمص بطرس.....	The article here doesn't explicitly state a timing but it mentions it implicitly. This is obvious in the sentence which indicated that Kafr El-Sheik governor congratulated the Egyptian copts for Christmas during the ceremony held in Margirgis church. This implies that the time we are interested in is the same time of the ceremony.
دروجبا لـتريكة: أنت دائمًا تبرهن على أنك أسطورة حرص ديديه دروجبا لاعب شنغهاي الصيني، ومنتخب كوت ديفوار على الاشادة بالنجم المصري محمد أبو تريكه لاعب النادي الأهلى، وصاحب لقب أفضل لاعب أفريقي داخل القارة، **خلال حفل التكريم الذي جمع الاثنين معًا في العاصمة الغانية أكرا** . وقالت صحيفة إيفواريه : إن الفيل الإيفواري حرص على تهنئة لاعب الوسط المصري، كما نشرت حديثاً مختصرًا دار بين دروجبا وأبو تريكة، أعرب خلاله العاجى عن إعجابه بأبو تريكة. وقال دروجبا لـأبو تريكة : أهنئك على جائزة أفضل لاعب في افريقيا، فأنت دائمًا تبرهن على أنك أسطورة.	The article mentions the time implicitly by noting that Drogba congratulated Abu Treka during the ceremony in Accra, however, our model didn't capture this.

5.3.1 Experiment 1: Comparing Multiple Classifiers

In this experiment we use all combinations of n-grams for $n \leq 3$ to compare the efficiency of KNN, NBC, and SVM classifiers and determine the best performing one. From Figure 5.4 (Table 5.12 shows the exact figures), It can be seen that KNN and NBC perform approximately similar in most of the combinations of n-grams except for trigrams where KNN with small values of k perform much worse as a result of the sparsity of trigrams in the small dataset. On the other hand, SVM performs best on all combinations of n-grams which agrees with literature.

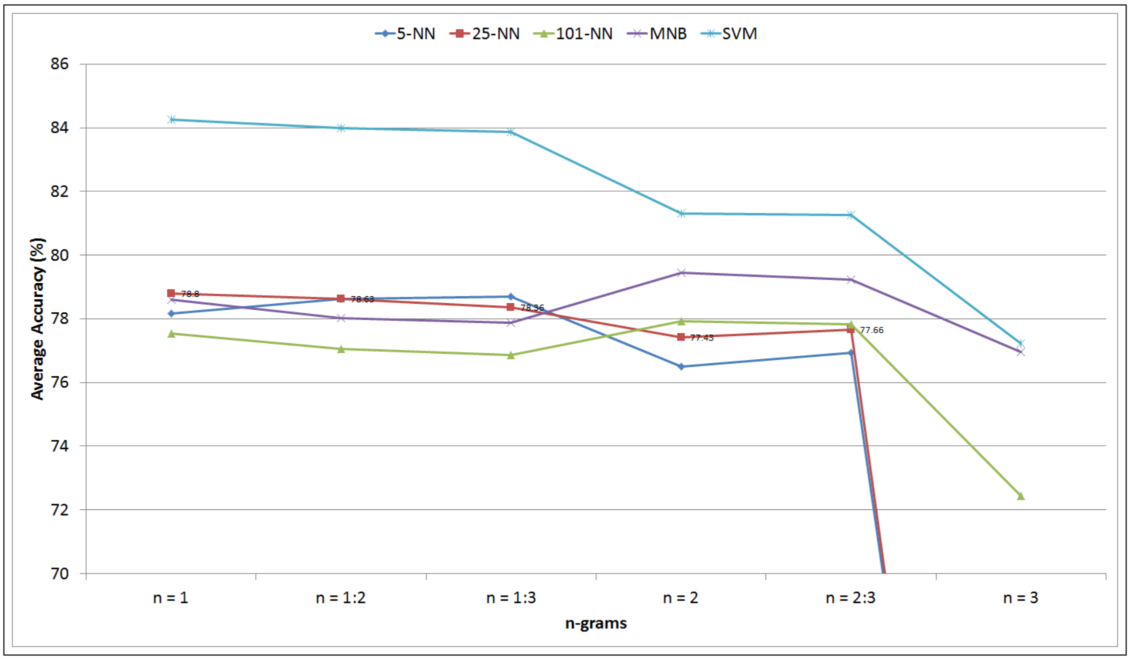


Figure 5.4: Comparison between KNN (n = 5, n = 25, n = 101), NBC, and SVM classifiers using all combinations of n-grams for $n \leq 3$

