Al al-Bayt University

Computer Science Department

Prince Hussein Bin Abdullah College for Information Technology

An Enhanced Classification Technique for Detecting Spam in Arabic Tweets

Student Name: Kholood Waleed Eid Olimat

Student ID: 1320901001

Supervisor:

Dr. Khaled Batiha

2017

An Enhanced Classification Technique for Detecting Spam in Arabic Tweets

By

Kholood Waleed Eid Olimat

Supervisor

Dr. Khaled Batiha

This Thesis was Submitted in partial fulfillment of the Requirement for the Master's Degree of Computer Science

Al al-Bayt University

Deanship of Graduate Studies

Al al-Bayt University

May/2017

COMMITTEE DECISION

This Thesis (Creating Large Scale of Tweet Collection for Detecting Hashtag Spam on Arabic Tweets and Finding all approval Communities) was successfully Defended and approved on / / 2017

Examination Committee Signature

Dr. Khaled Batiha, (Supervisor)

____

Dr. Akram Hamarsheh (Member)

____

Dr. Mefleh Thiabat (Member) ______________________

Dr. Ali Alawnah (Member) ______________________

(Philadelphia University)

Dedication

This thesis is dedicated to my dear father and mother, my husband, my child for their endless love, support and encouragement.

Acknoledgment

Prais to Allah, Lord of the world for this blessing, and who dose not thank people dose not thank God, so I would like to express my thanks and appreciation to the supervisor of this thesis Associate Prof. Khaled Batiha for his guidance, advice, patience, and usefull advices and how encouragement during preparing this thesis.

Thank so much and appreciation for all who supported me especially my father geology waleed olimat.

List of Contents

List of Tables

List of Figures

List of Abberviations

	Abbreviations
Twitter Spam Detection	SDA
Application Program Interface	API
World Wide Web	WWW
Search Engines	SE
Information Technologies	IT
Online Social Networks	OSN
Simple Mail Transfer Protocol	SMTP
Characterizing Automation of Twitter Spammers	CATS
Support Vector Machine	SVM
Expectation Maximization	EM
Training Dataset	TRD
Testing Dataset	TSD
Natural Language	NL
Term Frequency	TF
Inverse Document Frequency	IDF
Term Frequency-Inverse Document Frequency	TF-IDF
True Positive	TP
True Negative	TN
False Positive	FP
False Negative	FN
Classification Accuracy	Ai

List of Appendices

An Enhanced Classification Technique for Detecting Spam in Arabic Tweets

BY: Kholood Waleed Eid Olimat

Supervisor : Dr.Khaled Batiha

Abstract

The purpose of this study is to improvement a mechanism for detecting tweet spam. In this study, we have produced a dynamic method for collecting real tweets from real twitter accounts using twitter API. The system in the first step had made clustering of the tweets as spam or not spam according to is specific features, this phase allows us to create first dataset which called the training dataset. The second phase for the testing using system classifier of tweets according to text similarity using cosine algorithm. finally we produce the list of spam tweets and not spam tweets and the reason of classification as spam or not. because the evaluation of Twitter accounts and increase of Arabic tweets in few years, the researchers are portraying the twitter as one of the most important platforms to be used to apply detection techniques.

Our study focused on Arabic tweets and hashtags and it aimed for creating a larg scale of tweets collection for detecting spam hashtag on Arabic tweets using a integrated algorithm between cosine for comparing text, and stemming algorithm for text normalization process.

Key words:

Twitter,tweet, hashtag, Spam, Spam Detection,

Abstract in Arabic

الملخص

هدفت هذه الدراسة إقتراح آلية لاكتشاف التغريدات غير المرغوب بها.

في هذه الدراسة عملنا على إقتراح طريقة ديناميكية لجمع تغريدات حقيقية من حساب تويتر حقيقي باستخدام تطبيق التويتر.

يعمل النظام في الخطوة الاولى كتل للتغريدات المرغوب بها وغير المرغوب بها وفقا لميزات محددة، هذه المرحلة تسمح لنا بإنشاء أول قاعدة بيانات تسمى مجموعة بيانات التدريب، المرحلة الثانية مخصصة للفحص باستخدام تصنيف النظام على التغريدات وفقا لتشابه النص باستخدام خوارزمية ال cosine، في النهاية نزود بلائحة للتغريدات الغير مرغوب بها والتغريدات المرغوب بها والسبب في تصنيفها كمرغوب بها او لا.

ركزت الدراسة على التغريدات المكتوبة باللغة العربية وال hashtags، وتهدف لبناء مجتمعات كبيرة من التغريدات لاكتشاف ال hashtag الغيرمرغوب بها في التغريدات العربية باستخدام خوارزمية هجينة بين خوارزمية ال cosine لمقارنة النصوص والخوارزمية الجذعية لعملية تطبيع النص.

Chapter 1
Introduction

1.1General Overview

Twitter is among the fastest-growing microblogging and online social networking services (Atefeh and khrich, 2013). Twitter allows users to make tweets with 140 character messages which may be embedded with (Uniform Resource Locator) URLs (with the help of URL shortening services). Twitter's wide reach has also attracted spammers looking to mint financial gains through easy access to millions of users (Amleshwaram, Reddy, et al, 2013). A major problem in detecting spam stems from active adversarial efforts to thwart classification (Gomes, Castro, Almedia and Almeid, 2005), because of the negative effects of spam on twitter community. In our research have taked the first step to create a large scale of tweets collection for detecting spams hashtag on Arabic tweet. Among all different types of spamming, we mainly focus on the identification and annotation of tweets with hashtags, because hashtags serve as channels to increase the visibility of spam tweets. The cost of hijacking hashtags is very low with the availability of trending hashtags published on many web sites including Twitter. Our research is conducted to build new technique to detect spams in Twitter.

1.2 Background

Twitter is one of the most popular Platforms and the most popular sources for disseminating news and propaganda in the Arab region (Abozinadah, Mbaziira and Jones,2015), which allow users to make post and updates for different information, it has become particularly valuable for targeted advertising and promotions.

Because the users can post different tweets from a wide range of web-enabled services, and they are increasingly using Twitter to market or promote their products, it becomes the focus of merchants, governments and even malicious spammers. Spammers are working increasingly now to create abusive accounts to distribute adult content in Arabic tweets, which is prohibited by Arabic norms and cultures in most Arabic countries.

For example, when we search for specific topic or Hashtag, or when try to read the new updates or news for specific Hashtag, some of unrelated content will appear such as:

- Advertisement tweets.
- Duplicated and auto tweets attacking the country;
- Pornography tweets;
- Normal tweets but unrelated to the topic with multiple unrelated hashtags;

Researchers have regarded Twitter as a real platform for the real world and they have conducted numerous studies and researches on a different issues including analyzing mood and sentiment of people. One of the most important researches is to detecting spammed tweets and to classify them as spams or non-spam.

Our study aims to present an algorithm to filter the Arabic tweets and Arabic Hashtags in Twitter without dealing with any type of spams group, and to reduce the risk for the spammed tweet, which is considered as a one of the most problem in social media. It aims to building cosine similarity methods for producing an automated detecting spam hashtag on Arabic tweets. The proposed method will use a similarity measure to be applied for Arabic language tweets texts. This method will include a number of steps needed for improving the resulted tweets of research.

1.3. Problem Statement

1.3.1Research phases

(i) Providing a comprehensive review for social media, spams and their techniques, and the previous techniques that used to detect Twitter spams.

(ii) Filter groups of spam tweets based on heuristic selection.

(iii) Chunking the similar group of tweets into the same cluster.

(iv) Labeling, as much as possible, group of tweets or Hashtags that are non-spam.

(v) Building a prediction algorithm in order to predict which group of tweets is spam or not.

(vi) Finding all approval communities for each hashtag.

1.3.2Research Motivation

Because Twitter is one of the most important websites in social media, and because the evolution of Twitter accounts and increase of Arabic tweets in few years, the researchers are portraying the twitter as one of the most important platforms to be used to apply spam detections techniques.

1.3.3 Research Questions

The proposed research will investigate the Twitter Spam Detection (SDA) for Arabic tweet and will attempt to answer the following questions:

- Is the cosine similarity measure efficient to evaluate SDA?
- Is the proposed method of automated result efficient to get accurate results?
- Is the proposed approach enough and more accurate compared to already existing spam detection approaches?

1.1.1 Research Significance

This research adds a significant contribution by presenting Arabic tweet hashtags based on cosine algorithm. The proposed approach combines between Arabic tweets and pre-processing on Arabic text for detecting and classify the spam tweets and non-spam tweets and to find all expected and approval communities. The expected outcomes of this research is to reach more flexible SDA system when measuring the similarity between training dataset and testing dataset based on the Arabic text pre-processing and cosine similarity measure.

1.3.4 Research Limitation

Our proposed approach is focused on a twitter Application Program Interface (API) extraction based on Arabic text hashtags for detecting all spam tweets. Also, the proposed approach assumes to find the approval tweets (non-spam) to find all approval communities.

Chapter 2
Literature Review

2.1 Social Media

Long time ago, there were huge advancements in Telecommunications, Computer Technology, and Electronics. Particularly, advancements were made in the analysis, storage and retrieval of vast amounts of data have been happening at an exponential rate. This, in turn, has prompted the development of Database Technology that has permitted organizations to gather extremely helpful data on users and their purchasing conduct (Kaplan and Haenlein,2010).

With the advancement of expanding accessibility of Network Bandwidth and Internet Technology, Social Media has developed into multiplication in recent years. Social Media can display in various structures and has a few sorts. Social media dramatically changes our life, clients can interface with each different as well as can make and share content. Accordingly, Social Media Websites have produced a huge volume of information, which contains a considerable amount of helpful data and learning. In this way, social media analysis has turned into a basic issue for both academia and industry.

