نموذج رقم (1)

إقـــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

تعريف المحتوى السياسي الفلسطيني في التعريفات العربية

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وإن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

DECLARATION

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification

Student's name: اسم الطالب: حسام عبد الرزاق الكرد

Signature: التوقيع:

Date: التاريخ: 2015 / 12 / 14

بسم الله الرحمن الرحيم

The Islamic University of Gaza
Deanery of Graduate Studies
Faculty of Information Technology

الجامعة الإسلامية – غزة
عمادة الدراسات العليا
كلية تكنولوجيا المعلومات

Identifying Palestinian Political Content from Arabic Tweets

By

Hussam ElKurd

Supervised by:

Dr. Rebhi Baraka

October 2015

Thesis Submitted to the Faculty of Information Technology in Partial Fulfillments of the Requirements for the Degree of Master in Information Technology

بســـم الله الرحمن الرحيم

الجامعة الإسلامية – غزة
The Islamic University - Gaza

مكتب نائب الرئيس للبحث العلمي والدراسات العليا هاتف داخلي: 1150

الرقم: ج س غ/35/ Ref

التاريخ: 2015/10/26 Date

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ **حسام عبدالرازق أحمد الكرد** لنيل درجة الماجستير في **كلية *تكنولوجيا المعلومات*** **برنامج تكنولوجيا المعلومات** وموضوعها:

تعريف المحتوى الفلسطيني السياسي من التغريدات العربية

Identifying Political Palestinian Content from Arabic Tweets

وبعد المناقشة التي تمت اليوم الاثنين 13 محرم 1437هـ، الموافق 2015/10/26م الساعة الثالثة مساءً، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

د. ربحي سليمان بركة	مشرفاً و رئيساً	....
أ.د. علاء مصطفى الهليس	مناقشاً داخلياً	....
د. سناء وفا الصايغ	مناقشاً خارجياً	....

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في **كلية *تكنولوجيا المعلومات*/ برنامج** تكنولوجيا المعلومات.

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق ،،،

نائب الرئيس لشئون البحث العلمي والدراسات العليا

أ.د. عبدالرؤوف علي المناعمة

صدق الله العظيم

Dedication

To my peacemaker and teacher prophet Mohammed.

To my loving parents for give me support and push always.

To my beloved wife for being responsible and supportive.

To my daughters Solaf and Sham whose missing care while the study.

To my brothers and sister as they all kept me going.

To all my friends, colleagues for their endless support.

Acknowledgement

The completion of this thesis could not have been possible without Allah. Thanks to Allah first and foremost who bestowed me the awareness, perseverance and mercy.

I would like to express my sincere appreciation to my supervisor Dr. Rebhi Baraka for his steady guidance, help and relentless support.

In addition, I would like to extend my thanks to the academic staff of the Faculty of Information Technology who helped me during my master's study and taught me different courses.

Abstract

Using Twitter in news agencies and media has become widely popular, thus considerable proportion of tweets greatly reflects the social perspective in the real world. Further many people follow the news on twitter which attracts the news agencies to analyze and try to know what is happening on Twitter. Press and media agencies are looking to find efficient tools to analyze and classify tweets and this is due the difficulty and high cost of the manual approaches. Various works have discussed and provided effective solutions for processing tweets into understandable formats for machines to classify and analyze. While researches receive much attention in languages and locales such as English, some languages such as Arabic have not received much research attention despite the wide spread of Twitter in the Arab world in general and Palestine in particular.

In this thesis we propose an approach using machine learning to automatically classify Arabic tweets related to Palestinian political topics/content. The purpose of classifying Palestinian Arabic political topics is that Palestine receives great attention in Arab news and social media. The approach is based on collecting tweets using an application that we develop based on the TwitterPalPol. It collects tweets from different Twitter API's through specific factors like keywords, region and language. Then we process the collected tweets and classify part of them manually as Palestinian political and as not Palestinian political in order to be used as the learning data set for the selected machine learning. This is used in the algorithm to classify new tweets automatically. In addition we create two datasets for learning, the first one includes all the collected tweets prepared for learning, and the second include filtered tweets with creditability filter created to evaluate the creditability for each tweet and ignore fake tweets. The filter is dependent on many factors related to tweet properties, therefore we compare the results for the classification in both data sets and find out the importance of the filter. The results was sufficient as they preserve ranges between 97% and 80% in the main classification measurers like recall and precision.

Keywords: Machine Learning, Text Processing, Text Classification, Twitter.

الملخص

تعريف المحتوى الفلسطيني السياسي من التغريدات العربية

أصبح تويتر شائع الإستخدام خصوصا في وكالات الأنباء والإعلام. وذلك بسبب أن التغريدات تعكس نظرة المجتمع في العالم والأحداث الحقيقه, كما أن هناك أعداد كبيرة تتواجد وتتابع كل جديد على تويتر, مما يجذب جميع الوكالات الإخبارية إلى تحليل ومعرفة ما يحدث في هذه الشبكة. الوكالات والصحافة تبحث دائما على الأدوات التي تساعدها في تحليل التغريدات, نظرا لصعوبة وتكلفة الأمر من خلال المتابعة اليدوية, بحيث تستطيع من خلال هذه الأدوات معرفة ما يحدث في قضايا وأحداث أو مناطق محددة. هناك أعمال كثيرة ناقشت وقدمت حلول لتحليل التغريدات بلغات لصيغة تستطيع الآلة تصنيفها وتحليلها. ولكن هذه الأعمال والأبحاث تعتني في مناطق ولغات محددة, على عكس اللغات الأخرى مثل العربية والتي ليس هناك إهتمام بحثي كافي يرقى لمستوى أهمية تويتر في الوطن العربي والقضايا والأحداث الكثيرة والكبيرة وعلى رأسها القضية الفلسطينية.

في هذه الرسالة قدمنا منهجية بإستخدام التعلم الآلي يمكن من خلالها تصنيف التغريدات العربية إلى تغريدات سياسية فلسطينية أم لا بشكل آلي. ويرجع السبب إلى إختيارنا للتصنيف للقضايا الفلسطينية السياسية كونها واحدة من أهم القضايا والتي تهتم فيها أغلب وسائل الإعلام العربي وكذلك وسائل التواصل الإجتماعي. منهجيتنا تقوم على جمع التغريدات حيث قمنا بتطوير تطبيق TwitterPalPol يقوم بجمع التغريدات من خلال منافذ تويتر المختلفة عبر محددات معينة مثل الكلمات المتعلقة بالقضية الفلسطينية والمنطقة واللغة ومن ثم نقوم بمعالجة التغريدات وتصنيف جزء منها بشكل يدوي إلى تغريدات متعلقة بالقضية الفلسطينية من الناحية السياسية وتغريدات لا تتبع, بحيث يتم إستخدام هذه البيانات لتغذي خوارزمية التعلم الآلي بحيث يمكنها تصنيف تغريدات جديدة بشكل آلي. وفيما يتعلق بالبيانات قمنا بتجهيز قاعدتين من البيانات, حيث تحوي قاعدة البيانات الأولى جميع التغريدات التي قمنا بجمعها, بينما تحوي القاعدة الآخرى بيانات للتغريدات التي تم تصفيتها من خلال مصفي القيمة والذي يقوم على مجموعة من المحددات يتم فيها إستثناء التغريدات الوهمية أو الأقل أهمية, لنقارن النتائج فيما بعد ونعرف أهمية وجود مرحلة تصفية أم لا, حيث كانت النتائج مرضية وعالية حيث كانت تتراوح بين الممتاز والجيد جدا في وحدات القياس الأساسية للتصنيف مثل ال (Recall) و ال (Precision).

الكلمات المفتاحية: التعلم الآلي, معالجة النصوص, تصنيف النصوص, تويتر.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

FN	False Negative
FP	False Positive
TF-IDF	Term Frequency - Inverse Document Frequency
TN	True Negative
TP	True Positive
SVM	Support Vector Machine
JSON	JavaScript Object Notation
API	Application Programming Interface
NB	Naïve Bayes
K-NN	K-Nearest Neighbors

Chapter 1
Introduction

Recently Twitter become the most well-known microblogging social network in the internet world, people send over 400 million tweets every day (Tsukayama 2013). Over 15 billion twitter API calls per day are requested from different clients to retrieve tweets (Ciotec et al. 2014). Hence this shows how much importance of twitter today. Tweets have a natural touch, which make twitter fertile environment for public opinion. In the last few years many people use twitter for liberation from governmental imposed constraints. In the Arab world, revolutions took benefit from Twitter where protesters and activists tweet about events taking place, as well as wars erupted in the region in Syria, Palestine and many others countries.

Whereas this excite press agencies to become more interested in twitter not only in publishing news, but also in classifying and analyzing public opinion. However it is much difficult to discover tweets manually without using tools to support classification and analysis for the tweets. For example, the total rate of tweets from Egypt, and around the world, about political change in that country ballooned from 2,300 a day to 230,000 a day (O'Donnell 2011), and that figure how much difficult to detect and identify tweets manually.

Although there are many research contributions in classification and analysis of tweets, they are limited to some languages, while in the other hand some languages such as Arabic does not receive much research attention though in the Arab world about 62 million users write 10 million tweets every day (Jazra 2014).

In our research we propose an approach to identify political Palestinian topics from Arabic tweets. The purpose of selecting political Palestinian topics is that it among the threads receiving great interest. In addition Palestine is one of the hottest conflict area in the world, therefore receive much attention from the news and social media on the web.

Our approach relays on creating labeled trained data set as input for custom machine learning algorithm that take defined factors and equations to classify new political Palestinian Arabic tweets. The purpose of selecting machine learning is ability to learn new factors used to classify according to the changes. Thus it does not require large training data set.

We collect data using twitter API and identify tweets related to political Palestine Arabic tweets in order to define the set of factors to label new tweets. Furthermore other factors will be taken into account such as number of followers and verified account. Reformat the collected data to be used as training data, the next step is to format the data and submitted to the machine learning algorithm to classify new data and label each tweet to be an Arabic Palestine political tweet or not depending on the learning process. In each phase we evaluate the accuracy value for defining the topics and the classification result. We conduct

our results and have obtain different evaluation measure to evaluate the system from different perspectives.