Table 5.12: Exact figures for comparison between KNN (k = 5, 25, k = 101), NBC, and SVM classifiers using all combinations of n-grams for $n \leq 3$

	5-NN	**25-NN**	**101-NN**	**NBC**	**SVM**
n = 1	78.16	78.8	77.54	78.6	84.26
n = 1, 2	78.63	78.63	77.05	78.01	84
n = 1, 2, 3	78.71	78.36	76.87	77.87	83.88
n = 2	76.49	77.43	77.92	79.44	81.31
n = 2, 3	76.93	77.66	77.84	79.24	81.25
n = 3	39.98	39.16	72.43	76.96	77.22

From now on, we will carry out our experiments using SVM.

5.3.2 Experiment 2: N-grams selection

In this subsection we run experiments to determine the best setting of N-GRAMS on DEV. First we try all combinations of N-GRAMS for $n \leq 5$; we do this once using TF-IDF weighting and another time using TF without IDF weighting. Figure 5.5 shows a bar chart comparing both settings. The same conclusions drawn in Section 5.2.2 apply again: *Combinations of N-GRAMS that don't use IDF weighting give higher accuracies than those with IDF weighting.*

Table 5.13 shows the actual accuracy figures for N-GRAMS using TF only on both DEV and TEST. Unigrams give the highest accuracy on DEV. It exceeds the baseline by

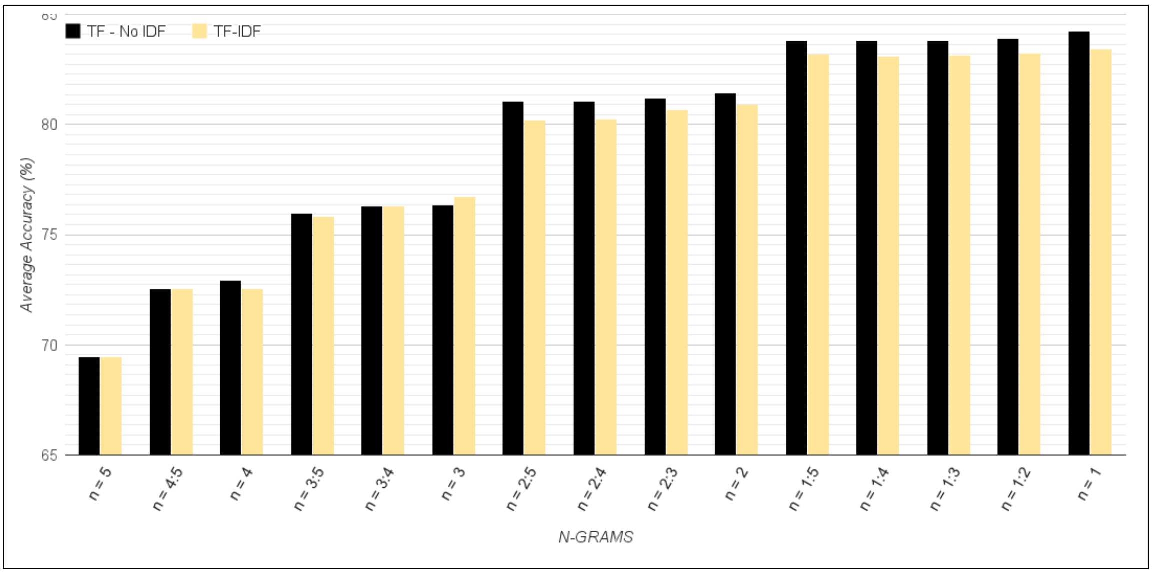


Figure 5.5: Comparing results for N-GRAMS when $n \leq 5$ using TF-IDF weighting versus TF only.

23.4%. Although unigrams + bigrams + trigrams together give an accuracy that exceeds unigrams alone by 0.8% on TEST, we still use the setting performing best on DEV i.e. unigrams in our next experiments.

Table 5.13: *has_source* accuracy(%) on both DEV and TEST using all combinations of *N-GRAMS* $n \leq 3$.