Kaplan and Haenlein (2010) characterized Social Media as a gathering of Internet-based applications that work with respect to the development of Web 2.0, and that permit the creation and trade of user- produced content. Kietzmann, Hermkens (2011). presented a structure that characterizes social media using so as to network seven utilitarian building blocks: personality, vicinity, connections, groups, discussions, sharing, and reputations.

Schrape (2011) examined the relations between digital Social Media and mass media from a frameworks hypothetical viewpoint, and the examinations led to the conclusion that social media and mass media are arranged on reciprocal levels of publicness. Yet social media is to a great extent unique in relation to traditional media in a few perspectives. The user generated content distinguishes it from the content created by professional journalists, broadcasters or other paid content providers. The user relationships, data dispersal and impact are additionally diverse.

Social Networking Services can be characterized as: Web-based administrations that permit people to:

(1) Develop an open or semi-open profile inside of a bounded system.

(2) Explain a rundown of different users with whom they share an association.

(3) View and cross their rundown of associations and those made by others inside of the framework. The nature and terminology of these associations might differ from site to site.

Social media contains video, content, pictures, sound, and so on. Discussions, blogs, microblog, wikis, interpersonal organizations, podcasts and content groups are basic social media frames. Kaplan and Haenlein (2010) depend on an arrangement of speculations in the field of media exploration (social vicinity, media extravagance) and social procedures (self-presentation, self-exposure), the two keys components of Social Media, separate social media into six sorts:

communitarian ventures, long range interpersonal communication destinations, blogs, content groups, virtual amusement universes and virtual social universes.

Weinberg and Pehlivan (2011) distinguished two variables (the depth and the half-life of data) that clarify well the variety in social media and locate the best sorts of social networking to serve different marketing targets: blogs, social networks, microblogs, and online groups.

Social advancement can be portrayed as a procedure of social change, whereby examples of human cooperation in fluence dynamic enhancements to the way of life in a social group. Since the commencement of the United Nations Millennium Development goals, the pace of global activity towards accomplishing social advancement has been moderate and conflicting. In 2014, the OECD recorded the most extensive improvement hole in 30 years. This pattern can be credited to the expanding absence of ability to address the bunch of social difficulties in the 21st century.

2.2.Spams

As more individuals depend on the abundance of data accessible online, expanded introduction on the World Wide Web (WWW) might yield significant financial picks up for people or associations. Most much of the time, Search Engines (SE) the entryways to the Web; that is the reason a few individuals attempt to deceive web indexes, so that their pages would rank high in query items, and in this way, catch user's consideration.

Pretty much as with messages, we can discuss the marvel of spamming the Web. The essential result of web spamming is that the quality of search result, as response for specific query, is decreased (Banday and Qadri,2006).

Some Web Sites contain just a couple lines of valuable data (predominantly some term definitions, likely replicated from a genuine word reference), however comprises of a large number of pages, each repeating the same substance and indicating many different pages. Every one of the pages were most likely made to help the rankings of some others, and none of them is by all accounts especially valuable for anybody searching for drug stores affiliated with Kaiser-Permanente. The auxiliary outcome of spamming is that web crawler lists are inflated with futile pages, expanding the cost of each processed query.

To give minimal cost, quality administrations, it is basic for internet SE to address web spam. However to the extent we know, despite everything they do not have a completely effective arrangement of devices for battling it. The first venture in detecting spam is understanding it that is, examining the systems the spammers use to misdirect SEs. A legitimate comprehension of spamming can then guide the improvement of fitting countermeasures.

Spam alludes to spontaneous bulk email. These messages are utilized to promote items and services for phishing goals or to drive recipients to compromised sites with the data or money theft.

These have progressed significantly since their first incarnation as text strings. At the outset, these were for the most part innocuous to security-conscious recipients however now keep on detect threats, as these have gotten to be focused on and, thus, more dangerous.

The platform of Web 2.0 keep on posturing unlimited conceivable outcomes for online communications, which tragically likewise mean more roads for cybercrime. Spam perseveres even in the platform of Web 2.0. Indeed, these messages remain an incredible disturbance for Internet users. Recent study reported that these messages cost European organizations an expected 2.8 US Dollar billion worth of profitability misfortune while U.S. based organizations reported lost 20 billion US Dollar (Banday and Qadri,2006).

The global spam volume in the main portion of 2011 demonstrates the shifting number of spam caught on a week by week premise. Spam has demonstrated expensive as far as resources, for instance, storage, server capacity, network infrastructure, bandwidth, and. These cost an organization technical expenses because of things like the amount of power required in email server processing important to handle their downpour and the measure of time IT bolster staff need to spend to deal with the problem.

We used the term spamdexing (also, spamming) to allude to any planned human activity that is intended to trigger an unjustifiably favorable pertinence or significance for some web pages, considering the page's real (true) value. We have utilized the modifier spam to check every one of those web objects (page contents and links) that are the consequence of some type of spamming. Individuals who perform spamming are called spammers.

2.3.Spamming Techniques

In this sub-section we have summarized spamming strategies that influence the ranking algorithm utilized via search engines (Gyongyi and Molina, 2014).

Term Spamming:

In assessing textual relevance, SE consider where on a website page query terms happens. Every kind of location is known as a field. The common text fields for a page P are the document title, body, the Meta tags in the Hyper Text Mark-up Language(HTML) header, and page P's URL.

Moreover, the anchor texts connected with URLs that indicate P are additionally considered fitting in with page P, since they regularly depict extremely well the substance of P. The terms in P's content fields are utilized to decide the importance of P as for a specific query, frequently with different weights given to different fields. Term spamming alludes to procedures that tailor the substance of these content fields so as to make spam pages pertinent for a few query.

Link Spamming

Close to term-based relevance metrics, SEs additionally depend on connection data to decide the significance of web pages. In this manner, spammers frequently make join structures that they trust would expand the significance of one or a greater amount of their pages.

Hiding Techniques

It is normal for spammers to hide the telltale signs (e.g. list of links, and rehashed terms) of their activities. They utilize various strategies to hide their misuse from regular web users visiting spam pages, or from the editors at SEs organizations who attempt to define spam cases.

Cloaking

On the off chance that spammers can obviously distinguish web crawler, they can receive the accompanying technique, called Clocking: given a URL, spam web servers return one specific HTML record to a consistent web program, while they give back a different document to a web crawler. Along these lines, spammers can exhibit the eventually planned substance to the web users (without hints of spam on the page), and, in the meantime, send a spammed record to the search engines for indexing.

Redirection

Another method for spam on a page is via naturally redirecting the browser to another URL when the page is loaded. Along these lines the page still gets indexed by the search engines, yet the client won't ever see it—pages with redirection go about as intermediates for a definitive targets, which spammers attempt to serve to a client achieving their destinations through Internet searches. Redirection can be accomplished in various ways. A straightforward methodology is to exploit their fresh Meta-tag in the header of an HTML archive. By setting the revive time to zero and the invigorate URL to the objective page, spammers can accomplish redirection when the page gets stacked into the browser:

<Meta Http-Equiv"refresh" content= 0; "URL=target.htm">

2.4Spam on Social Media

The Online Social Networks (OSN) spam issue has officially gotten consideration from analysts. In the interim, email spam, an apparently fundamentally the same issue, has been widely concentrated on for quite a long time. Shockingly, the group of the current arrangements are not straightforwardly relevant, as a result of a progression of particular attributes relating to the OSN spam. In any OSN, all messages, including spam, begin from records enlisted at the same site (Zhu, et al, 2012).

Interestingly, email spam is not as a matter of course sent from records enlisted at true providers. The generally utilized email server notoriety based recognition approaches depend on the suspicion that the spamming Simple Mail Transfer Protocol(SMTP) servers keep running on bot machines, and are therefore inapplicable in OSNs. Understanding that this supposition is not generally genuine, researchers have proposed to distinguish accounts signed up by spammers from honest to goodness email service providers. Spamming account identification is likewise the center of the current OSN spam location work.

In any case, OSN spam is that the larger part of spam messages originates from fake accounts, as opposed to accounts made and solely controlled by spammers. It basically implies that spammers and honest to goodness clients are sharing accounts. Along these lines, recognizing spamming accounts is not sufficient to fight OSN spam (Stringhini, et al, 2010).

Messages in OSNs, spam or not, are short. The recognition that honest to goodness messages have variable size while spam has a tendency to be little no more holds in OSNs.

2.5 Detection Techniques for Spam in Twitter

A Twitter Hashtag is just a keyword phrase, illuminated without spaces, with a pound sign (#) before it. It ties the conversations of various users into one stream, which one can discover via seeking the Hashtag in Twitter Search. For illustration, #HP. We characterize hashtag hijacking as abuse of a Hashtag for the reason it is not expected to.

Hijacking Hashtag can happen through ways (McCord, Chuah, 2011):

- Attaching an injurious link.

- Attaching an irrelevant connection.

- Discussing random conversations.

This slants the client to look at whatever is left of the discussion happening around that hashtag paying little mind to the actuality whether it is in arrangement with the hashtag or not. This induces the brand contenders and spammers to hijack the trending hashtag for business addition and slander. Notwithstanding the above expressed likewise the hashtag tweets are focused on:

- To look for consideration and make once junk well known.

- To attack the matter of a specific prevalent brand where the spoilers express their assumptions in a snarky way.

• Posting injurious and sullied content on social forms by means of mainstream Hashtags.