1.1 Problem Statement

Currently, classifying and analyzing social media is a hot research topic as a result of too much demand from different organizations like media and news agencies to get useful information. Although twitter plays important role in Arab world especially in areas of conflict and high political interest like Palestine, Arabic language has not received much attention in such research topics.

The problem is how to identify political content with acceptable accuracy and creditable content from Arabic Palestinian tweets in order to get a general view from news agencies that reflect Palestinian issues from a social perspective.

1.2 Objectives

1.2.1 Main Objective

To develop an approach using machine learning to identify Arabic political Palestinian tweets topics with acceptable accuracy and creditable content.

1.2.2 Specific Objectives

- To collect proper training data set for training machine learning algorithm based on suitable learning methods.
- To define a set of factors and terms to process the collected training data.
- To select and use an efficient algorithm that is able to identify tweet topics.
- To remove fake tweets and content and provide creditable content.
- To implement the algorithm and conduct several experiments on various tweets and accuracy based on measures such as precision and recall.
- To evaluate the result and justify it from different perspectives.

1.3 Importance of the Research

Now a days, news agencies are looking to get overview from social perspective about different events and circumstances. Therefore microblogging social networks such as twitter help them to get fast communication and information shared. In contrast social networks are huge space and volume and demand auto classification to get specific useful information. Many researches address the problem in specific language like English. While Arabic language not getting much attention in this field. Our research help to cover the gap for Arabic language classification in twitter. Moreover we define political Palestinian topics support news agencies get general view in this hot news topic. However classification rules can be used in any other classification label.

1.4 Scope and Limitation

The work is applied with some limitations and assumption such as:

1. Our work is limited to Arabic Palestinian tweets with pre-defined labels.

2. Not all tweets will identified, the algorithm will have a degree of uncertain results.

3. The factors and data for classifications are subject to twitter features and policy.

4. The produced classified tweets are labeled in single class label (political), therefore it provide news media with political news.

1.5 Research Methodology

In our thesis, we devote our study on classification based on timely fashioned unlabeled instances. In our process we shall use adaptive supervised learning technique with delayed labeling. This is done to change and update the training set by what is called formation methods. We follow a research methodology that consists of the following:

- **Research and survey:** this include reviewing the recent literature closely related to the thesis problem statement and the research question. After analyzing the existing methods, identifying the drawbacks or the lack of existing approaches, we

formulate the strategies and solutions how to overcome the drawbacks.

- **Building the Approach:** the structure of our proposed approach include the following general step:
  1. Gathering tweets from different twitter by API.
  2. Prepare the collected data set into proper format.
  3. Use part of the data for training throw label it manually into Arabic Palestinian Political class or not.
  4. Use machine learning algorithm with supervised train set and apply it to unlabeled data set to label the new tweets.

- **Implementation:** we implement the learning phase using PHP and JavaScript to collect the data from twitter API's and use rapidminer tool to apply the learning and classification procedures.

- **Evaluation:** To evaluate our approach, we use different measures discussed in chapter 3 elaborate evaluation from different perspectives. Confusion Matrix one of the most important ways in which we relied upon to calculate the metrics

- **Results and discussions:** in this stage we will analyzed the obtained results and justify our model feasibility.

1.6 Thesis Format

The thesis is organized as follows: Chapter 2 present related work in classification in machine learning and Arabic content, Chapter 3 covers the theoretical and technical background related to twitter, machine learning, classification and its evaluation, Chapter 4 presents the proposed approach for identifying tweets, Chapter 5 describes the implementation and the experiments, Chapter 6 presents the evaluation of the results. Finally chapter 7 concludes the thesis and suggests some future work.

Chapter 2
Related Works

Basically, our methodology in considering related works is to look for researches investigate in classification tweets using machine learning algorithms, and the studies discuss Arabic classification techniques.

2.1 **Tweets Classifications Studies**

Many studies classify tweets in different assumptions. In (Pennacchiotti & Popescu 2011), they identify classification model for English tweets and apply it in different experimental domains such as political affiliation democrats (positive set) or Republicans, Ethnicity classifying users as either African-Americans or not. Their work depends on an equation with four main terms introduced in the following: 1) User profile features: get user profile features such as user, bio and many others to discover user type and reality. 2) Tweeting behavior: number of tweets posted by the user. 3) Linguistic content: explore a wide variety of linguistic content features prototypical words. 4) Social network: explore the social connections established by the user with others he follows. This study depends on set of tweets to identify weather this is positive or negative while our research is classify each tweet individual, the common thing between the study is to identify the vector of features to be used to remove noise tweets.

(Lee et al. 2011) Propose classification approach and evaluate accuracy of different machine learning algorithms. The approach classify tweets in two stages for classification 1) text based classification start with clearing each tweet from lingo or hyperlinks to estimate the (term frequency–inverse document frequency) weights in order to evaluate the importance of a word(term) to a document. The importance is proportional to the number of times a word appears in the document. 2) Network based classification depend on using twitter specific social network information that provide linkage indicates common interest between two users, so that applying algorithm from user similarity model and weighted page rank algorithm with the twitter information predict topics similar to trending topic. The study aim to classify twitter trending topics (a list of most popular topics people tweet about provided by twitter) into 18 general categories. The proposed methodology is to collect sample of twitter trending topics data using twitter API as Data Collection stage and label. Lastly the approach is evaluated as shows in the next paragraph.

Their approach evaluates classification in different algorithms to get the most accurate machine learning algorithm. For the text based Naive Bayes Multinomial gives best classification accuracy (65.36%), while C5.0 decision tree classifier gives best

classification accuracy (70.96%), they look to integrate text-based classification using Naive Bayes Multinomial (NBM). The study is miss up the SVM as one of the most important algorithms in the text classification, in our study we use SVM as many studies refer and it is one of the most popular classification algorithms.

In the words of (Pennacchiotti n.d.) Build a user classification architecture to classify user by interest based on the attributes of data and the graphic, they solve it using an architecture with two components: the first component is a machine learning algorithm that learns a classification model from labeled data and user-centric features. The model is then used to classify new incoming users (e.g., as being a Democrat or a Republican). The second component of the architecture is a graph-based label updating function which uses social graph information in order to revise the classification label assigned to each user by the initial machine learning component. In our thesis we focus on classifying tweets not user, therefore we have use machine learning on tweets content, further the classification is based on Palestinian political and not too generic as this work.

(Cheng et al. n.d.) Proposed and evaluated a probabilistic framework for estimating a Twitter user's city-level location based purely on the content of the user's tweets, even in the absence of any other geospatial cues. The content-based approach relies on two key requirements: (i) a classification component for automatically identifying words in tweets with a strong local geo-scope; and (ii) a lattice-based neighborhood smoothing model for refining a user's location estimate. The proposed approach are not using machine learning, therefore they have to update the words dictionary manually to identify the tweets reflecting location, while we have use machine learning to enable auto discover of new words and content reflecting the classification needs.

In addition (Piao & Whittle 2011) present their research on the feasibility of extracting Twitter users' interests for suggesting serendipitous connections using natural language processing. The approach is to use three main sources for extract user interest 1) user number of tweets, 2) user shared tweets with other users 3) hyperlinks inside the tweets we will add those factors and more others to extract topic. This also another research find the interests of user throw set of tweets. In contrast our research is classifying each tweet individually.

Adopting twitter as information source has two main concerns, the first concern is that twitter policy impose 140 char for each tweet that is not enough to provide sufficient information to extract interest if it depend on a single tweet. However collecting tweets provide a wealth of information about user's thoughts and interests as shows in study of (Sakaguchi et al. 2010). In contrast (Kim et al. 2010) found that "even though tweets are brief, they contain enough information to express identifiable characteristics, interests and sentiments". This study proof that the tweets can be classified individually and this what our research do.

2.2 **Arabic Tweets Classifications Studies**

For Arab studies we find a study for (Al-Eidan et al. 2010) aim to measures Arabic content credibility published in Twitter, propose a system to classify the tweets as creditable, not credible and questionable, thus, the tweets are processed and then classified throw measuring the credibility rate, the rate is estimated by the similarity of tweet content and the Arab article news content also the account verified on twitter or not, and the URL related to the tweet if exist. The research was evaluated by Appling two different topics and find how much creditable are tweets in this topic, this may be useful in our future work in order to detect wither the tweet contain a creditable news or not.

Although this study is useful, but it's limited in focusing on content credit not topic identification, however hash tags not used. In our study we take some properties of creditability to remove fake tweets or users described in Section 4.2.

In the word of Ingmar (Weber & Garimella 2013) study the phenomenon of secular vs. Islamist polarization in Egyptian tweets, start by create a labeled data set (secular or Islamic) of users and obtain each one tweets with their bio and location, aggregate hashtag and estimate each hashtag polarity as it secular or Islamist or neutral, adds tweet URL as a factor from inspecting the most polarized domains, moreover the content of tweets presented from basic dictionary-based approach compiled a list of 609 terms (381 Arabic and 228 English) related to Islam, Hence the relative usage of religious terms increases as a user closer to Islamic, this is fully depend on statistics and there is no real algorithm used to predict or mining, for our approach we can apply approaches starting with labeled data and obtain the user bio and location, the URL attached with tweets, for the tweet content our approach is a pretty different hence it is processed as terms with correlated to each other.

An argument in favor of study (Ahmed Ibrahim & Salim 2013) investigated studies of twitter opinion mining and Arabic tweets, many studies focus on sentiment analysis, therefore NB, SVM is the most used in classification in general and in Arabic also, our approach use SVM to apply it in classification phase.

Another study of (Duwairi & Qarqaz n.d.) Present supervised learning for sentimental analysis of tweets (positive, negative, natural). The study use different machine learning classifiers such as NB, SVM and KNN and apply it on prepared dataset. We find that the study have many sides of limitation in order to get better accuracy. However in our approach is classifying each tweet individually and this is the challenge, while sentimental analysis depends on set of tweets that similar to document.