	DEV	TEST	Baseline
n = 1	**84.23**	82.54	60.8
n = 1, 2	83.91	82.89	60.8
n = 1, 2, 3	83.82	82.77	60.8
n = 1, 2, 3, 4	83.79	82.77	60.8
n = 1, 2, 3, 4, 5	83.79	83.54	60.8
n = 2	81.43	79.16	60.8
n = 2, 3	81.19	79.28	60.8
n = 2, 3, 4	81.08	79.74	60.8
n = 2, 3, 4, 5	81.05	79.74	60.8
n = 3	76.37	73.57	60.8
n = 3, 4	76.29	73.22	60.8
n = 3, 4, 5	75.99	73.46	60.8
n = 4	72.95	66.47	60.8
n = 4, 5	72.57	66.71	60.8
n = 5	69.45	64.49	60.8

5.3.3 Experiment 3: Effect of Domain Dependent Features

First of all, we run the SFS algorithm to determine the entries of VERB lexicon. Table 5.14 shows each step in the execution of the SFS algorithm.The final entries of the lexicon are shown in Table 4.4.

Table 5.14: Using SFS to develop the *VERB*-lexicon.

ناقشت	نفي	تابعت	انتقدت	اضافت	اشارت	اكدت	اوضحت	قالت	اكد	تابع	اضاف	انتقد	اشار	اوضح	قال
60.84	60.84	60.84	60.84	60.84	60.84	60.84	60.84	60.84	60.84	60.84	60.84	60.84	60.84	60.84	**67.9**
68.08	68.72	68.28	68.02	69.22	69.07	69.80	60.81	70.44	**73.27**	68.60	72.34	68.22	71.5	71.17	-
73.42	73.57	73.57	73.39	74.33	74.10	74.48	74.04	**75.23**	-	73.51	75.03	73.45	74.68	74.77	-
75.35	75.50	75.38	75.29	75.32	75.59	75.50	75.64	-	-	75.50	**76.87**	75.35	76.61	76.73	-
76.99	77.05	76.99	76.93	76.93	77.19	77.14	77.25	-	-	76.87	-	76.90	77.72	**77.84**	-
77.92	77.95	77.92	77.87	77.89	78.13	77.90	78.22	-	-	77.75	-	77.87	**78.42**	-	-
78.48	78.45	78.51	78.45	78.48	78.71	78.36	**78.74**	-	-	78.3	-	78.45	-	-	-
78.80	78.77	78.86	78.77	78.74	**78.89**	78.57	-	-	-	78.62	-	78.77	-	-	-
78.95	78.92	**78.98**	78.89	78.89	-	78.77	-	-	-	78.77	-	78.92	-	-	-
79.03	79.00	-	78.98	78.95	-	78.86	-	-	-	78.86	-	79.00	-	-	-
-	**79.06**	-	79.03	79.00	-	78.92	-	-	-	78.92	-	79.06	-	-	-
-	-	-	79.06	79.03	-	78.98	-	-	-	78.95	-	**79.09**	-	-	-
-	-	-	79.09	79.06	-	79.01	-	-	-	78.98	-	-	-	-	-

The best accuracy achieved on DEV using VERB feature is - as shown in Table 5.14 - 79.09%

Next, we combine VERB with SOURCE to get an accuracy of 80.03%. Table 5.15 presents the execution of SFS algorithm to build the SOURCE lexicon.

Table 5.15: Using SFS to develop the *SOURCE*-lexicon.

الاناضول	اسوشيتد برس	فوكس نيوز	رويترز	ا ش ا
79.09	79.09	79.09	79.09	**80.23**
80.32	80.23	80.26	80.26	-
-	80.32	**80.35**	80.35	-
-	80.35	-	**80.37**	-
-	**80.38**	-	-	-

Table 5.16 shows accuracies for domain dependent features. Although, best results for domain dependent features *VERB + SOURCE* exceeds baseline by 19.5% and 14.2% on DEV and TEST respectively, it is still lower than results of unigrams by 3.8% for DEV and 7.4% for TEST. Combining both of them with unigrams gives our best setting on both DEV and TEST with accuracies exceeding baseline by 24% and 21.6%, respectively.

As can be seen, domain dependent features for *"has_source"* weren't as good as their corresponding for *"answers_when"*. We attribute this to the lack of **NoSourceInfo** training examples; we had a total of 1341 examples which is slightly more than 35% o f what we had for **NoWhen**.

Table 5.16: *has_source* accuracy(%) on both DEV and TEST.