• Posting undesirable URL's through trending Hashtags spamming the social media all things considered.

2.6 Twitter Spamming Techniques

Twitter Spamming techniques can be divided into two categories (Thomas, Grier, paxson, et al, 2011):

I. Profile-Based Spamming Techniques:

• Follow Spam: It is the demonstration of following mass number of people, not on the grounds that a client really inspired by their tweets, however essentially to pick up consideration, get perspectives of a particular users' profile, or (in a perfect world) to get followed back. Automated programs are utilized to make this undertaking less demanding; along these lines they can follow a large number of clients with in a small amount of seconds. In great cases, these automated accounts have followed such a large number of individuals and they are danger to the execution of the whole framework. In less-amazing cases, they just disturb a large number of honest to goodness clients who get a notice about this new follower just to discover their advantage may not be totally sincere. These sorts of accounts can be analyzed by checking the tweets posted by the clients and looking at their conduct.

- Mention Spam: Spammers specify the username of a focused on user before tweeting. Focused on user’s consideration can be gotten by this strategy.

II. Content-Based Spamming Techniques (Thomas, et al, 2011):

- Trend Abuse Spamming: Twitter's API likewise gives a rundown of the top trends every hour. Spammers utilize these trending topics in their tweets and it gets posted in the course of events making disturbance every one of the clients since open records can be seen by anybody on the twitter.
- Trend Setting Spamming: Here spammers post a substantial number of tweets containing a particular word in it, making the word or Hashtag another trending topic.
- Fake Re-tweets: In this procedure spammers exploit the Twitter's Re-Tweet tradition to make it give the idea that a Spammer's tweet was initially distributed by another client. These can be recognized by twitter's search where re-tweets can be recognized from original tweets.
- Embedding The Most Popular Search Topics/Terms: In this strategy spammers act exceptionally savvy. They incorporate well known search terms in their tweets and when a user searches the same terms, these tweets gets showed in the result set, which is again an irritating background for a true blue client, who does not get the normal results.
- Direct Message: This is customary spamming system where spammers send individual message to another client.

2.7 Previous Studies

Mazzia and Juett . gave the issue of how to successfully arrange and search for posts. Taking a gander at Twitter, they saw that clients might order their posts utilizing Hashtags, and any word or expression might be utilized as the classification. Endeavoring to scan for tweets about Facebook, a client would need to attempt a wide range of Hashtags, as #FB, #Facebook.com, #Facebook, or #Zuckerberg. To battle this, they proposed, implemented and assessed a tool for recommending relevant Hashtags to a client, given a tweet. Starting analysis propose dataset is sufficiently rich to remove informative distributions of words for some Hashtags that will encourage an Naive Bayes model for Hashtag suggestion given a query post (Allie, James,).

El-Mawass and Alaboodi (2015). broke down of spam content on Arabic Trending Hashtags in results in an appraisal of around three quarter s of the aggregate generated content. This disturbing rate makes the improvement of adaptive spam detection strategies an undeniable and pressing need. They analyzed the spam content of trending Hashtags on Twitter, and evaluate the execution of past spam detection systems on accumulated dataset. Because of the raising control that describes more up to date spamming accounts, straight forward manual labeling at present prompts erroneous results. With a specific end goal to get reliable and trusted ground-truth information, they proposed an upgraded manual classification procedure that maintains a strategic distance from the deficiencies of more seasoned manual methodologies. They additionally adjusted the already proposed components to react to spammers avoiding strategies, and utilize these elements to build a new data-driven detection system (El-Mawass and Alaboodi, 2015).

Guo and Chen (2014) added to a methodological framework to: (1) extract client attributes taking into account geographic, graph based and content-based elements of tweets; (2) build a dataset by manually examining and marking a vast sample of twitter users; and (3) infer reliable guidelines for detecting non-personal users (automateds) with supervised classification strategies. The separated geographic qualities of a client incorporate most extreme rate, mean speed, the quantity of various regions that the client has been to, and others. Content-based qualities for a user incorporate the quantity of tweets every month, the rate of tweets with URLs or Hashtags, and the rate of tweets with emotions, distinguished with sentiment analysis. The proposed framework consists of the following steps (Guo and Chen , 2014):

1. Extract client attributes taking into account the geographic, graph- based and content data in tweets;
2. Construct training datasets by physically inspecting tweets and labeling an extensive example of twitter clients.
3. Conduct supervised classification and infer principles and learning for identifying non-personal users; and
4. Assess the inferred rules with new manual inspection and training data.

Chu, Widjaja and Wang (2012). misused the aggregate recognition way to deal with catching spams with multiple accounts. Their work utilizes the features consolidating both substance and conduct to recognize spam campaigns from honest to goodness ones, and fabricate an automatic classification framework. Moreover, their work can be connected to other informal organizations by incorporating application-specific highlights (Chu, Widjaja and Wang, 2012).

McCord M. and Chauh M (2011). examined some content-based and user-based components that are distinctive in the middle of spammers and genuine users. At that point, they utilized these components to encourage spam recognition. Utilizing the API methods gave by Twitter, they crawled active Twitter users, their followers/following data and their latest 100 tweets. At that point, they dissected the gathered dataset and assessed our discovery plan in view of the recommended client and substance based elements. The outcomes demonstrate that among the four classifiers they assessed, the Random Forest Classifier delivers the best results. The outcomes in view of the 100 latest tweets likewise demonstrate that spam detection taking into account their recommended components can accomplish 95.7% precision and 95.7% F measure utilizing the Random Forest Classifier (McCord and Chuah, 2011).

Amleshwaram et al (2013). proposed a new characteristic to identify spam from legitimate accounts. The characteristics dissect the behavioral entropy as well as content entropy, bait- methods, and profile vectors describing spammers, which are then bolstered into supervised learning algorithms to produce models for their tool, Characterizing Autommation of Twitter Spammers (CATS). Utilizing the proposed framework on two Twitter data sets, they watched a 96% percent of detection rate with around 0.8% percent of false positive rate beating detection technique (Amleshwaram, et al, 2013).

Meda , BiSIO , Gastaldo. et al, (2014) bring up the application of three algorithms related to machine learning, concentrating on the execution performance of these algorithms, with a specific end goal to distinguish the best algorithm and the best parameters that join both satisfactory detection results and impressive performance level. Test results confirm the effectiveness of using Random Forest Algorithms contrasted with the Support Vector Machine (SVM) and the Extreme Learning Machines: the performance of Random Forest increase with the minimizing number of characteristics contradicted to the next two method. This conduct underlines the favorable position to pick few characteristics in the interest of computational cost as well as detection cost (Claudia, et al, 2014).

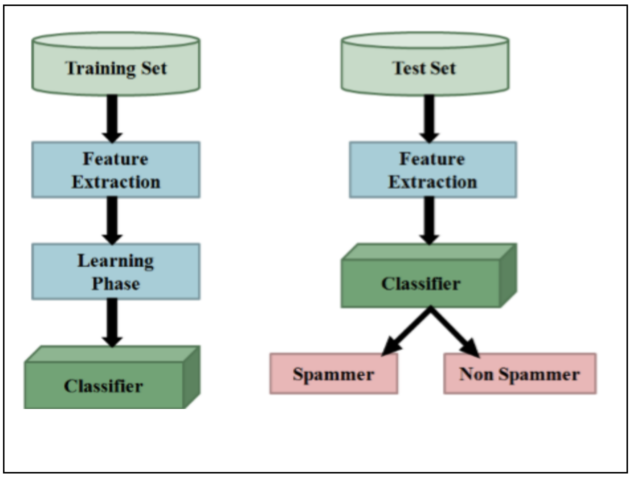


Figure 2.1: ML System (Claudia, et al, 2014).

Sedhai and Sun (2015) gathered 14 million tweets that coordinated some trending Hashtags in two months and after that directed systematic annotation of the tweets being ham (i.e., non-spam) and tweets being spam. They name the dataset HSpam14. Their explanation process incorporates four noteworthy steps (Surendra and Aixin, 2015):

(i) heuristic-based selection to look for tweets that will probably be spam.

(ii) near-duplicate cluster based comment to firstly amass comparative tweets into groups and afterward name the groups.

(iii) reliable ham tweets identification to label tweets that are non-spam.

(iv) Expectation-Maximization (EM) based label to predict the labels of staying unlabeled tweets.

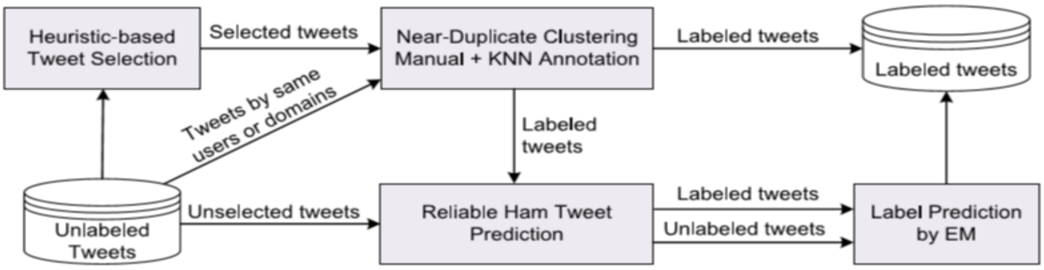


Figure 2.2: HSpam14 framework (Surendra and Aixin, 2015).

Miller, Dickinson, Deitrick, et al (2014) have made three new contribution to the field spam of spam detection on twitter. First they viewed spam identification as an anomaly detection problem. Secondly they introduce 95 one-gram features from tweet text to the task of spam detection in twitter. Finally they used the stream of real-time tweets as well as user profile information with two stream-based clustering algorithms, they used DenStream and StreamKM++ and they found that these algorithms independently demonstrated good detection, the combination of the two further improved all there metrics particularly recall and false positive rate .