In recent study of (Bekkali & Lachkar 2014) proposed method for Tweets Representation based on Rough Set Theory. Arabic Tweets Categorization using NB and SVM classifiers. They add tool for data analysis and classification to estimate the Precision of classification.

The research have no creditability check it depends on tweet. In our approach we are going to include creditability and support learning for new classifiers.

In addition the study (El-Halees 2007) classifying Arabic text based on maximum entropy, while there is a difference on the methodology and the purpose of the classification in contrast to our study, therefore the study using maximum entropy for classifying documents with many words and content, thus they have more prepressing and more information to classify. The general steps looks similar to the first phase as they have preprocessing and training, in the other hand our research are classifying tweets individually however each document represents no more than 140 character and we use SVM as machine learning algorithm to classify the tweets.

In (Kourdi et al. 2004), They classify Arabic documents using the NB algorithm into different categories like sport and health and culture etc., they have collected 300 web documents for each of five categories and execute NB algorithm for classifying and evaluate the result using cross validation, the study look retrieval as there is no clear approach to maintain preprocessing and data processing also there is no clear justification to use NB as algorithm, our research look different from the type and content and algorithm of classification, the common thing is that our research use the same approach of machine learning.

We find the Arabic studies focus on sentimental analysis, and depend on set of tweets to maintain classification, for our approach we classify each tweet individually and this make classification more complicated, further we have use creditability filter to remove fake contents.

Chapter 3
Theoretical and Technical Foundation

This chapter describes Twitter environment. It introduces the most common phrases and terms, discusses the Twitter user layout and its main parts, and describes Twitter API and how it works. It presents the learning and evaluation methods.

3.1 Twitter Phrases and Terms

Twitter has its own expressions. We introduce the common ones in the next paragraphs.

A Tweet: is a text message containing no more than 140 characters (less in some language like Arabic 80 character) to put down thoughts were people write what is happing or what they like to share. Thus interested people can view and interact with it (Twitter 2014b).

In our research tweet present the main source for the learning and classification, therefore we proceed each tweet individually, thus the classification is performed on tweet text content and the related twitter properties. Part of the tweets is neglect because of it has less words, or fake tweets. We maintain different filters introduced in chapter 4 to filter fake tweets.

Hashtags (#): used to create groupings and help generate popularity around particular keyword or topic. The hashtag is created by preceding word with the hash mark (#). For example if someone talk about Palestine he can add #Palestine inside the tweet, therefore people can search for the tweets contain Palestine and also the tweet are generally belongs to this words (Twitter 2014c).

Hashtags are used indirectly in our work. Because it contains indication words provide initial classification if the tweet belongs to political Palestinian class or not. In the other hand hashtags may cause deception, thus the hashtags are not related to the tweet content. However we have not fully depend on the hashtag, but also the full tweet content.

In addition we have use another factors to classify tweets, in the next section we will present other properties that can be used as creditability to the tweet.

@Replies: when tweets show up to users any user can reply into particular tweet. Similar to posts comments. The reply is offered by reply button or use @ with user name in the tweet text. The reply is considered as a tweet, thus we remove ”RT” (Reply Tweet) from tweet and perform the same procedures of classification (Twitter 2014b).

Mentions: mention users in tweet, the technique is similar to reply. Use @ and the user name when writing a tweet. In the data preparation we remove user mentions from the tweet text to classify the text content not usernames (Twitter 2014b).

Retweet: A Retweet is a re-posting of someone else's Tweet. Twitter's Retweet feature helps user quickly share that Tweet with all followers (Twitter 2014a). We have use the retweet number as one factor from different factors in credibility check to verify the tweet source as discussed in chapter 4.

Follow/Unfollow: an action to follow particular user profile and get his tweets feed. In contrast unfollow is not get the user feed. Each user have a number of followers. Further we have use the number of followers in credibility check to verify the tweet source.

Favorite tweets: favorite someone else tweet. This gives a credit to the number of favorites to particular tweet. Direct messages: Also called DMs, direct messages let user communicate privately with other Twitter users. To send a direct message, type the letter D followed by the username want to reach, and then enter message. In the credibility check number of favorites are also taken into account to verify the tweet source.

Link shorteners: If user have 140 characters, user do not want to use 50 of them by including a long URL. He need to shorten the URL so that he can save some characters.

Most URL shorteners shrink the links to anywhere from 16 to 20 characters. We have remove any html or links in the tweet in preparing data, thus not effect the classification process.

Verified Account: Any account with a blue verified badge on their Twitter profile is a verified account. Verification is currently used to establish authenticity of identities of key individuals and brands on Twitter. We consider any verified account are verified source for the tweets, therefore we didn't maintain any creditability check if the account is verified.

The **user profile** include custom parameter defined by customer, part of them are custom. The parameters appear in screen user name, bio, location and many others. In our research some parameters will be an argument for classification.

3.2 Twitter Layout and Interface

In this section we like to visualize the properties mentioned in the previous section, therefore we can find it in twitter layout. Twitter is designed to meet user's interest, therefore the layout is presented in many components with different functionality. Figure 3.1 shows the view of the twitter user layout.

Figure 3.1 Twitter User Layout

General profile information takes up part of the layout. Figure 3.2 shows such profile. It consists of the following components:

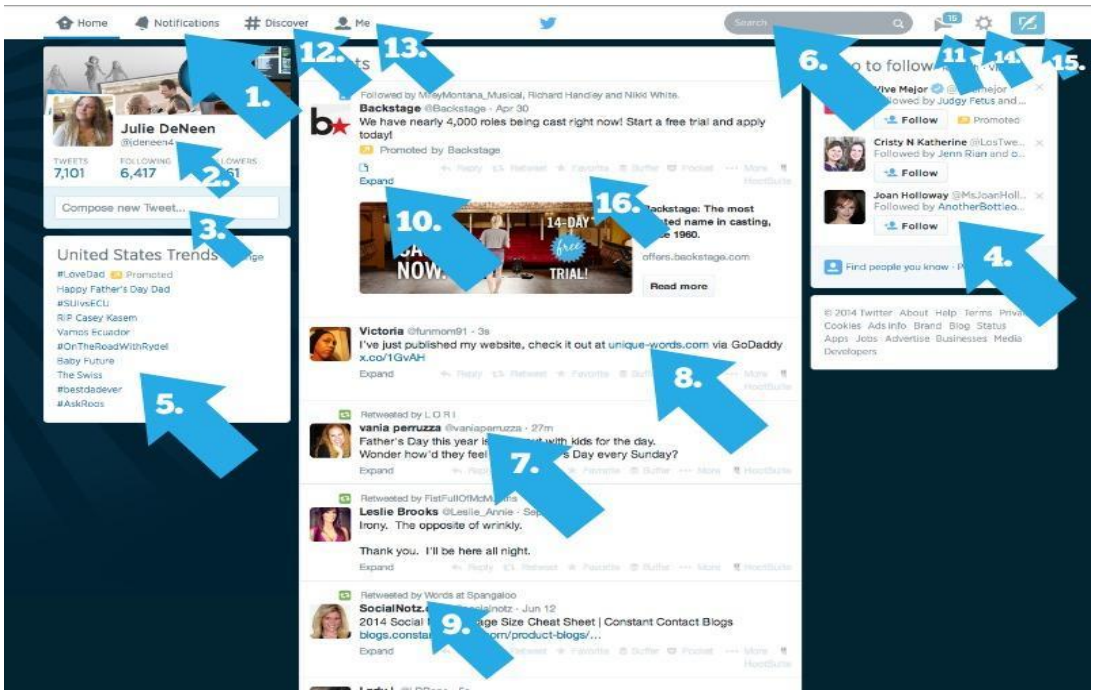

Figure 3.2 Twitter User Layout Description

1. Notification button. When clicking on it, it shows a list of tweets where the user account is mentioned. Any new follower, or retweeted, It's basically seems like Twitter inbox. Anyone who has said something to the user directly will appear in the connect area.
2. Clicking on the link will take the user into his profile page. This is where he set up your profile photo, bio, etc. It also has a list of all tweets plus his following and followers users.
3. Composing new tweet the user can write a tweet.
4. Twitter suggests some accounts that have similarity with user interests in order to follow. Therefore user can pick up new people to follow.
5. The box points to hash tags that are trending. In other words, what's the most popular hashtags at the current time. User can customize the recent popular hashtags according to location.
6. This is for general search as users can search for specific hashtag, users, and subjects.
7. Next to each tweet is be the person's twitter handle. If the user wants to speak directly to someone, he types in their twitter handle and it will show up on their connect screen.
8. Those are links that people have pasted. If the user clicks on the link, it takes him to another site.
9. "Retweeted by…" That means that someone the user follow took another person's tweet and retweeted it.
10. The view summary button is a quick way to examine the tweet without having to go to another page. If someone tweets out, "Check out my blog!" and then pastes a link, the user can hit view summary and it will give him a snapshot of the website. If he like what he sees, clicks on the link to go there. Sometimes, it won't say view summary but rather, "Expand" or "View conversation". When the user sees a tweet where one person is talking to another, he can hit that button to see what the whole conversation was. It helps to place the tweet in context.
11. This is the direct message inbox. It's the place where someone can privately message the user.
12. The Discover tab allows the user to search hashtags and keywords. It's basically a way to filter through various conversations.
13. The ME tab simply takes the user to his profile (showed in figure 3.1).
14. The setting is where the user can adjust his settings, email notifications, background colors, widgets, etc.
15. That's the compose tweet button, same functionality as point number.
16. This button to favorite any tweet the user views in the timeline, therefore the tweet will increase the number of favorites.

Throw getting the general properties for the tweet and how we can use these properties to have more classification accuracy, the next is to know how to collect this tweets within specific criteria.

3.3 Twitter API'S

Twitter has online API that allows different applications to interact with it, according to (Ciotec et al. 2014) over 15 billion twitter API calls per day requested from different clients, to retrieve tweets and interact with. This shows how intensively website or application in the worldwide conversation happening on Twitter.