	NG (**n = 1**)	$VERB$	$VERB + SOURCE$	$NG + VERB + SOURCE$
DEV	84.23	79.09	80.38	**85.02**
TEST	82.54	75.44	75.09	**82.42**
Baseline	60.8	60.8	60.8	60.8

5.3.4 Experiment 4: Effect of Preprocessing

This experiment demonstrates the effect of preprocessing on the accuracy of the model. We report the average accuracy of our model on DEV once without any preprocessing, another time with normalization only and a third time with a combination of normalization and stripping of ال and و. From Table 5.17, we can see that normalization have a positive effect on the accuracy of the model which conforms with the results previously presented in literature. We also note that stripping و and ال further increases the accuracy for the reasons mentioned in Section 4.3.

Table 5.17: Comparing results of different preprocessing.

No Preprocessing	**Normalization**	**Normalization + Stripping of و and ال**
84.37	84.87	85.02

5.4 Credibility Score per News Site

In this last experiment we compute the credibility score of each news site in the same way as MCE Watch for all news articles evaluated in April 2013. Basically, for each news site we classify all articles from this site according to *has_source* and *answers_when* criteria. For all articles violating a criteria we subtract the corresponding points of the violated criteria from the monthly pool of 500 points assigned to each news site. The percentage of remaining points corresponds to the credibility score of the news site, Equation 5.1.

$$\begin{aligned} violations &= P_{has\_source} \times NS_i + P_{answers\_when} \times NW_i \\ Score_i &= \frac{500 - violations}{500} \times 100\% \end{aligned} \tag{5.1}$$

Where

- $Score_i$: the credibility score assigned to news site i at the end of April 2013.
- P_{has_source}: the points assigned to *has_source*. From Table 3.2, $P_{has\_source} = 2$.
- $P_{answers\_when}$: the points assigned to *answers_when*. From Table 3.2, $P_{answers\_when} = 1$.
- NS_i: the number of articles from source i that violate *has_source* criteria at the end of April 2013.
- NW_i: the number of articles from source i that violate *answers_when* criteria at the end of April 2013.

Figure 5.6 shows the score calculated by MCE Watch versus the score calculated based on our model. The Mean Absolute Error (MAE) is 17.56% and the Root Mean Square Error (RMS) is 17.98. Two factors contribute to this error (i) the error introduced by our classification (ii) and the error due to the criteria we don't consider. To get a closer picture, we adjust the scores assigned by MCE Watch by adding back points deducted due to criteria that we don't consider. Figure 5.7 shows the comparison between our scores and the adjusted MCE Watch scores, it is clear that the gap between the scores is much less. Consequently, MAE drops significantly to 2.55% and RMS also drops significantly to 3.16%. We note a couple of issues here: (i) The larger portion of the error is introduced by classification errors for *has_source*. This is very clear for youm7.com, shorouknews.com, rassd.com, fj-p.com, and elfagr.org with highest absolute error in the range [4, 6.4] (ii) The drop in MAE is not due to our model accuracy alone but there is still a hidden error that contributes towards this drop; this happens when erroneously classified articles contribute positively in penalizing (or not penalizing) a news site. However, this can be used as a high level indicator for a news site credibility.

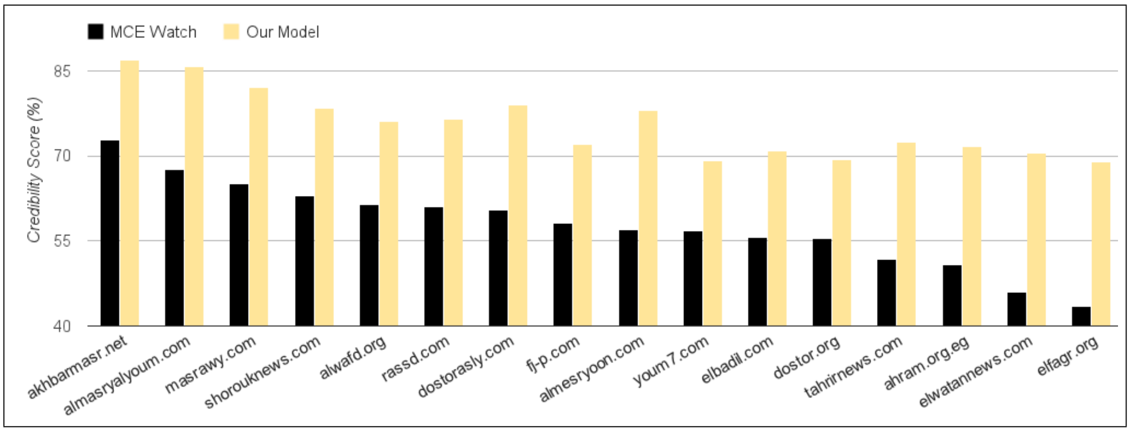


Figure 5.6: Credibility score of each news site measured by MCE Watch vs. score based on our model