Al-smadi, Jaradat, Al-Ayyoub, et al (2017). They proposed approach employs a set of extracted features based on lexical, syntactic, and semantic computation, the approach used word alignment features to detect the level of similarity between tweets pairs. The best achieved results in both tasks is when using the lexical overlap features with the word alignment and topic modeling features.

Chapter 3
Methodology

3.1Introduction

This section presents the methodology that we will use for creating a large scale of tweets collection for detecting spam hashtags on Arabic tweets and then finding the approval communities. The first step in the proposed methodology is to collect both of training dataset (TRD) and the testing dataset (TSD) directly form twitter API. The second phase is to convert both of TRD and TSD into separated keywords, each of which could have a preprocessing; such as the tokenization, stop word checking and delete suffix and prefix if found. Next phase is for system classifier; in this phase we will convert each word to vectors.

3.2.Overall Research Design

Figure 3.1 shows the overall research design which represents the general overview of framework which starting from Twitter API connection to detection and filtering spams and approval tweets.

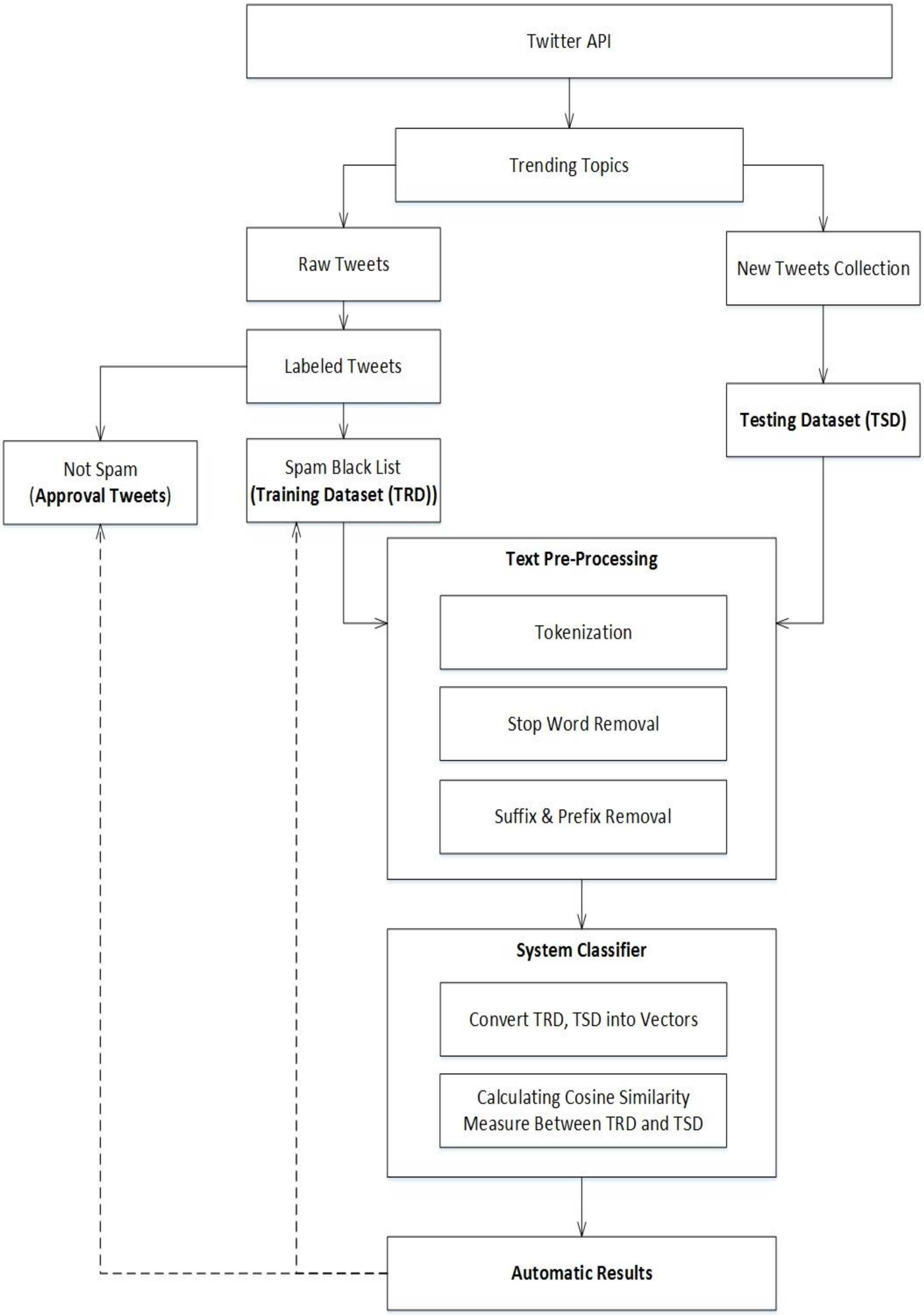


Figure 3.1: General framework

3.3 Data Sample

In our research, we have used real dataset from twitter using REST APIs, which it provides programmatic access to read and write Twitter data. Author a new Tweet, read author profile and follower data, and more. The REST API identifies Twitter applications and users using OAuth. In our research we have four variant datasets:

- Training Dataset: This dataset is the spam tweets collection, which we got it after spam labeling process or after using system classifier, all spam detection will be located in training dataset.
- Testing Dataset: Tweets collection which we need to classify it using system classifier, we will get it from twitter API directly.
- Approval Tweets: All tweets which are not classified as spam in labeling phase or in system classifier phase.
- Arabic Stop Words: We have used a sub list of El-Khair (2006) stop list that contains 1,377 words of Arabic stop words (El-Khair, 2006).

3.4 Research Phases

For reach to the best filtering for spams on tweet and testing this we must pass the following steps:

3.4.1 REST API

REST API allow us to access a group’s Tweets by using statuses/group timeline. The API returns a Tweet as shown in Figure 3.2.

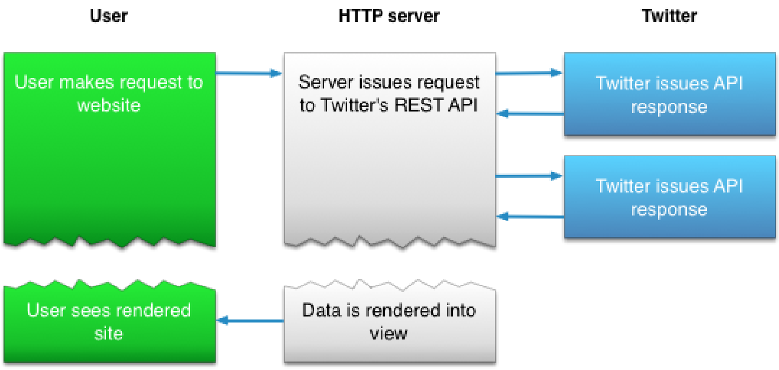

User
HTTP server
Twitter
User makes request to website
Server issues request to Twitter's REST API
Twitter issues API response
Twitter issues API response
User sees rendered site
Data is rendered into view

Figure3. 2: REST API

Before we get started in this level, we need to create an application key, because of that we need also a Twitter account. The steps below discuss the all required steps to make configuration for our API:

1. Create new Twitter accounts.
2. Create an API key for our application https://dev.twitter.com/apps (we should fill all requirements to get successful application, such as: **Name, Description, Website, Callback URL**).
3. After creating our application, we can access what we need to authenticate to twitter using OAuth, namely (Consumer key, Consumer secret, Access token, Access token secret)
4. Now we can connect with twitter API.

3.4.2 Collection of Trending Topics

The proposed model gathers a group of tweets and group of user accounts which related to a specific trending topic. The trending topics are periodically retrieved in order to gather a heterogeneous set of available tweets and user accounts. In our research, the trending topics are related about Arabic hashtags, so we select specific hash tags concerned with such as:

- Public opinion issues and polls.
- Topics related to health organizations.
- Contemporary issues.
- Tourism topics.

3.4.3 Text Preprocessing

The first step is to define a set of training dataset (TRD) and testing dataset (TSD), as input for the system. Text preprocessing is vital for successful automatic results for detecting processes because every word or term in the tweet does not have the same significance. This may affect the term's selection and indexing. Therefore, preprocessing helps attain an efficient utilization of resources.

Some terms play a more crucial role in tweets than other terms. For example, if some terms have a higher frequency than others in the tweets, this does not mean that they are more important, or be used as index terms in automatic detection results. Probably, these words seem to be stopping words, which will be discussed in this section. In this work, text preprocessing methods needed are the following:

- Tokenization.
- Stop words Removal.
- Normalization.

3.4.3.1 Segmentation and Tokenization

Sentence segmentation is considered as a problem in identifying the boundaries of sentences. Arabic sentences are segmented using punctuation marks that define the end of each sentence. A set of punctuation marks, including commas (،), semicolons (؛), question marks (؟), exclamation marks (!), colons (:), and periods (.) are selected to split the text into sentences.

Tokenization is breaking the tweet text into sets of words or phrases or other meaningful elements called tokens. These token will be used later in some processes like text mining or segmentation. Normally, tokenization occurs at a word level and white-spaces are used to separate the tokens. The tokens are then used as inputs to represent the training dataset and testing dataset.