In conjunction twitter platform has developed different types to help developers develop application and connect it to twitter, we have try more than API and technology to collect the tweets described as the following:

3.3.1 REST APIs

(Twitter n.d.) The REST APIs provides programmatic access to read and write Twitter data. Author a new Tweet, read author profile and follower data, and more. Responses are available in JSON. We have use REST API to collect tweets with specific criteria; therefore the criteria are defined with set of parameters passed to the API. Described below :

The Query Operator (q): the query operator look similar to any search field, as you write the keywords to get the tweets that contain those words. Twitter query operator has operators that modify its behavior such as when we set the q= love OR hate we find tweets containing either "love" or "hate" (or both), another example when set q= beer –root this means containing "beer" but not "root", likewise many operators available for modifying behavior. We have use query operator in defining what are the keywords that we look to find in the tweets. However the keywords are to important to get tweets within specific context.

Language: the lang parameter restricts tweets to the given language. We set the lang param to Arabic to collect Arabic tweets.

Geolocalization: the geocode parameter specified with the template "latitude,longitude,radius", for example, "37.781157,-122.398720,1mi". When conducting geo searches, the search API will first attempt to find tweets which have lat/long within the queried geocode, and in case of not having success, it will attempt to find tweets created by users whose profile location can be reverse geocoded into a lat/long within the queried geocode, meaning that is possible to receive tweets which do not include lat/long information.

Number of tweets: the "count" parameter define the number of maximum tweets to be collected, we use the maximum limit, which is 100 tweets.

Result type: the result type parameters can be defined as popular that collect popular tweets only, recent that collect the new tweets and mixes contain both, we have set it to be both.

In addition there are more parameters available but we actually use these parameters to define the criteria of collecting tweets. In the other hand twitter require authentication to collect tweets, therefore we have use use OAuth to authorize our script.

3.3.2 **OAuth**

OAuth is an open standard for authorization (Twitter 2015).. OAuth provides client applications a 'secure delegated access' to server resources on behalf of a resource owner. It specifies a process for resource owners to authorize third-party access to their server resources without sharing their credentials. It designed specifically to work with Hypertext Transfer Protocol (HTTP), OAuth essentially allows access tokens to be issued to third-party clients by an authorization server, with the approval of the resource owner. The client then uses the access token to access the protected resources hosted by the resource server. Twitter use OAuth 2 to verify API request, therefore it require a valid twitter account to allow create an application with 2 keys named as consumer and secret keys, these keys are sent to twitter authorization and it will response with bearer key that can be used to authorize any request via twitter API, figure 3.3 describe the authorization process

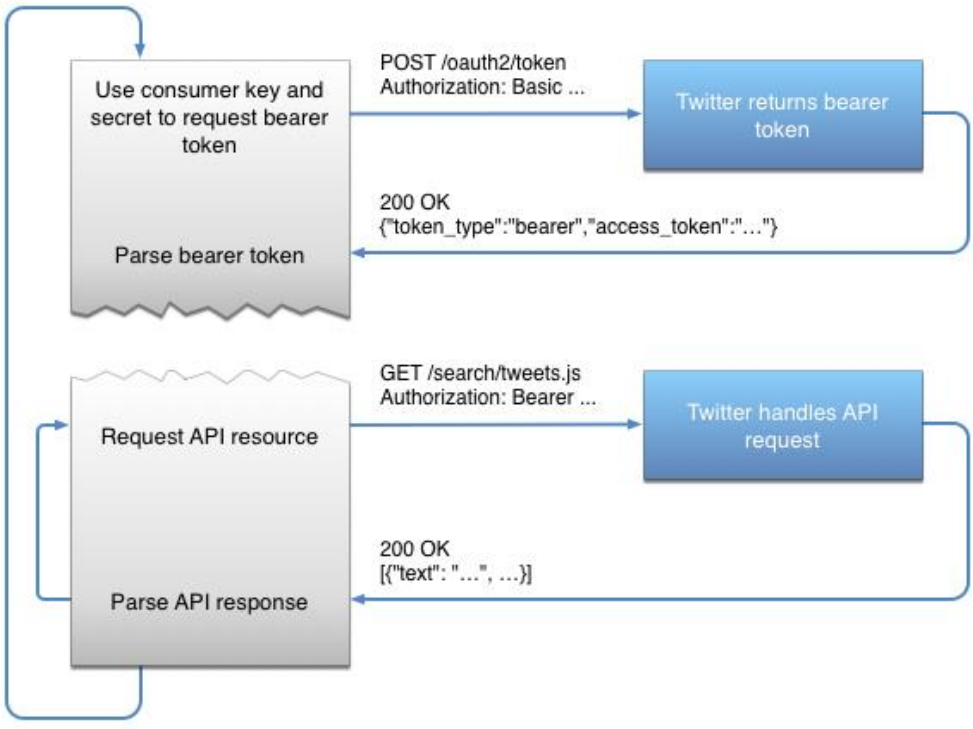

Figure 3.3 Twitter API Authorization Process.

We have use Twitter API to collect the tweets, but we find the number of the responded tweets are limited and doesn't supply our research another more concern is that twitter API

provide you with tweets within 8 past days maximum and no historical tweets can collect, we search for an alternative to provide bulk tweets and we find Topsy API.

3.3.3 **Topsy API**

Topsy is a real-time search engine powered by the Social Web. Unlike traditional web search engines, topsy indexes and ranks search results based upon the most influential conversations millions of people are having every day about each specific term, topic, page or domain queried.

Topsy look similar to twitter API of defining the criteria, the main difference is how to authorize your request as topsy just use an API Key and they add additional information for the response tweets (Topsy 2015).

We use topsy to get bulk tweets from previous dates, as twitter does not allow collecting tweets with less than 8 days from the current day.

3.4 **Supervised vs. Unsupervised Learning**

Machine learning is a class of algorithms which is data-driven (At n.d.), i.e. unlike "normal" algorithms it is the data that "tells" what the "good answer" is. In our case we look to have an algorithm to classify the tweet text content set. A machine learning algorithm would not have such coded definition, but would "learn-by-examples": you'll show several text classifications and a good algorithm will eventually learn and be able to predict the class for the new text.

This particular example of our case is **supervised**, which means that examples must be *labeled*, or explicitly say which ones are belongs to your class and which ones aren't.

Supervised learning: classification is seen as supervised learning from examples.

- Supervision: The data (observations, measurements, etc.) are labeled with pre-defined classes. It is like that a “teacher” gives the classes (supervision).
- Test data are classified into these classes too.

Thus supervised learning require two main phases, the first is the learning process and the second one is to maintain the classification using machine learning algorithm.

In an **unsupervised** algorithm examples are not *labeled*, i.e. you don't say anything. Of course, in such a case the algorithm itself cannot "invent" what a class is, but it can try

to cluster the data into different groups, e.g. it can distinguish that text are very different from sport, which are very different from politics.

Unsupervised learning (clustering)

- Class labels of the data are unknown.
- Given a set of data, the task is to establish the existence of classes or clusters in the data.

We have use supervised learning because in our case the problem is to label single class not a lot of classes or cases, further supervised learning is often faster than reinforcement learning technique, thus it depends on defined rules.

In the other hand the selecting of the algorithm to apply the learning and classification process had to be justified, thus in the next section we introduce the most well-known algorithms in machine learning.

Selecting the suitable learning algorithm for classification reflected directly of how to setup the experiment and the final results. However many algorithms issued specific problems. In the below paragraphs we introduce set of algorithms that could be applied in our classification problem.

3.4.1 **Support Vector Machine (SVM)**

Used for classification, regression, or other tasks. Intuitively, the hyper plane that has the largest distance to the nearest training-data point of any class achieves a good separation, support linear and non-linear classification.

SVMs are helpful in text and hypertext categorization as their application can significantly reduce the need for labeled training instances in both the standard inductive and settings. Because it makes this linear approximation, it is able to run fairly quickly. Where it really shines is with feature-intense data, like text or genomic. In these cases SVMs are able to separate classes more quickly and with less over fitting than most other algorithms, in addition to requiring only a modest amount of memory.

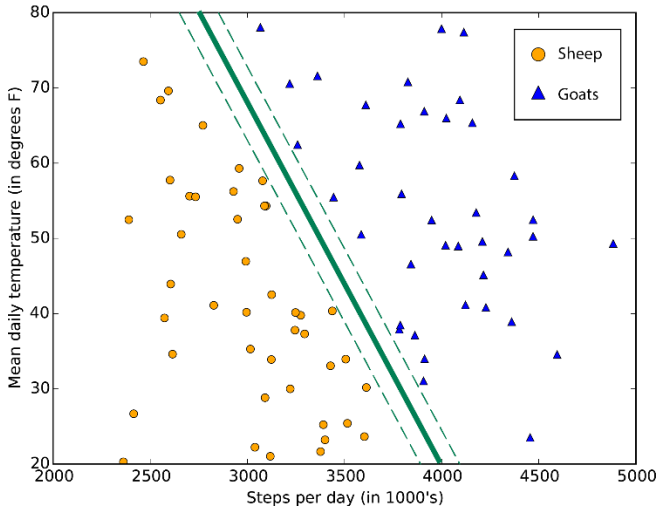

Figure 3.4 Typical SVM Class Boundary Maximizes the Margin Separating Two Classes

SVM can be formulated as the following, given some training data D, a set of n points of the form

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\ y_i \in \{-1, 1\}\}_{i=1}^{n}$$

Where the Yi is either 1 or −1, indicating the class to which the point Xi belongs. Each Xi is a P -dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having Yi = 1 from those having Yi =2. Any hyperplane can be written as the set of points X satisfying

$$\mathbf{w} \cdot \mathbf{x} - b = 0$$

Where **.** denotes the dot product and $\mathbf{w}$ the (not necessarily normalized) normal vector to the hyperplane. The parameter $\frac{b}{\|\mathbf{w}\|}$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$.