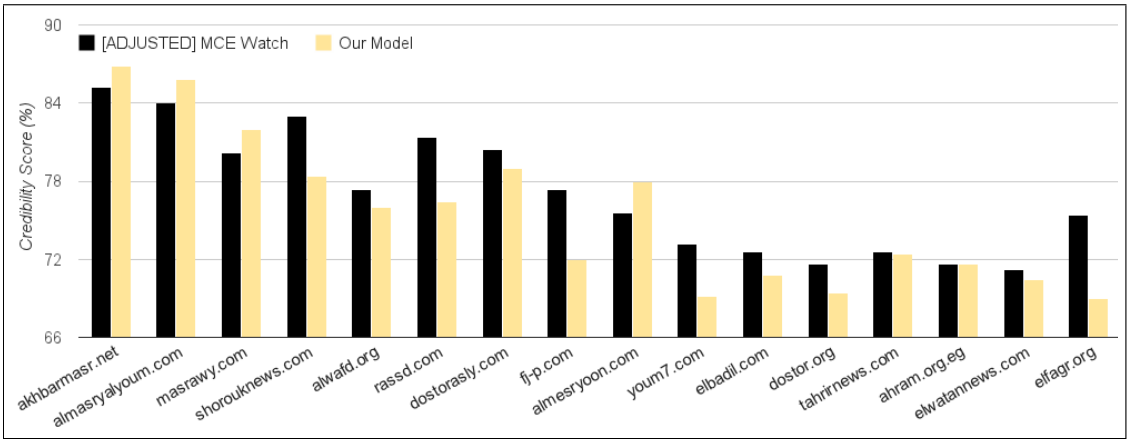


Figure 5.7: Adjusted credibility score of each news site measured by MCE Watch vs. scored based on our model

The last experiment modifies Equation 5.1 slightly such that the constant 500 in the equation is replaced with the maximum possible value that could be achieved by a news site if all the sample articles taken from this news sites don't violate any of the two credibility criteria, Equation 5.2.

$$Score_i = \frac{sample_i * (P_{has\_source} + P_{answers\_when}) - violations}{sample_i * (P_{has\_source} + P_{answers\_when})} \times 100\% \tag{5.2}$$

where $sample_i$ is the number of articles assessed for news site i at the end of April 2013. All the other variables in the equation are defined similar to Equation 5.1.

From Equation 5.2, if all articles coming from a news site violate both credibility criteria then the credibility score for the news site will be 0% and if all articles coming from a news site are credible with respect to the two credibility criteria then the credibility score for the news site will be 100%. Thus $Score_i$ is in the range [0, 100]. Figure 5.8 shows the comparison between our scores and adjusted MCE Watch scores after applying Equation 5.2 to both of them. The MAE in this case is 3.10% which is approximately 0.6% higher than that produced by Equation 5.1. We also compute the RMS which has the value of 3.83 and is higher than that produced by Equation 5.1 by 0.7 approximately.

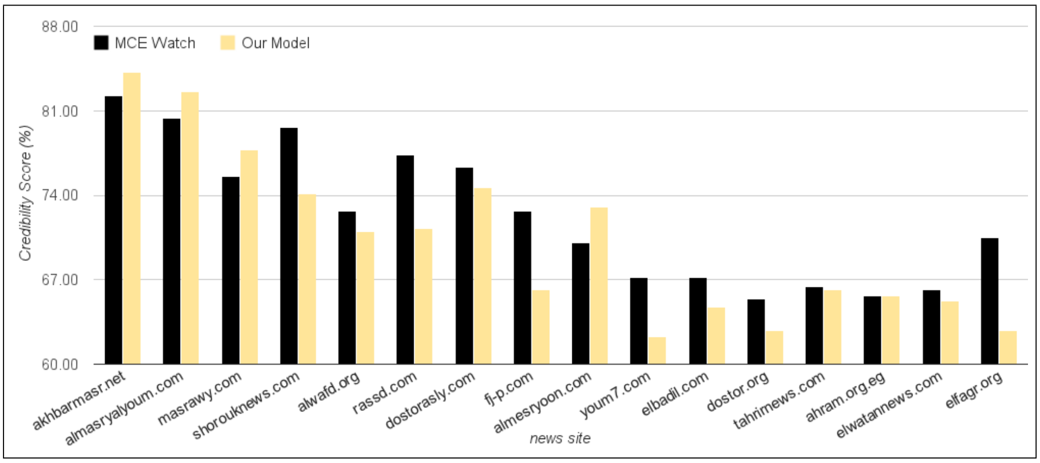


Figure 5.8: Adjusted credibility score of each news site measured by MCE Watch vs. scored based on our model using Equation 5.2