3.4.3.2 Stop Words Removal

Stop words are the words that occur too frequently in tweets and do not carry any meaning in the natural language processing. These words are not useful in the detection and classification process, and they are not used as index terms. An example of stop words is prepositions and conjunctions. Filtering out stop words will improve performance because of fewer terms in our dictionary and more relevant search results. In our work, we will use a sub list of El-Khair (2006) stop list that contains 1,377 words. This list is listed in Appendix A.

3.4.3.3 Normalization

Normalization means modifying the text to make it consistent by removing unneeded characters of suffix and prefix, such as removing non-alphanumeric characters or diacritical marks, removing unacceptable repeated characters and normalization of some Arabic letters such as the normalization of (أ) or (إ) in all forms to (ا), as regular expressions would suffice (Alsaleem, 2011). This leads to the perspective that normalization process will increase the performance of the work (Alsaleem, 2011).

In this phase we will use the stemming algorithm to remove suffixes and prefixes that are added to the word (Finn and Portrait, 2011). Figure 3.3 shows how stemming algorithm working for deleting the suffix and prefix:

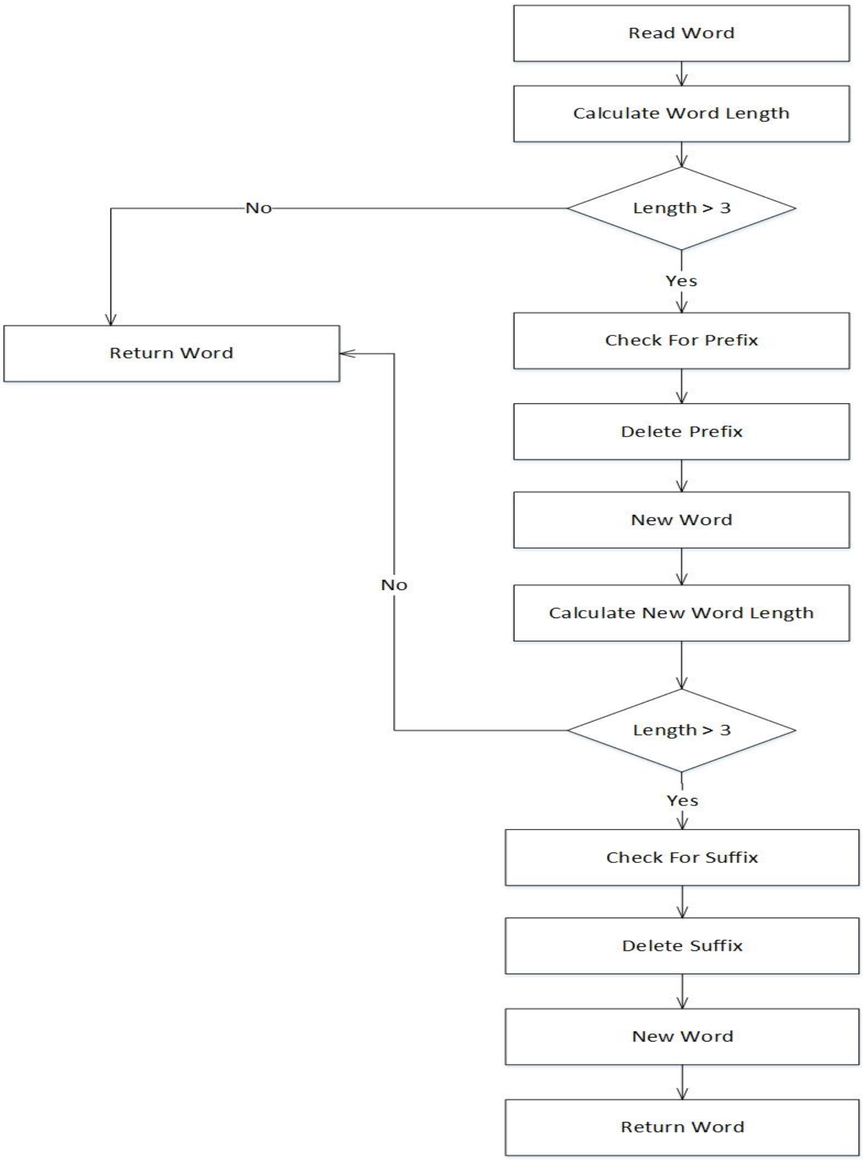


Figure 3.3: Prefix and Suffix Removal

3.3.4 Spam Labeling

The second stage of the proposed system is the labeling of spams that related to trending topics, where the proposed system employs different blacklists in order to detect group of spam in tweets and label the entire collection in this manner. The labeled tweets set acquired in this stage will be used in order to train the system and to be used in detecting any new group of spam tweets. This stage is considered as pre-classified stage into group spam of tweets and group non-spam of tweets.

By using Twitter API, we built a crawler to gather trending topics and their related group of tweets. For our dataset, we took into our considerations that it may still has some bias group spam tweets. However, even for that set of group spam tweets, we will use them in order to evaluate the performance of our proposed system on detecting these groups of spam tweets.

3.3.5 Feature Extraction

A features extraction process is used to represent each labeled group of tweets by using Natural Language (NL) processing and by using content analysis methods. Then, the final dataset, containing the labeled set of group tweets and each group tweet totally represented by a set of specific features, will be used by the classifier in order to train the model and get significant knowledge in the detection of groups of spam tweet.

3.3.6 Spam Detection

The classification algorithm gets a group of tweets from a user as input and then it will notify him whether this tweet is spam or not. In turn, the user can directly inform the system of a possible wrong in the classification. In this case, the system's admin considers this error and decides to alter the dataset in.

Similarity between testing dataset (TSD) and training dataset (TRD) can be measured by using text-to-text similarity methods. This makes it clear that it is mainly dependent on the comparison of the text tweets between the testing dataset (TSD) and training dataset (TRD) using several methods, covering string-based similarity. String similarity measures operate on string sequences and character composition in order to judge the similarity between two text strings.

Each tweets in TSD and TRD will be represented as vectors, where each tweet in both of TSD and TRD are a set of terms; each term has a weight which reflects its importance on that tweets in testing dataset or training dataset. There are several ways to calculate this weight, such as the Term Frequency-Inverse Document Frequency (TF-IDF), where the Term Frequency (TF) refers to the term frequency in the training dataset, and the Inverse Document Frequency (IDF) represents the importance of a term with respect to the entire corpus. It is calculated by the number of tweets in the corpus divided by the number of tweets containing a term.

The formulas of TF, IDF and TF-IDF are illustrated bellow as follows (Dongen and Enright, 2012):

$$\text{TF} = \frac{number\ of\ occurrence\ of\ term\ in\ tweet}{number\ of\ term\ in\ tweet}$$

IDF = log (1 +$\frac{N}{nj}$)

Where N is the total number of tweets, n_j is the number of tweets containing the term.

TFIDF = TF * IDF

The main idea behind this model is to calculate the weight of each term in each tweet with respect to the entire corpus. First, the model calculates the TF of each term in a tweet. A high TF value indicates that this term will play a crucial role in that tweet. Second, IDF is calculated based on (Equation 2). Finally, TF is multiplied by IDF to get TF-IDF.

moreover, cosine-similarity measure will be used in this study. The TF-IDF is used to compare a tweets in testing dataset vector with a training answer vector using cosine similarity measure. The cosine measure is a function that has proved reliable in decades of experimentation (Dongen and Enright, 2012). Cosine similarity measures the cosine of the angle between two vectors. This shows that the value of cosine similarity is bounded by the interval [0, 1]. In fact, this measure has been used in information retrieval and text mining.

Two vectors of attributes, training dataset (TRD) and testing dataset (TSD), the cosine similarity and cos (θ) (Van Dongen, and Enright, 2012) are represented by using a dot product and magnitude as follows:

$$cosine\ similarity(D.Query) = \frac{Dot\ product(TRD.TSD)}{||TRD|| * ||TSD||} \qquad (4)$$

Where the dot-product is:

$$Dot\ product(TRD.TSD) = TRD[0] * TSD[0] + \cdots + TRD[n] * TSD[n] \qquad (5)$$

And distant, $\left||\mathrm{TRD}|\right|$ and $\left||TSD|\right|$ is defined as:

$$\left||\mathrm{TRD}|\right| = \sqrt[2]{\mathrm{TRD}[0]^2 + \mathrm{TRD}[1]^2 + \cdots + \mathrm{TRD}[\mathrm{n}]^2} \qquad (6)$$

$$\left||TSD|\right| = \sqrt[2]{TSD[0]^2 + TSD[1]^2 + \cdots + TSD[n]^2} \qquad (6)$$

For text matching, the vectors TRD and TSD are usually the term frequency vectors of the tweets.

After the cosine similarity between the training dataset and testing dataset are calculated, resulted values are assigned. If the result of cosine similarity is 1, the tweet will be classified as spam. But if the cosine similarity is 0, then the tweet will be classified as not spam. Otherwise, the result values will be between one and zero.

We have adopted a dynamic method to calculate the classification for each tweet, this method depends on the average of cosine values for all previous tweets which it measured by system classifier, if cosine value for specific tweet is greater than cosine limit (average) or Threashold, then the system classifier will classify this tweet as spam, but if cosine value is less than cosine limit, then the system classifier will classify this tweet as not spam (approval tweet). Table below show an example for classification method.

Table 3. 1: Classification method for spam and not spam tweets.

Tweet ID	Cosine Value	Cosine Limit (Threashold)	Classify
1	0	0.0228595	Not Spam
2	0.05	0.00849401	Spam
3	0.117647	0.0128342	Spam
4	0.142857	0.192327	Not Spam
5	0.15	0.00346117	Spam

3.5 Research Tools

In our work we will use PHP programming language as main language and MySQL for database to apply both of training and testing dataset and word list.