If the training data are linearly separable, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin". These hyperplanes can be described by the equations

$\mathbf{w} \cdot \mathbf{x} - b = 1$ And $\mathbf{w} \cdot \mathbf{x} - b = -1.$

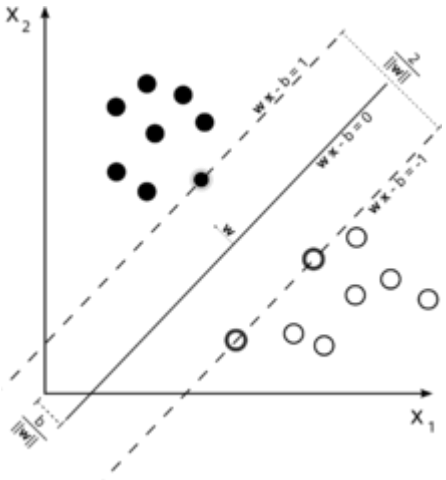

Figure 3.5 Maximum-Margin Hyperplane and Margins for an SVM Trained with Samples from Two Classes.

Geometrically, the distance between these two hyperplanes is $\frac{2}{\|\mathbf{w}\|}$, so to maximize the distance between the planes we want to minimize $\|\mathbf{w}\|$. As we also have to prevent data points from falling into the margin, we add the following constraint: for each i either

$\mathbf{w} \cdot \mathbf{x}_i - b \geq 1$ for $\mathbf{x}_i$ Of the first class

Or

$\mathbf{w} \cdot \mathbf{x}_i - b \leq -1$ for $\mathbf{x}_i$ Of the second.

This can be rewritten as:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \quad \text{for all } 1 \leq i \leq n.$$

We can put this together to get the optimization problem:

Minimize (in $\mathbf{w}, b$)

$$\|\mathbf{w}\|$$

Subject to (for any $i = 1, \ldots, n$)

$$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1.$$

There is many optimization for SVM and mathematical forms, but generally the main goal is to Maximum margin hyperplane and margins for both classes as described above.

In the other hand many researches compare support vector machine with other popular algorithms and find that SVM is preferable in the performance, in the words of (Basu et al. 2002) compare SVM in text classification problem Artificial Neural Network Algorithm (ANN) and precision indicate significantly differences in the performance of the SVM algorithm over the ANN algorithm and of the reduced feature set over the larger feature set. Furthermore (Joachims 1998) empirical evidence that SVMs are very well suited for text categorization and proof its faster than K-NN. In addition (Pilászy 2005) apply SVM in text classification and discuss aspects of SVMs are covered that reflect why they are suitable for text classification. In the words (Corresponding et al. 2009) find also SVM is one of the most efficient linear machine learning algorithms.

Thus SVM is suitable machine learning algorithm in text classification, and it meets the requirements for linearity and supervised learning, moreover many studies indicates SVM performance beats many popular algorithms in performance. Thus we use it to perform learning and classification.

3.5 Evaluation Methods

Many methods validate and evaluate the performance and results for the classifier, in the below paragraphs we introduce cross validation and confusion matrix.

N-fold cross-validation (holdout): The available data is partitioned into n-equal-size disjoint subsets. Use each subset as the test set and combine the rest n-1 subsets as the training set to learn a classifier. The procedure is run n times, which give n accuracies. The final estimated accuracy of learning is the average of the n accuracies. 10-fold and 5-fold cross-validations are commonly used. The error estimates are averaged to yield an overall error estimate. For example we have a data set with size n, therefore we need to create a K-fold partition of the dataset as shown in Figure 3.5

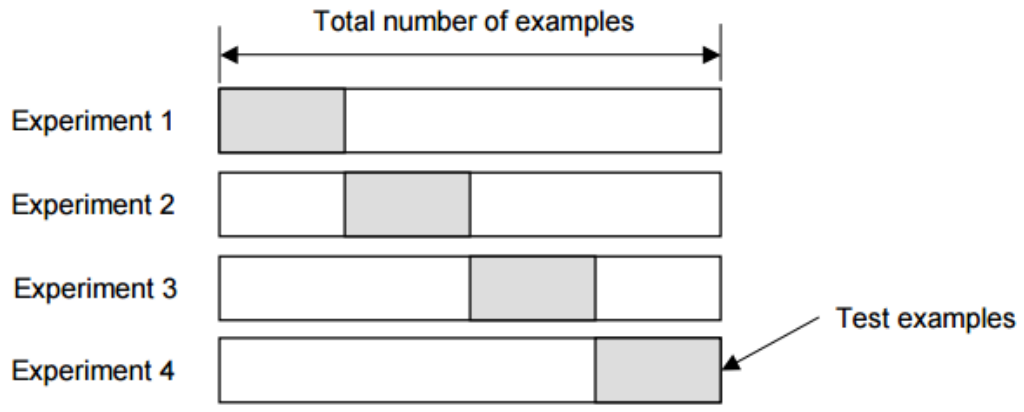

Figure 3.6 Perform K-1 Fold in each Experiment to Validate the Dataset Overall Error

The dataset n for each of K experiments, use K-1 folds for training and the remaining one for testing g K-Fold Cross validation is similar to Random Subsampling. The advantage of K-Fold Cross validation is that all the examples in the dataset are eventually used for both training and testing.

3.5.1 **Confusion matrix**

Confusion matrix is a useful tool for analyzing how well your classifier can recognize data of different classes. Confusion matrix for two –class classification problem is:

Table 3.1 Confusion Matrix

True Class			
Negative	Positive		
False Positive (FP)	True Positive (TP)	Positive	Class Predicted
True Negative (TN)	False Negative (FN)	Negative	

- **True positive (TP)** refer to positive instances that correctly labeled the classifier.
- **True negatives (TN)** refer to negative instances that correctly labeled the classifier.
- **False Positive (FP)** are the negative instances that were incorrectly labeled.
- **False Negative (FN)** are the positive instances that were incorrectly labeled.

3.5.2 **Performance Measures**

From the matrix we can estimate precision, accuracy and recall that will evaluate the performance of the classification. In chapter 5 we introduce how to apply and evaluate the results practically. In addition we have estimate the following to evaluate the performance:

Precision *p* is the fraction of retrieved documents that are relevant to the query. Formulated as the number of correctly classified positive examples divided by the total number of examples that are classified as positive.

$$Precision = \frac{number\ of\ true\ positive}{total\ number\ of\ classified\ positive} = \frac{TP}{TP + FP} \quad [1]$$

Recall *r* is the fraction of the documents that are relevant to the query that are successfully retrieved. Formulated as the number of correctly classified positive examples divided by the total number of actual positive examples in the test set.

$$Recall = \frac{number\ of\ true\ positive}{total\ number\ of\ actual\ positive} = \frac{TP}{TP + FN} \quad [2]$$

F-measure A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F = 2 \times \frac{Precision\,.\,Recall}{Precision + Recall} \quad [3]$$

There are several reasons that the F-score can be criticized in particular circumstances due to its bias as an evaluation metric

Accuracy in the word of (Huang & Ling 2005) The predictive ability of the classification algorithm (or error rate, which is 1 minus the accuracy). Formulated as the number of correctly classified positive examples and the number of correctly negative examples divided

$$Accuracy = \frac{number\ of\ true\ positive + number\ of\ true\ negative}{total\ number\ of\ actual\ postive\ and\ negative}$$

$$= \frac{TP + TN}{TP + TN + FP + FN} \quad [4]$$

In conclusion, the chapter presents the background and techniques used in the research. It started with description of Twitter and tweets and general properties that may be utilized to support classification process, then it take over the source to collect the data from twitter by different API's and what is useful to use in the research problem. It discussed the difference between the supervised learning and the unsupervised learning and our justifications to use supervised learning. Moreover we introduced the SVM algorithm as supervised machine learning algorithm and how it is suitable for the research problem. Finally it presented the evaluation methods and metrics to evaluate the classification results.

Chapter 4
Approach for Identifying Political Topics from Tweets

In this chapter we present the design of the approach for identifying Palestinian political topics from tweets. The chapter is divided into three section, the first section is a general overview about the learning and phases to identify the tweets, while the second section discuss the learning phase and the data preparation, the third section describe the classification and the evaluation phase.

4.1 Overview

Basically, supervised machine learning algorithms required two main stages 1) preparing labeled training data set and 2) applying the algorithm with the training data to classify new unlabeled data. In Section 3.4 we justify why supervised learning is suitable to the research problem.

Our approach follow the main idea of the supervised learning, therefore it illustrate the classification procedure with two phases as the following:

Phase 1: prepare training data and manual or semi manual set the labels, include collecting the data and define the related properties with formulation described in Section 4.2.

Phase 2: performing classification with machine learning algorithm, include the evaluation for the results, phase 2 is described in section 4.3.

However the steps looks straightforward we face many challenges to decide how the approach defined with reasonable justification depends on previous studies or a real experiment.

4.2 Preparing Training Data Set

Classification based on supervised machine learning required suitable predefined dataset to perform the learning stage for the machine learning algorithm and maintain classification the new unlabeled data, thus we perform the following steps to prepare the dataset:

1. Collect tweets: in this step we collect tweets and create the dataset using twitter API and topsy API as described in Section 3.3.

2. Define the features and the properties and considered in the classification procedure.

3. Transform data into understood format to be used as input for the second phase, related to classification.

4. Get part of data to be initial labeled and manual or semi manual classified and used in training.

5. Label part of data according to Arabic Palestinian political topics.

After perform the 5 step the training data are ready to enter the algorithm learning stage. However we may continuously made modifications on the training data set such as size in order to enhance the results, below figure 4.1 shows conceptual diagram for preparation phase. In the next paragraphs we discuss each step in details.