Chapter 6: Conclusion

In this work we presented a new approach to automate Arabic news credibility measurement. Our approach is a first step towards automating a manual solution for the same problem presented by MCE Watch. The solution depends on evaluating each news article on a preset collection of criteria developed by media experts. Criteria are evaluated according to the text content of the news article. To automate this, we crawled news articles previously labeled by MCE Watch. Then, we picked two criteria of the most violated ones, and modeled each one as a separate binary classification problem dividing the crawled data into training and test sets.

The first criterion we chose is whether the article answers the question *"When?"*. Using SVM and a combination of unigrams, bigrams and domain dependent features we were able to reach an accuracy of 87% on a total of 1875 articles representing 20% of the crawled data. Our baseline is 50.29% which is the majority class.

The second criterion we chose is whether the news article mentions the *“Source Info”*. Using SVM, unigrams and domain dependent features we got an accuracy of 82% on a total of 859 articles representing 20% of all the data we have for this problem and 9.1% of the crawled data. Our baseline is 60.8% representing the majority class.

As for future work, we aim at collecting more data to be able to automate the assessment of more of the credibility criteria. Also, enhance the models for *answers_when* and *has_source*. Design stronger baselines probably using the state-of-art automated feature extractors.

References

[1] Abdulla, R. A., Garrison, B., Salwen, M., Driscoll, P., and Casey, D. The credibility of newspapers, television news, and online news. In *Education in Journalism Annual Convention, Florida USA* (2002).

[2] Akamine, S., Kato, Y., Inui, K., and Kurohashi, S. Using appearance information for web information credibility analysis. In *Second International Symposium on Universal Communication, ISUC* (2008), IEEE, pp. 363–365.

[3] Akamine, S., Kawahara, D., Kato, Y., Nakagawa, T., Inui, K., Kurohashi, S., and Kidawara, Y. Wisdom: a web information credibility analysis system. In *Proceedings of the ACL-IJCNLP Software Demonstrations* (2009), Association for Computational Linguistics, pp. 1–4.

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

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

[6] Al-Shalabi, R., Kanaan, G., and Gharaibeh, M. Arabic text categorization using knn algorithm. In *Proceedings of The 4th International Multiconference on Computer Science and Information Technology* (2006), vol. 4, pp. 5–7.

[7] Al-Shalabi, R., and Obeidat, R. Improving knn arabic text classification with n-grams based document indexing. In *Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt* (2008).

[8] Aljazeera. `http://www.aljazeera.net`, April 2013.

[9] Alsaleem, S. Automated arabic text categorization using svm and nb. *International Arab Journal of e-Technology 2*, 2 (2011).

[10] Associated Press. Ap. `http://www.ap.org`. [Online; accessed 08-May-2013].

[11] Blanzieri, E., and Bryl, A. Instance-based spam filtering using svm nearest neighbor classifier. *Proceedings of FLAIRS-20* (2007), 441–442.

[12] Blanzieri, E., and Bryl, A. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review 29*, 1 (2008), 63–92.

[13] Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (2011), ACM, pp. 675–684.

[14] Chhabra, P., Wadhvani, R., and Shukla, S. Spam filtering using support vector machine.

[15] Çiltik, A., and Güngör, T. Time-efficient spam e-mail filtering using| n|/-gram models. *Pattern Recognition Letters 29*, 1 (2008), 19–33.

[16] Clark, J., Koprinska, I., and Poon, J. A neural network based approach to automated e-mail classification. In *Proceedings of International Conference on Web Intelligence (WI)* (2003), IEEE/WIC, pp. 702–705.

[17] Darwish, K., Magdy, W., and Mourad, A. Language processing for arabic microblog retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (2012), ACM, pp. 2427–2430.

[18] Dictionary, W. C. What is spam? `http://www.webopedia.com/TERM/S/spam.html`. [Online; accessed 08-May-2013].

[19] Drucker, H., Wu, D., and Vapnik, V. N. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks 10*, 5 (1999), 1048–1054.

[20] Farghaly, A., and Shaalan, K. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP) 8*, 4 (2009), 14.

[21] Fecyk, G. Designated mailers protocol. may 2004. *Internet Engineering Task Force 11* (2004).