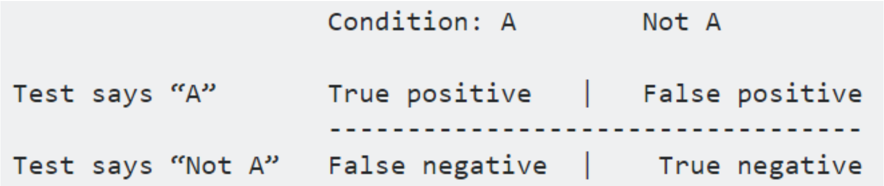
Chapter 4
Results and Analysis

To evaluate the performance of proposed classifier we collect several tweets for different trending Arabic hashtags.

We ran the classifier for each retrieved tweet for specific hashtag. The number of the true classification (i.e. the actual class is spam and it is classified as spam by our proposed classifier, or the actual class is not spam and it is classified as not spam by our proposed classifier).

The most common performance metrics consider the classifier ability to recognize one specific class versus all others. The class of interest is known as the positive class, while all others are known as negative.

The use of the terms positive and negative is not intended to imply any value judgment (that is, good versus bad), nor does it necessarily suggest that the outcome is present or absent (such as birth defect versus none). The choice of the positive outcome can even be arbitrary, as in cases where a model is predicting categories such as sunny versus rainy or dog versus cat.



```
                    Condition: A         Not A

Test says “A”       True positive   |   False positive
                    ----------------------------------
Test says “Not A”   False negative  |    True negative
```

Figure 4.1: Confusion metrics of TP, FP, FN, and TN.

The relationship between the positive class and negative class predictions can be depicted as a 2 x 2 confusion matrix that tabulates whether predictions fall into one of the four categories:

- True Positive (TP): Correctly classified as the class of interest.
- True Negative (TN): Correctly classified as not the class of interest.
- False Positive (FP): Incorrectly classified as the class of interest.
- False Negative (FN): Incorrectly classified as not the class of interest.

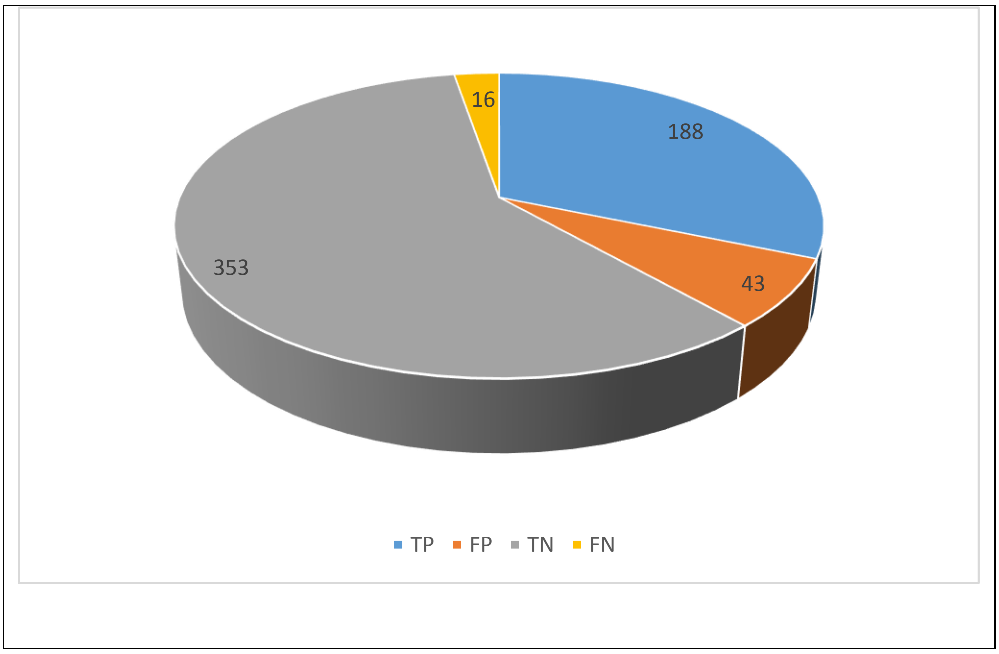


Figure 4.2: The overall TP, FP, TN, and FN for all retrieved tweets (Proposed classifier).

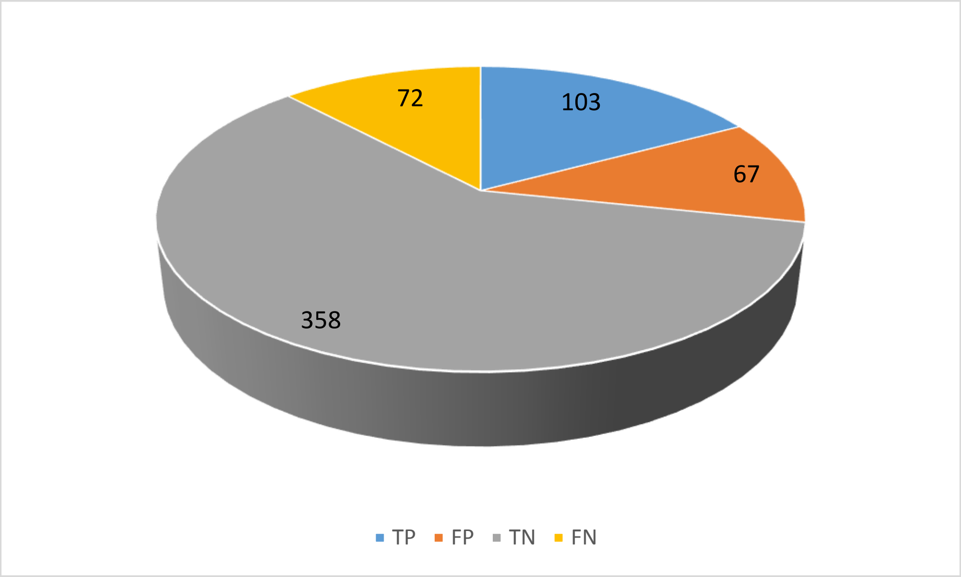


Figure 4.3: The overall TP, FP, TN, and FN for all retrieved tweets (KNN classifier).

To evaluate our proposed classifier, we used different evaluation metrics, such as (Atefeh and Khrich, 2013):

- TP rate (sensitivity) measures the proportion of positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition). It is measured as:

$$\text{TP rate} = TP/(TP+FN)$$

- TN rate (Specificity) measures the proportion of negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition). It is measured as:

$$\text{TN rate} = TN/(FP+TN)$$

- Precision (or Positive predictive value) is the proportion of predicted positives which are actual positive. It is measured as:

Precision= TP/(TP+TN)

- Recalls the proportion of actual positives which are predicted positive.

Recall= TP/(TP+FN)

- Classification Accuracy (A_i) of an individual program i depends on the number of samples correctly classified (true positives plus true negatives) and is evaluated by the formula:

$$A_i = \frac{t}{n} \cdot 100$$

Where t is the number of sample cases correctly classified, and n is the total number of sample cases.

- F-measure based on precision and recall values

F= 2*Precision *Recall / (Precision + Recall)

Table 4.1: Precision for proposed classifier and KNN classifier

	Precision for proposed classifier	Precision for KNN classifier
5 tweets - 50 hashtags	63%	86%
5 tweets - 100 hashtags	53%	88%
10 tweets - 50 hashtags	76%	83%
10 tweets - 100 hashtags	71%	91%
15 tweets - 50 hashtags	75%	85%
15 tweets - 100 hashtags	99%	87%
Average	81%	86%

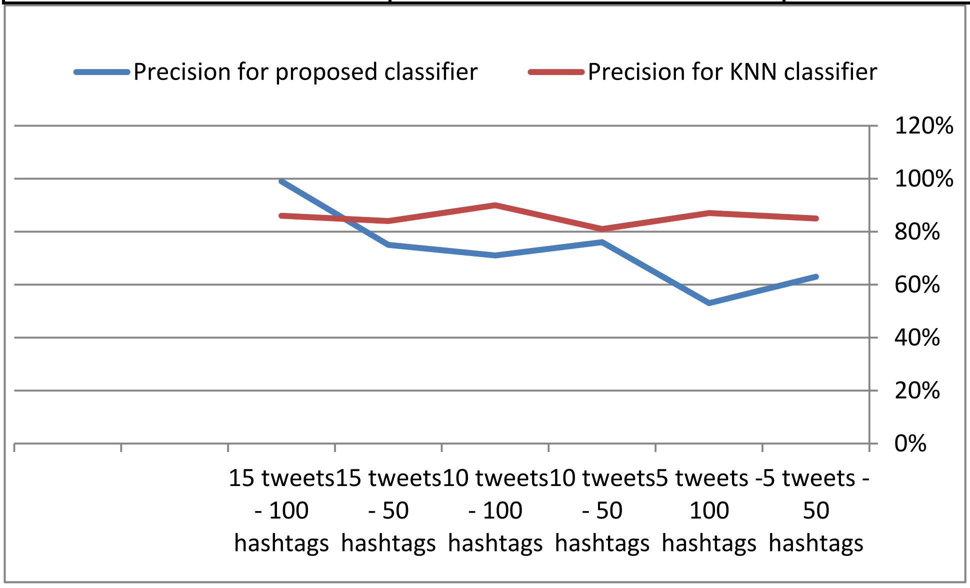


Figure 4.4 : Precision for proposed classifier and KNN classifier

In precision measure the average result in our proposed classifier was 81% while it was in KNN classifier 86%.

Where KNN classifier achieved result better than our proposed classifier result because of multiple stages which tweets passed in before classifying process.