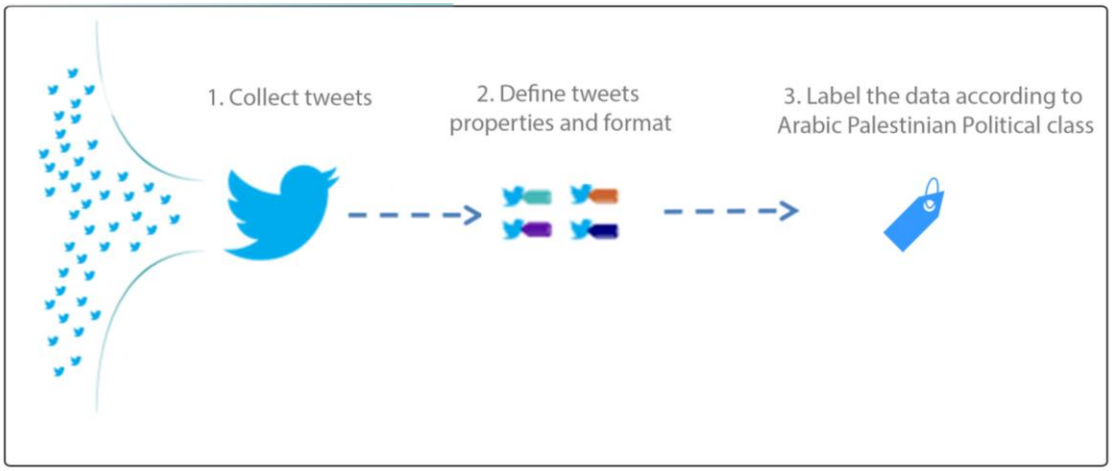

Figure 4.1 Preparing Training Dataset with Properties

4.2.1 **Collecting Tweets**

Collect tweets is the first step to prepare the full data set that part of it is labeled initially for the learning stage and the other part classified in the next phase, we need to take the following points into account:

- Source to collect tweet within specific criteria, therefore twitter is an open space and required criteria definition to select tweets.

- Select type of tweets for collection. However tweets probably have variant contexts and languages.

We collect tweets using topsy API as REST discussed in Section 3.3, we try to use the original twitter API but the data was limited because we filter the tweets according to the scope for the Palestinian Arabic tweets and twitter API response for the tweets within previous 6 to 8 days only, therefore we look to find bulk API that have feature of responding to previous tweets, we find topsy API as good source and provide reasonable number of data could be used in machine learning.

In addition the API need to have set of variables in order to get the exact tweets. Thus we have defined the variables of location (latitude, longitude), language, and common words. However setting this variables allow the API to get the tweets that we look to classify. In chapter 5 we introduce more details about the experiment.

4.2.2 **TwitterPalPol**

We have develop a PHP script to collect the tweets throw the previous API'S, TwitterPalPol collect the tweets and save it in files, therefore the files are assembled incrementally in a single data warehouse were the data are prepared and cleaned to be ready to the learning, we have discuss TwitterPalPol in Section 5.1

The second part is to create the learning data set and provide the selected algorithm, and therefore we need to distinguish between the learning types. In the next section present the types of learning and what is suitable to the research problem.

Figure 5.2 show the pseudo code for TwitterPalPol collecting tweet logic.

```
For radius := 1000 to 5000 do
  For tweetDayCounter := 0 to 8 do
    For pageCounter := 0 to 50 do
```

```
       If tweeter id already exist in the repository then
        Continue looping;
       if max id is not null then
        if current id not equal to max id then
           set since id to max id
    set query operator to the keywords;
    set query operator date to today – tweetDayCounter;
    set geolocation code to the center of Palestine lat/long and set
the radius distance to radius;
    set count parameter tdo max;
    set language to Arabic;
    set result to mixed;

    request twitter with parameters setting (query operator +
geolocation + count + language + result type + paging paramters)

    set twitterResponse from the request;
    if twitterResponse contain tweets then
for tweetCounter:= 0 to twitterResponse size do
          format twitterResponse[tweetCounter];

          if the current twitterResponse[tweetCounter]equal to the
twitterResponse size – 1 then
              set max id to twitterResponse[tweetCounter] id;

       end;
       save the current twitterResponse and add it to Repository;
     end;
   end;
end;
```

Figure 4.2 Psudo Code of TwitterPalPol to Collect Tweets

4.2.3 **Properties Definitions and Format**

The properties are important regarding to creditability filterfor the tweets in order to remove fake tweets. We get the general properties for the tweet, user, and profile. Each tweet text has the associated properties as a row because the classification is performed for each tweet individually, thus the properties support filter creditable tweets that will be considered in classification or not. In Table 4.1 shows sample of formatted data.

Table 4.1 Sample of Formatted Dataset and Properties

id	text	hashtags	created_at	user_created_at	verified	profile_location	description	followers_count	favourites_count	timestamp_ms	listed_count	tweets_count
4.82487E+17	الإحتلال يخشى عملية كبيرة من غزة... موقع والا العبري/ قيادة فرقة غزة تخشى من تنفيذ عملية كبيرة إنطلاقاً من القطاع... http://t.co/JjASxZ0ssF	NULL	Fri Jun 27 11:33:11 +0000 2014	Fri Jun 08 08:27:04 +0000 2012	FALSE	NULL	NULL	41	0	0	3	9976
4.8249E+17	@Sousou_Officiel , وفلسطسن وخاصة غزة تهديك التحية والسلام :)	NULL	Fri Jun 27 11:46:20 +0000 2014	Sun Jul 14 14:29:01 +0000 2013	FALSE	gaza	مبدئي ان الهوان لغيرنا والعز لديني ولبلادي ولامتي ولنا لا نستسيغ الذل او نرد الردى فالموت في زمن الهوان مرادنا	338	4	0	0	2796
4.8249E+17	أسير فلسطيني من غزة يرزق بتوأم عن طريق النطف المهربة - http://t.co/Xj2m5PhyNt #غزة	غزة	Fri Jun 27 11:46:28 +0000 2014	Thu Apr 17 10:29:12 +0000 2014	FALSE	NULL	موقع اخباري، ثقافي، اجتماعي	33	0	0	0	9926
4.8249E+17	الفقر والحصار يزيدان الخناق على "غزة" بحلول رمضان http://t.co/h5wevVLzAj	NULL	Fri Jun 27 11:46:44 +0000 2014	Wed Feb 02 12:47:59 +0000 2011	FALSE	تونس	NULL	515	2	0	12	194017

4.2.4 **Credibility Filter**

Perform credibility check are important to save effort of classify fake tweets. Before run classification we estimate the source user and the tweet itself as we can see in figure 4.3.

The credibility filter is applied on the collected dataset before performing the classification to produce another dataset with creditable content.

However the first step is to check if the twitter account is verified, because this is a confirmation sign from twitter prove that the user is not a fake user and all his tweets will also be actual, if not verified we need to check the number of followers, favorites and following if they meet the normal ranges, we evaluate the ranges depend on the data set we collect for example we take the range for number of followers from the Palestinian political tweets that classified in the learning dataset in manual classifying collected tweets of political or not.

If the ranges of counts are not within the acceptable criteria it will be classified not creditable, therefore it will not consider in the classification. In contrast if any of the checks meet the criteria, we remove all extra words like links and special character to estimate the tweet actual content length if it is larger than three words the tweet will be considered as creditable and move to the classification.

However we can conclude the credibility filter steps as following:

1. Check if the tweet source user is verified from twitter or not, if the user is verified we move to the next step.

2. Remove any stopping words and clean content.

3. Check if the tweet content is less than three words or not, if the tweet content more than three words the tweet is set as creditable tweet.

4. If the source user is not verified we get the number of followers and favorites and following to check if it meets the averages for the creditable user.

5. If it meets the general average for real user we follow the same process from step 2 to 4.

We have use the credibility filter in the collected dataset to filter the data before we perform the classification, in chapter 5 we introduce both results with and without credibility filter.

4.2.5 **Data Labeling**

The number of collected tweets has to and for limit, however we have to define two different data sets the first one training data set for learning and the second one test data for classification. The availability of the data in the specific classification class is another important factor that affect the results. In our problem we set 30% for training data set and 70% for the test dataset, this percent

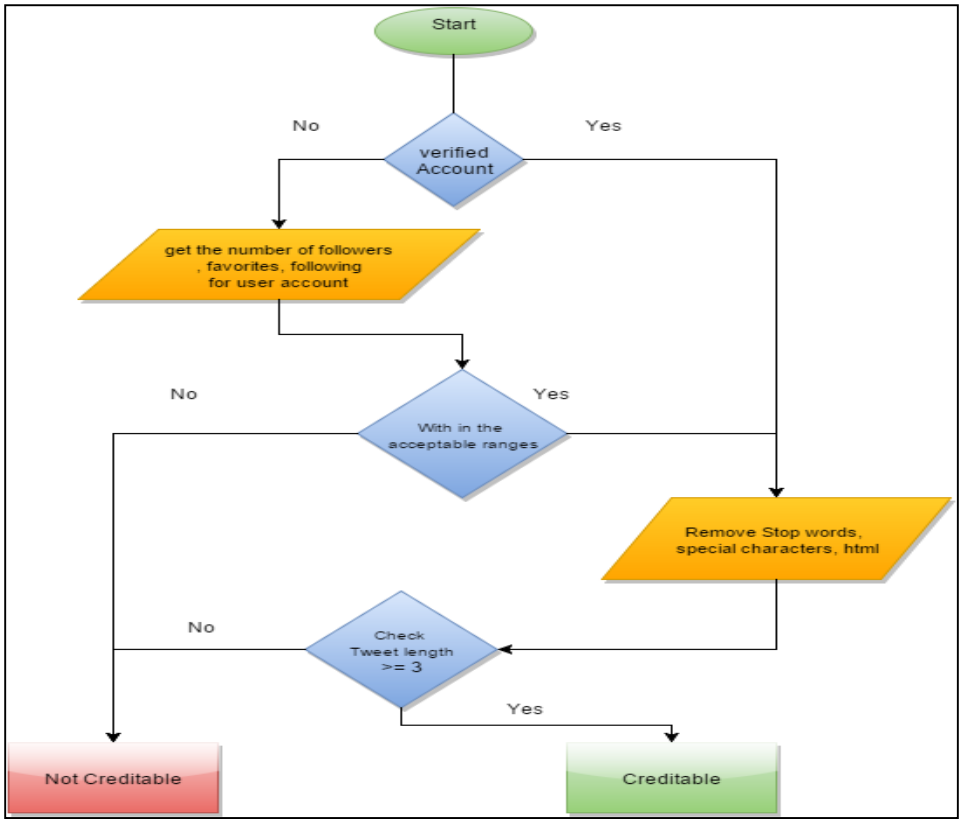

Figure 4.3 Credibility Check to remove Fake Tweets

Could be changed according to different cases. In another hand we have create two different dataset, the first dataset is maintained without creditability filter and the other one with the filter.