[22] Gaziano, C., and McGrath, K. Measuring the concept of credibility. *Journalism Quarterly 63*, 3 (1986), 451–462.

[23] Gharib, T. F., Habib, M. B., and Fayed, Z. T. Arabic text classification using support vector machines. *International Journal of Computers and Their Applications 16*, 4 (2009), 192–199.

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

[25] Guzella, T. S., and Caminhas, W. M. A review of machine learning approaches to spam filtering. *Expert Systems with Applications 36*, 7 (2009), 10206–10222.

[26] Hammad, M., and Hemayed, E. Automating credibility assessment of arabic news. In *Social Informatics*, A. Jatowt, E.-P. Lim, Y. Ding, A. Miura, T. Tezuka, G. Dias, K. Tanaka, A. Flanagin, and B. Dai, Eds., vol. 8238 of *Lecture Notes in Computer Science*. Springer International Publishing, 2013, pp. 139–152.

[27] Islam, M. R., Chowdhury, M. U., and Zhou, W. An innovative spam filtering model based on support vector machine. In *International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce* (2005), vol. 2, IEEE, pp. 348–353.

[28] Joachims, T. *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[29] Juffinger, A., Granitzer, M., and Lex, E. Blog credibility ranking by exploiting verified content. In *Proceedings of the 3rd workshop on Information credibility on the web* (2009), ACM, pp. 51–58.

[30] Kang, B., O'Donovan, J., and Höllerer, T. Modeling topic specific credibility on twitter. In *Proceedings of the ACM international conference on Intelligent User Interfaces* (2012), ACM, pp. 179–188.

[31] Kawahara, D., Kurohashi, S., and Inui, K. Grasping major statements and their contradictions toward information credibility analysis of web contents. In *International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (2008), vol. 1, IEEE/WIC/ACM, pp. 393–397.

[32] Kawai, Y., Fujita, Y., Kumamoto, T., Jianwei, J., and Tanaka, K. Using a sentiment map for visualizing credibility of news sites on the web. In *Proceedings of the 2nd ACM workshop on Information credibility on the web* (2008), ACM, pp. 53–58.

[33] Khreisat, L. Arabic text classification using n-gram frequency statistics a comparative study. In *Conference on Data Mining— DMIN* (2006), vol. 6, p. 79.

[34] Kibriya, A. M., Frank, E., Pfahringer, B., and Holmes, G. Multinomial naive bayes for text categorization revisited. In *AI 2004: Advances in Artificial Intelligence*. Springer, 2005, pp. 488–499.

[35] Kim, S.-B., Han, K.-S., Rim, H.-C., and Myaeng, S. H. Some effective techniques for naive bayes text classification. *IEEE Transactions on Knowledge and Data Engineering 18*, 11 (2006), 1457–1466.

[36] Kuipers, B. J., Liu, A. X., Gautam, A., and Gouda, M. G. Zmail: zero-sum free market control of spam. In *25th IEEE International Conference on Distributed Computing Systems Workshops.* (2005), IEEE, pp. 20–26.

[37] Lan, M., Tan, C.-L., Low, H.-B., and Sung, S.-Y. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *Special interest tracks and posters of the 14th international conference on World Wide Web* (2005), ACM, pp. 1032–1033.

[38] Larkey, L. S., and Croft, W. B. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* (1996), ACM, pp. 289–297.

[39] Lee, R. S. Credibility of newspaper and tv news. *Journalism Quarterly 55*, 2 (1978), 282–87.

[40] Lewis, D. D. Naive (bayes) at forty: The independence assumption in information retrieval. In *Machine learning: ECML-98*. Springer, 1998, pp. 4–15.

[41] Lewis, D. D., and Gale, W. A. A sequential algorithm for training text classifiers. In *SIGIR94* (1994), Springer, pp. 3–12.

[42] Maier, S. R. Accuracy matters: A cross-market assessment of newspaper error and credibility. *Journalism & Mass Communication Quarterly 82*, 3 (2005), 533–551.

[43] Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to information retrieval*, vol. 1. Cambridge University Press Cambridge, 2008.

[44] MCE Watch. Media credibility in egypt. `http://www.mcewatch.com`, April 2013.

[45] Miniwatts Marketing group. Internet world stats. `http://www.internetworldstats.com`, April 2013.

[46] Newhagen, J., and Nass, C. Differential criteria for evaluating credibility of newspapers and tv news. *Journalism Quarterly 66*, 2 (1989), 277–84.