Table 4.2: recall for proposed classifier and KNN classifier

	Recall for proposed classifier	Recall for KNN classifier
5 tweets - 50 hashtags	63%	85%
5 tweets - 100 hashtags	91%	87%
10 tweets - 50 hashtags	95%	81%
10 tweets - 100 hashtags	100%	90%
15 tweets - 50 hashtags	92%	84%
15 tweets - 100 hashtags	92%	86%
Average	92%	86%

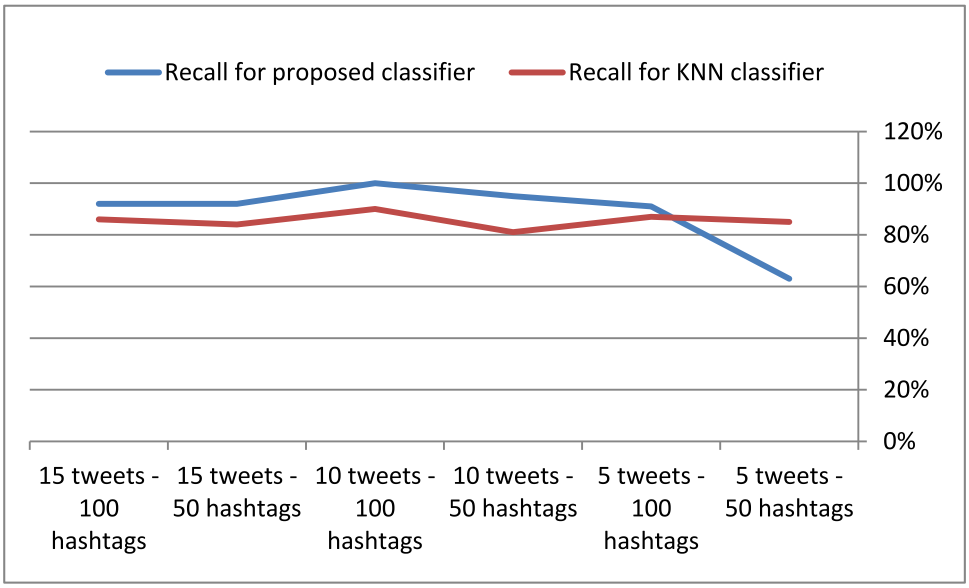


Figure 4.5: recall for proposed classifier and KNN classifier

In recall measure the average result in our proposed classifier was 92% while it was in KNN classifier 86%.

Our proposed classifier achieve result better than KNN classifier result because of recall focuses on dose the algorithm returned topics more related to the search process, and as we know whenever recall result increase it will be much better, whereas recall talk about (True positive rate).

Table 4.3: Accuracy for proposed classifier and KNN classifier

	Accuracy for proposed classifier	Accuracy for KNN classifier
5 tweets - 50 hashtags	94%	85%
5 tweets - 100 hashtags	90%	87%
10 tweets - 50 hashtags	93%	81%
10 tweets - 100 hashtags	87%	90%
15 tweets - 50 hashtags	86%	84%
15 tweets - 100 hashtags	91%	86%
Average	90%	86%

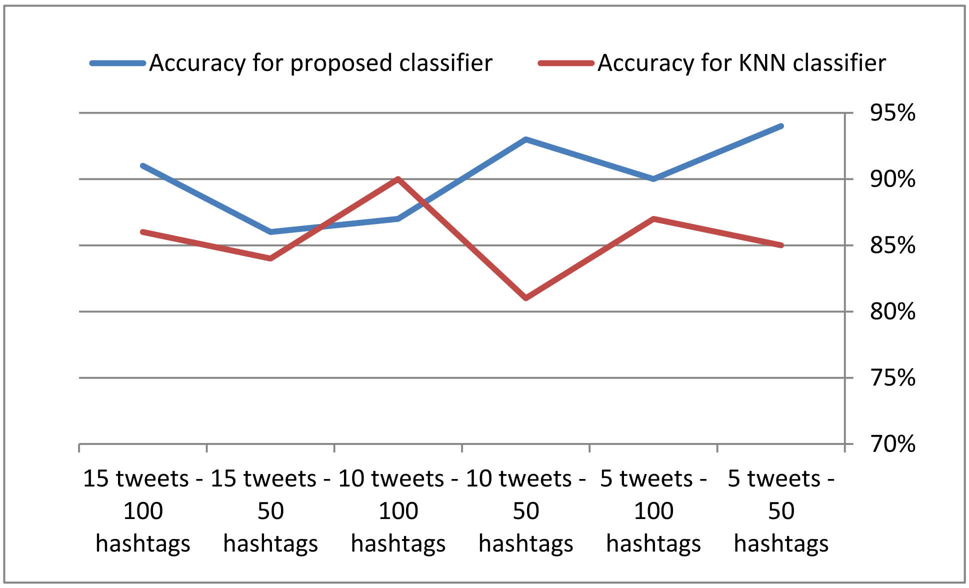


Figure 4.6: Accuracy for proposed classifier and KNN classifier

In Accuracy measure the average result in our proposed classifier was 90% while it was in KNN classifier 86%.

Whereas our proposed classifier achieved result better than KNN classifier result because of multiple stages which tweets passed in pre processing before classifying process.

Table 4.4: F-measure for proposed classifier and KNN classifier

	f-measure for proposed classifier	f-measure for KNN classifier
5 tweets - 50 hashtags	62%	85%
5 tweets - 100 hashtags	67%	87%
10 tweets - 50 hashtags	84%	81%
10 tweets - 100 hashtags	83%	90%
15 tweets - 50 hashtags	83%	84%
15 tweets - 100 hashtags	95%	86%
Average	79%	86%

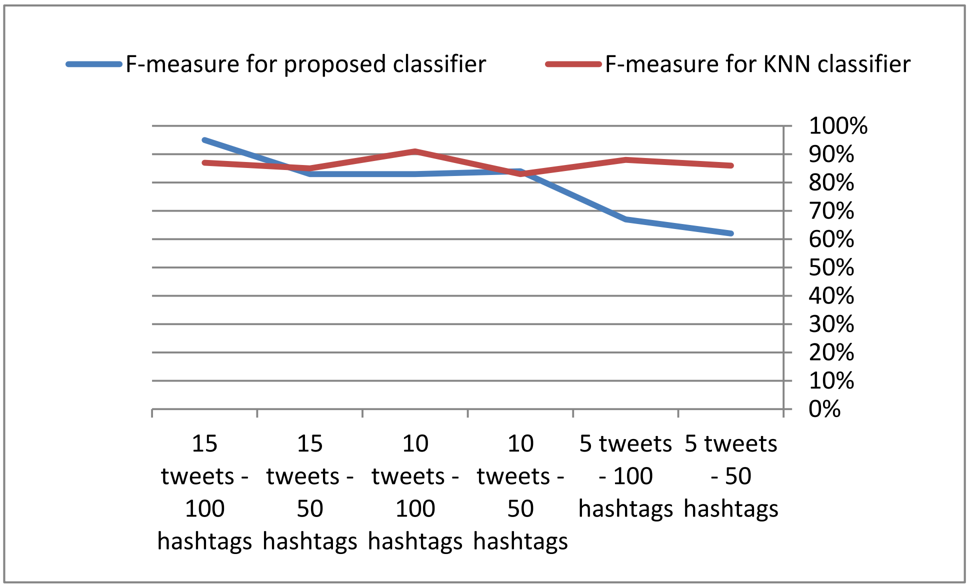


Figure 4.7 : F-measure for proposed classifier and KNN classifier

In F- measure the average result in our proposed classifier was 79% while it was in KNN classifier 86%.

KNN classifier achieve result better than our proposed classifier result because of precision result affects on F-measure result since it depends on precision and recall results.

Conclusions

Spam detection is a fundamental task in the social media requirements, because spam's make user anxiety and discomfort. At the same time, many of researcher are working on spam detections with different algorithms, our study focused on Arabic tweets and hashtags and it aims for creating a larg scale of tweet collection for detecting hashtag spam on Arabic tweets using a hybrid algorithm between cosine for comparing text, and stemming algorithm for text normalization process.

Our proposed classifier overpowered on KNN classifier in recall and accuracy results where our recall result is 92% and KNN recall result is 86%, and our accuracy result is 90% where KNN accuracy result is 86%.

Future Works

It is hoped that this platform can be assessed in the future by working on:

1.Improving the scalability of the system so as to benefit from newer technologies in this field.

2.Using new libraries that support a specific ontology to analyze texts more efficiently.

3.Apply our methodology for English text language.

4. compare our classifier to more classifiers.

References:

[1] Abozinadah E, Mbaziira A, Jones J, (2015), **Detection of Abusive Accounts with Arabic Tweets,** International Journal of Knowledge Engineering, Volume. 1,Number. 2.

[2] Abu-Errub A, Odeh A, Shambour Q and Al-Haj Hassan O, (2014**), Arabic Roots Extraction Using Morphological Analysis,** IJCSI International Journal of Computer Science Issues, Volume. 11, Number. 1.

[3] Alotaibi M, (2013), **The Impact of Twitter on Saudi Banking Sectors in the Presence Of Social Media: An Evaluative Study,** International Research: Journal of Library & Information Science, Volume. 3, Number. 4.

[4] Alsaleem, S (2011**), Automated Arabic Text Categorization Using SVM and NB,** International Arab Journal of e-Technology, Volume. 2, Number. 2.

[5] Al-Smadi M, Jaradat Z, Al-Ayyoub M, Jararweh Y (2017), **Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features,** Elsevier journal.