4.3 **Classification and Results**

The next stage after prepare the train and test datasets is perform classification and enhance the results. This phase revolves the classification steps. The phase consist of 6 steps as following:

1. Get the prepared dataset (unlabeled and labeled) data.
2. Select proper machine learning algorithm that has ability to learn and classify.
3. Adopt the algorithms to handle the required input and get output.
4. Apply the predefined features and train data set into the modified algorithm.
5. Classify the unlabeled data set into political label or not, and evaluate the result.

The conceptual diagram in Figure 4.4 below shows the main steps in classification phase.

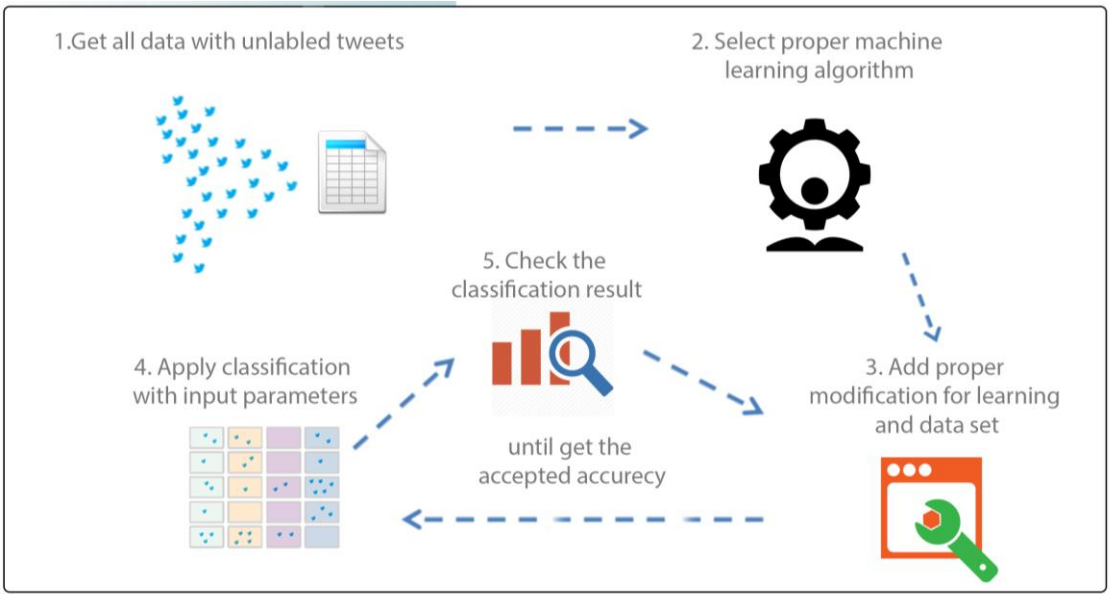

Figure 4.4 Classify Unlabeled Tweets and Evaluation

4.3.1 **Get Unlabeled Tweets**

The dataset is prepared before in the first phase of learning therefore part of data is used for learning and labeled initially while the other part is going to be classified in this phase.

4.3.2 **Machine Learning Algorithm Selection**

Many types of machine learning algorithms have different mentality and specification. Our approach look for algorithm has good performance and accuracy, and can classify the data into single label into two categories (Political Palestinian or not), thus we use SVM as many studies find its one of the most efficient algorithms specially in text classification,

we introduce set of studies and more justification of why to use this algorithm in Section 3.5.

4.3.3 Add Modification for Learning and Dataset

There is no available standard for how big of size dataset required to maintain the learning with good performance of classification, further in the learning how much true class required and false class in the manual labeling to get better performance, in another hand how big the dataset required for testing to judge on the results, therefore we follow the general rules and made the modifications continuously to improve the as there is no standard to follow.

4.3.4 Apply the Classification with the Input Parameters

The execution is performed in this step as the data already prepared, the data are processed to extract the most valuable words from each record using TF-IDF that intended to reflect how important a word is in the document, therefore the tweets transformed into words that indicate whether the tweet belongs to the political Palestinian class or not, the next is to apply the learn data (labeled data) into the machine learning algorithm and the new data (unlabeled data) to the algorithm to execute and classify the data, after execution we need to determine the classification performance and accuracy.

4.4 Check the Classification Results

After execute the classification we need to check the results, Section 3.6 shows the evaluation methods and measures, the main evaluation method is the confusion matrix as it represents the main source for the measures, the confusion matrix is very helpful for two class classification, therefore our problem is classifying tweets into two classes (political Palestinian tweet) or (not political Palestinian tweet), in the other hand the recall and precision are important, recall is present the ability for our approach to find the positive cases from the dataset, while precision is the accuracy of defining the class of the tweets, we try to improve result throw get back to the point number 3 as figure 4.2 shows.

This chapter presented the main framework for experiment, therefore it shape the approach for identifying political Palestinian Arabic tweets, the chapter describes the learning phase include the preparation for the data with formatting the properties and data labeling. Moreover it describes the creditability filter and how we use it to remove fake tweets. Further it take over the classification phase, including the selection for the machine learning algorithm and the improvement stages to reach to the acceptable results, thus the next chapter will apply the experiment on the approach.

Chapter 5
Implementation and Experiments

In this chapter we implement the proposed approach through some experiments and describe the details of the practical application, further the chapter includes the evaluation and results justification and classification samples.

5.1 Collecting Tweets

We use TwitterPalPol to collect certain tweets regarding to our classification criteria, TwitterPalPol is a web application written in PHP language that connect to twitter REST API using JSON format for data exchanging and OAuth for authentication.

First of all we create twitter application on twitter development platform in order to enable TwitterPalPol access twitter over OAuth. After authenticate the app, we look to find tweets that used in training data set and classification unlabeled tweets, therefore we have define two types of parameters, the first one is fixed parameters used in each request, and the second one is dynamic that changed in each request, the below list contain the exact definition for the fixed parameters:

1. **The Query Operator (q):** In TwitterPalPol the query operator contain words that related to Palestine as a location and well-known words in the same area, we have collect words gradually, therefore we start with few words and get trend words from the collected tweets, the following table contain sample of words we have use to collect tweets related to Palestine. We try to create dynamic part in the query operator is the tweets date, as twitter enable to get historical tweets from 6 to 8 days. However we use to collect tweets from the day of running the collection process until the 8th day.

Table 5.1 Query words samples

فلسطين	غزة	الصهانية
الضفة الغربية	فلسطين المحتلة	يهود
القدس	إسرائيل	قطاع غزة
عملية طعن	كتائب القسام	كتائب الأقصى
سرايا القدس	الرئيس أبو مازن	هنية
نتنياهو	دحلان	معبر رفح
السلطة الفلسطينية	حماس	محمد الضيف
الأسرى	قناة الأقصى	الجهاد الإسلامي
المفاوضات	ميناء غزة	العصف المأكول
الجرف الصامد	الرصاص المصبوب	حرب الفرقان

2. **Language:** the lang parameter restricts tweets to the given language. We set the lang param to Arabic to collect Arabic tweets.

3. **Geolocalization:** We specified the parameter geocode =31.603726,34.911230 this geo location approximately represent the center of Palestine, in twitter API, unfortunately this parameter has limit the response, therefore we have decline the parameter to get more tweets in the topsy API, while we use it in Twitter API.

4. **Number of tweets:** the "count" parameter define the number of maximum tweets to be collected, we use the maximum limit, which is 100 tweets. This is also was a limitation, therefore we have use topsy as bulk API.

5. **Result type:** the result type parameters can be defined as popular that collect popular tweets only, recent that collect the new tweets and mixes contain both, we have set it to be both.

6. **Radius**: in the geo we set the exact location to collect tweet using longitude and latitude while we set another dynamic parameter, which is the radius. The radius defines the cycle distance from the point of map (lat,long). However in each request we attempt to increase the radius to get more tweets (in twitter API).

7. **Max ID & Since ID:** use both max_id & since id parameters correctly minimize the amount of redundant data they fetch and process, while retaining the ability to iterate over the entire available contents of a timeline. In figure 5.1 we can fine the mission for max id parameter. However we set the since id is the last id for the current request iteration in order to start with it for the next iteration and get the next tweets.

Similarly, we gather tweets from topsy bulk twitter API discussed in Section 3.3 with pretty different in parameters and values, therefore no radius and max id parameters passed.