[47] Pantel, P., and Lin, D. Spamcop: A spam classification & organization program. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization* (1998), pp. 95–98.

[48] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research 12* (2011), 2825–2830.

[49] Rimmer, T., and Weaver, D. Different questions, different answers? media use and media credibility. *Journalism & Mass Communication Quarterly 64*, 1 (1987), 28–44.

[50] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. A bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop* (1998), vol. 62, pp. 98–105.

[51] Saudi press agency. `http://www.spa.gov.sa`, April 2013.

[52] Schaivone, V., et al. Trusted email open standard–a comprehensive policy and technology proposal for email reform. *EPrivacyGroup Website, http://www. eprivacygroup. com* (2003).

[53] Schwarz, J., and Morris, M. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the annual conference on Human factors in computing systems* (2011), ACM, pp. 1245–1254.

[54] Sculley, D., and Wachman, G. M. Relaxed online svms for spam filtering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (2007), ACM, pp. 415–422.

[55] Sebastiani, F. Machine learning in automated text categorization. *ACM computing surveys (CSUR) 34*, 1 (2002), 1–47.

[56] Smola, C. D., and Vishwanathan, S. *Introduction to machine learning*, vol. 1. Cambridge University Press Cambridge, 2008.

[57] SPAMHAUS. The definition of spam. `http://www.spamhaus.org/consumer/definition/`. [Online; accessed 08-May-2013].

[58] SPAMHAUS. The spam definition and legalization game. `http://www.spamhaus.org/news/article/9`. [Online; accessed 08-May-2013].

[59] Sundar, S. S. Effect of source attribution on perception of online news stories. *Journalism & Mass Communication Quarterly 75*, 1 (1998), 55–68.

[60] Thomson Reuters. Reuters. `http://www.ap.org`. [Online; accessed 08-May-2013].

[61] Von Ahn, L., Blum, M., Hopper, N. J., and Langford, J. Captcha: Using hard ai problems for security. In *Advances in CryptologyEUROCRYPT 2003*. Springer, 2003, pp. 294–311.

[62] Weerkamp, W., and de Rijke, M. Credibility-inspired ranking for blog post retrieval. *Information retrieval 15*, 3-4 (2012), 243–277.

[63] Wikipedia. Egyptian arabic. `http://en.wikipedia.org/wiki/Egyptian_Arabic`. [Online; accessed 12-June-2013].

[64] Wikipedia. K-nearest neighbors algorithm. `https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm`. [Online; accessed 12-June-2013].

[65] Wikipedia. Lebanese arabic. `http://en.wikipedia.org/wiki/Lebanese_Arabic`. [Online; accessed 12-June-2013].

[66] Wikipedia. List of languages by number of native speakers. `http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers`. [Online; accessed 12-June-2013].

[67] Wikipedia. Naive bayes classifier. `https://en.wikipedia.org/wiki/Naive_Bayes_classifier`. [Online; accessed 12-June-2013].

[68] Wikipedia. Polysynthetic language. `http://en.wikipedia.org/wiki/Polysynthetic_language`. [Online; accessed 12-June-2013].

[69] Wikipedia. Semitic languages. `http://en.wikipedia.org/wiki/Semitic_languages`. [Online; accessed 12-June-2013].

[70] Wikipedia. Spam (electronic). `https://en.wikipedia.org/wiki/Spam_(electronic)`. [Online; accessed 08-May-2013].

[71] Wikipedia. Support vector machine. `http://en.wikipedia.org/wiki/Support_vector_machine`. [Online; accessed 12-June-2013].

[72] Wong, M., and Schlitt, W. Sender policy framework (spf) for authorizing use of domains in e-mail, version 1. Tech. rep., RFC 4408, april, 2006.

[73] Xu, J., Yang, X., and Wang, L. Evaluation method of information credibility based on the trust features of web page. In *Eighth Web Information Systems and Applications Conference (WISA)* (2011), IEEE, pp. 69–72.

[74] Xu, L., Ma, Q., and Yoshikawa, M. Credibility-oriented ranking of multimedia news based on a material-opinion model. In *Web-Age Information Management*. Springer, 2011, pp. 290–301.

[75] Zhang, J., Kawai, Y., Nakajima, S., Matsumoto, Y., and Tanaka, K. Sentiment bias detection in support of news credibility judgment. In *44th Hawaii International Conference on System Sciences (HICSS)* (2011), IEEE, pp. 1–10.