[6] Amleshwaram A, Reddy N, Yadav S, Gu G, Yang C,(2013), **CATS: Characterizing Automation of Twitter Spammers.**

[7] Amleshwaram A, Reddy N, Yadav S, Gu G and Yang C, (2013), **CATS: Characterizing Automation of Twitter Spammers ,** IEEE 978-1-4673-5494-

[8] Arun Kuma, (2012), **Twitter Spamming: Techniques and Defense Approaches,** International Journal of Applied Engineering Research, Volume. 7, Number. 11.

[9] Atefeh .F, W. Khrcich,"(2013), **A Survey of Technologies for Event Detection in Twitter",** International Journal of Computational Intelligence**,** volume. 0, Number. 0.

[10] Banday M, and Qadri J, (2006), **Spam: Legal and Technical Aspects,** Department of Low University of Kashmir Hazratble, Srinagar, India, Volume. XIII, Number. XIII.

[11] Boyd D., (2007), **Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life,** Cambridge MA: MIT Press, Volume (ed. David Buckingham).

[12] Chu Z, Widjaja I and Wang H, (2012), **Detecting Social Spam Campaigns on Twitter,** Department of Computer Science, The College of William and Mary, Williamsburg, LNCS 7341, 455-472.

[13] Dash M, and Liu H, (1997), **Feature Selection for Classification,** Intelligent Data Analysis Journal, 131-156.

[14] El-Khair, I.A, (2006), **Effects of stop words elimination for Arabic information retrieval: a comparative study,** International Journal of Computing & Information Sciences, Volume. 4, Number. 3.

[15] El-Mawass N and Alaboodi S, (2015), **Hunting for Spammers: Detecting Evolved Spammers on Twitter**, College of Computer and Information Science, King Saud University, Riyadh, Saudi Arabia.

[16] Finn G., and Portrait A, (2011), **Of Who Uses Social Networks in the US (And How Social Media Affects our Lives),** Pew Internet and American Life Project.

[17] Golesorkhi L, (2015), **Cases of Contention: Activism, Social Media and Law in Saudi Arabia,** Arab Media & Society, Issue 20.

[18] Gomes L, Castro F, Almeida V, Almeida J, Almeid R,(2005), **Improving Spam Detection Based on Structural Similarity.**

[19] Greenhow C., (2011), **Online Social Networks and Learning,** on the horizon, Volume. 19, Number. 1.

[20] Gyongyi Z, and Molina H, (2014), **Web Spam Taxonomy**, computer science Department Stanford University.

[21] Guo D and Chen C, (2014), **Detecting Non-personal and Spam Users on Geo-tagged Twitter Network,** Department of Geography, University of South Carolina, 18(3): 370-384.

[22] Joachims Th, (1998), **Text Categorization with Support Vector Machines: Learning with Many Relevant Features,** University of Dortmund Informatik LS8, Baroper Str. 30144221 Dortmund, Germany.

[23] Kaplan M. and Haenlein M, (2010), **Users of the world, unite! The challenges and opportunities of Social Media,** Kelly School of Business, Indiana University**,** 53, 59-68

[24] Kaplan M. and Haenlein M, (2010), **Users of the world, unite! The challenges and opportunities of Social Media,** Kelly School of Business, Indiana University**,** 53, 59-68.

[25] Kietzmann H., Hermkens K., et al, (2011), **Social media? Get serious! Understanding the functional building blocks of social media,** Kelly School of Business, Indiana University**,** 54, 241-251.

[26] Kumar Sh, Morstatter F and Liu H, (2013), **Twitter Data Analytics,** Springer.

[27] Lenhart A, Purcell k, Smith A and Zickuhr K, **Social Media & Mobile Internet Use among Teens and Young Adults,** An initiative of the Pew Research Center, 1615 L St., NW – Suite 700 Washington, D.C. 20036, 202-419-4500.

[28] Mazzia A, Juett J, **Suggesting Hashtags on Twitter,** Computer Science and Engineering, University of Michigan.

[29] McCord M and Chuah M, (2011), **Spam Detection on Twitter Using Traditional Classifiers**, IEEE 1-58113-000-0.

[30] McCord M, (2011), **Spam Detection on Twitter Using Traditional Classifiers,** IEEE.

[31] Meda C, BiSIO F, Gastaldo P and Zunino R, (2014), **Machine Learning Techniques applied to Twitter Spammers Detection,** Recent Advances in Electrical and Electronic Engineering.

[32] Miller Z, Dickinson B, Deitrick W, Hu W, Wang A, (2014), **Twitter Spammer detection using data stream clustering,** Elsevier Journal

[33] Mourtada S and Salem F, (2012), **Social Media in the Arab World: the Impact on Youth, Women and Social Change,** Culture and Society Development and Cooperation, 269-274.

[34] Philippa C, (2011), **The Benefits of Social Networking Service,** University of Western Sydney.

[35] A Platinum Equity Company, **Explanation of Common Spam Filtering Techniques.**

[36] Schrape, F. (2011), **Social media, mass media and social reality construction,** Berliner Journal Fur Sociologies**.**

[37] Sedhai S, and Sun A, (2015), **HSpam14: A Collection of 14 Million Tweets for Hashtag-Oriented Spam Research,** ACM.

[38] Stringhini G, Kruegel C, Vigna G, (2010), **Detecting Spammers on Social Networks,** Published by the ACM, University of California, Santa Babara.

[39] Thomas K, Grier C, Paxson V, Song D, (2011), **Suspended Accounts in Retrospect: An Analysis of Twitter Spam,** International Computer Science Institute ACM 978-1-4503-1013-0.

[40] Van Dongen, and Enright, A.J, (2012), **Metric distances derived from cosine similarity and pearson and spearman correlations,** E-point archive, http://arxiv.org ,last visited 20-3-2016.

[41] Weinberg D. and Pehlivan E, (2011), **Social spending: Managing the social media mix,** Business Horizons, Volume. 54.

[42] Zhai Ch, and Lafferty J, (2001), **A Study of Smoothing Methods for Language Models Applied to Information Retrieval,** ACM 1-58113-331-6.

[43] Zhu Y, Wang X, Zhong E, Liu N, Li H, and Yang Q, (2012), **Discovering Spammers in Social Networks, Association for the Advancement of Artificial Intelligence,** Hong Kong University of Science and Technology, Hong Kong Renren Inc., China.

[44] http://www.wsj.com/articles/SB10001424052970204791104577107733831343976, (Last visited in: 4/7/2016).

[45] http://www.statista.com/statistics/284451/saudi-arabia-social-network-penetration/, (Last visited in: 4/7/2016).

[46] http://thenextweb.com/twitter/2012/01/07/interesting-fact-most-tweets-posted-are-approximately-30-characters-long/ , (Last visited in 11/15/2016)

Appendices

1. Arabic Stop Word

The following table shows the sub list of Arabic stop words.

Table 1: a sub list of Arabic stop words

Number	Stop Word	Number	Stop Word	Number	Stop Word
1	ان	38	دون	75	منه
2	بعد	39	مع	76	بها
3	ضد	40	لكنه	77	وفي
4	يلي	41	ولكن	78	فهو
5	الى	42	له	79	تحت
6	في	43	هذا	80	لها
7	من	44	والتي	81	أو
8	حتى	45	فقط	82	إذ
9	وهو	46	ثم	83	علي
10	يكون	47	هذه	84	عليه
11	به	48	أنه	85	كما
12	وليس	49	تكون	86	كيف
13	أحد	50	قد	87	هنا

14	على	51	بين	88	وقد
15	وكان	52	جدا	89	كانت
16	تلك	53	لن	90	لذلك
17	كذلك	54	نحو	91	أمام
18	التي	55	كان	92	هناك
19	وبين	56	لهم	93	قبل
20	فيها	57	لأن	94	معه
21	عليها	58	اليوم	95	يوم
22	إن	59	لم	96	منها
23	وعلى	60	هؤلاء	97	الى
24	لكن	61	فإن	98	اصبح
25	عن	62	فيه	99	امسى
26	مساء	63	ذلك	100	اضحى
27	ليس	64	لو	101	ستكون
28	منذ	65	عند	102	مما
29	الذي	66	اللذين	103	ابو
30	أما	67	كل	104	لدي
31	حين	68	بد	105	وهي

32	ومن	69	لدى	106	الذي
33	لا	70	وثي	107	هن
34	ليسب	71	أن	108	يمكن
35	وكانت	72	ومع	109	فإن
36	أي	73	فقد	110	اليها
37	ما	74	بل	111	انه
112	عنه	113	هو	114	بدلا
115	حول	116	عنها	117	اي

2. Prefixes Matrix

The following table shows the sub list of Arabic prefix.

Table 2: a sub list of Arabic prefix

Number	Prefix	Number	Stop Word	Number	Stop Word
1	ت	11	فال	21	با
2	ول	12	يتس	22	كال
3	وبم	13	فك	23	فلل
4	سن	14	است	24	ي
5	ولت	15	ال	25	فل
6	وال	16	فبال	26	سيت
7	فب	17	بال	27	سيست
8	اف	18	و	28	سيس
9	تست	19	ست	29	لل
10	فكال	20	ستت	30	اس

3. Suffix Matrix

The following table shows the sub list of Arabic suffix.

Table 3: a sub list of Arabic suffix

Number	Prefix	Number	Stop Word	Number	Stop Word
1	ا	11	ان	21	تما
2	ت	12	هـ	22	هم
3	يون	13	كم		
4	اتية	14	وا		
5	ون	15	كما		
6	تم	16	ها		
7	نا	17	ات		
8	ين	18	هما		
9	يه	19	هن		
10	كن	20	تن		