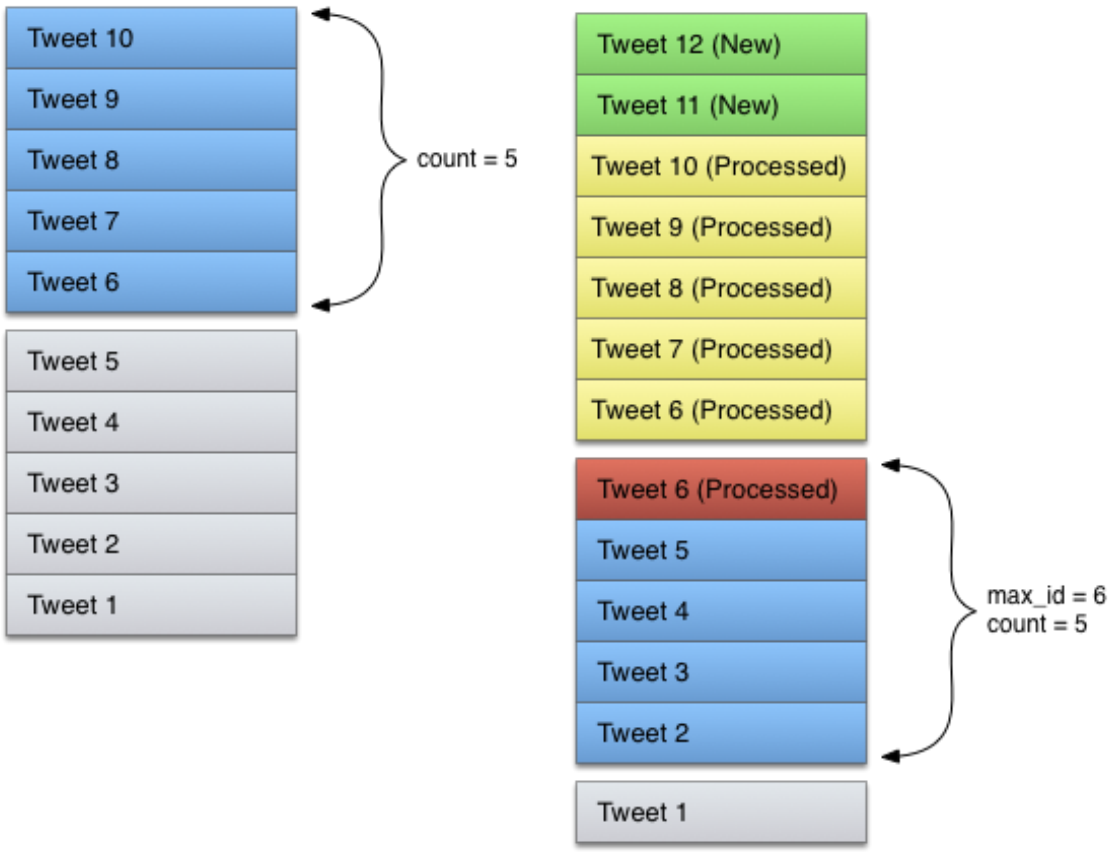

Figure 5.1 Max id used in Paging to Move from One Page to the Next

The example below contains a request API URL with parameters for Twitter API and snap for response via browser. It consist of the main key words and the geo code to define the location and tweet language also the max id and result type. Snap of response showed in Figure 5.3

```
https://api.twitter.com/1.1/search/tweets.json?q=20%فلسطينOR%2020%غزة
فلسطين%20المحتلة%22%20%2220%الضفة%20الغربية%22%20%2220%OR%20الصهانية%2020%
20OR%20قطاع%20غزة%22%20%2220%OR%20إسرائيل%2020%OR%20القدس%2020%OR%20يهود%2020
كتائب%20الأقصى%222%20%OR%كتائب%20القسام%22%2220%20%OR%20عملية%20طعن%22%2220%
20%2OR%20هن%20OR%20الرئيس%20أبو%20مازن%22%2220%20%OR%20سرايا%20القدس%22%2220%20%OR
ORمعبر%20رفح%22%2220%20%OR%20الحرب%2020%OR%20دحلان%2020%OR%20نتنيا هو%2020%OR%20ية%20
%20%2220%22%20الفلسطينية%20السلطة%20%2020%OR%20حماس%20%2020%OR%20محمد%20الضيف%22%20%2220%OR
%20%2020%OR%20الجهاد%20الإسلامي%22%20%2220%OR%20قناة%20الأقصى%22%20%2220%OR%20الأسرى%2020%
20%المفاوضات%20%2220%22%20%OR%20غزة%20ميناء&geocode=31.603726,34.911230km&lan
g=ar&count=100&result_type=mixed&max_id=637655299453640704
```

```
object(stdClass)[1249]
  public 'iso_language_code' => string 'ar' (length=2)
  public 'result_type' => string 'recent' (length=6)
public 'created_at' => string 'Sat Aug 29 15:54:19 +0000 2015' (length=30)
public 'id' => float 6.376545221946E+17
public 'id_str' => string '637654522194604033' (length=18)
public 'text' => string 'بـ 3000 قيمة سهم اطعام 250 حاجا من غزة يكلف ..
0599680769 فمن أحب أن يبادر لإكرام ضيوف الرحمن فليتواصل معي على الخاص او واتس اب
رت' (length=230)
public 'source' => string '<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>'
public 'truncated' => boolean false
public 'in_reply_to_status_id' => float 6.3727042212424E+17
public 'in_reply_to_status_id_str' => string '637270422124236800' (length=18)
public 'in_reply_to_user_id' => int 400240425
public 'in_reply_to_user_id_str' => string '400240425' (length=9)
public 'in_reply_to_screen_name' => string 'jhelles' (length=7)
public 'user' =>
  object(stdClass)[1250]
    public 'id' => int 400240425
    public 'id_str' => string '400240425' (length=9)
    public 'name' => string 'جهاد حلس' (length=15)
    public 'screen_name' => string 'jhelles' (length=7)
    public 'description' => string 'رئيس قسم تمويل المشاريع في الادارة العامة للزكاة. واتس اب: 00970599680769' (length=
    public 'url' => string 'https://t.co/e5zbNCGLa7' (length=23)
    public 'entities' =>
      object(stdClass)[1251]
        ...
    public 'protected' => boolean false
    public 'followers_count' => int 112786
    public 'friends_count' => int 260
    public 'listed_count' => int 548
    public 'created_at' => string 'Fri Oct 28 18:25:50 +0000 2011' (length=30)
    public 'favourites_count' => int 3827
    public 'utc_offset' => int 10800
    public 'time_zone' => string 'Riyadh' (length=6)
    public 'geo_enabled' => boolean true
    public 'verified' => boolean false
    public 'statuses_count' => int 10128
    public 'lang' => string 'en' (length=2)
    public 'contributors_enabled' => boolean false
    public 'is_translator' => boolean false
```

Figure 5.2 Snap of Response include Tweet and Related Properties for Twitter API

The example below is a request API with parameters for Topsy API and snap for response via browser. It consist of the words to retrieve tweets contain same words and also some other parameters need to authorize request such as API Key. Snap of response showed in Figure 5.4.

http://api.topsy.com/v2/content/bulktweets.json?q=20%فلسطينOR%2020%غزةOR%2020%الصهاينةOR%20%2220%22%الغربيةالضفة20%OR%20%2220%22%المحتلةفلسطين20%OR%2020%يهودOR%2020%القدسOR%2020%إسرائيلOR%20%2220%22%قطاعغزة20%'&apikey=09C43A9B270A470B8EB8F2946A9369F3&include_enrichment_all=1

```
[71] => Array
    (
        [text] => Jaz: حزب الله يعتبر إسرائيل عاجزة عن تنفيذ تهديداتها
        [id] => 895169999
        [id_str] => 895169999
        [timestamp_ms] => 0
        [favorite_count] => 0
        [created_at] => Fri Aug 22 04:34:48 +0000 2008
        [screen_name] => pic_rss
        [profile_location] =>
        [description] =>
        [followers_count] => 0
        [favourites_count] => 0
        [verified] =>
        [user_id] => 8766012
        [tweets_count] => 0
        [time_zone] => Santiago
        [listed_count] => 0
        [user_name] => pic_rss
        [user_created_at] => Sun Sep 09 14:59:51 +0000 2007
        [hashtags] =>
    )
```

Figure 5.3 Snap of Response include Tweet and Related Properties for Topsy API

5.2 Preparing Data sets and Format

In this section we discuss the collected datasets and how we split the data for the learning and the data for the testing. Further the section describes the schema format for each record in the dataset and the meaning of every property included.

5.2.1 Data Size

We have collected 6408 tweets that are related to the keywords we use in defining the criteria of collection for the API as mentioned in the previous section. However we have classify the data into do sets training set and validation set, we follow the traditional style of splitting the data approximately 30% for learning and 70% for validation, thus the training data set 1752 record.

In the other hand we have create the creditability filter as mentioned in section 4.3, therefore the data set size become 2776 and 963 for training data set.

We have evaluate the two data set, the first one without creditability filter and the other with and get the results as discussed in the next section.

5.2.2 Properties Definitions

In the previous section we attempt gathering tweets from twitter APIS, therefore we need to format the collected data into proper format. The formatted properties based on each tweet independently, thus the classification are operated on each tweet separately.

In Table 5.2 we can find the set of properties and the description for each one:

Table 5.2 Tweet Properties Description

Tweet Property	Description	Type
id_str	The id for the tweet on twitter used to identify the tweet.	Integer
text	Tweet text content.	String
followers_count	The number of users follows the user who writes the tweet.	Integer
favourites_count	The number of user's favorite tweets for the user who writes the tweet.	Integer
verified	Profile verification status for the user who writes the tweet.	Boolean

Hashtags	Contain the hashtags inside the tweet text	String
created_at	The date of tweet	Date
user_created_at	Date for creating for the user who writes the tweet.	Date
listed_count	Number of the people list the tweet	Integer
tweets_count	Number of tweets user have.	Integer
time_zone	The time zone for the place tweet was submitted.	String

We use the properties to identify our equation for classification process. In Table 5.2 we show a sample of actual tweet and the result for each property.

Table 5.3 Example of Actual Tweet and Properties

Tweet	**.. نساوم ولم نفرط لم .. فلسطين# لنكبة 67 الذكرى مايو/ أيار 15 اليوم لأرضنا محررين فاتحين سنعود #كالشمس_حق_العودة http://t.co/AouPCeCRdW**
Tweet Property	**Value**
id_str	599056576817733632
text	محررين فاتحين سنعود .. نساوم ولم نفرط لم .. فلسطين# لنكبة 67 الذكرى مايو/ أيار 15 اليوم لأرضنا #كالشمس_حق_العودة http://t.co/AouPCeCRdW
profile_location	null
location	غزة / فلسطين
description	بوست نون و توك الجزيرة في إعلامي ناشط .. غزة أسكن المحتلة عسقلان من فلسطيني لاجئ
followers_count	1560
favourites_count	1076
verified	null
hashtags	كالشمس_حق_العودة,فلسطين
created_at	Mon Jun 29 21:29:05 +0000 2015
user_created_at	Fri Sep 21 21:50:10 +0000 2012
listed_count	0
tweets_count	2796
time_zone	Palestine

The gathered tweets are prepared as table 5.3, therefore all tweets are saved in tweet repository to be processed for the learning and classification process.

In the other hand we have clean the text from stop words and links that will not considered in the classification.

5.3 **Model Training**

The training data set was initially labeled manually, therefore we have labeled each tweet with Boolean class (true or false) that it's an Arabic political Palestinian tweet or not, 389 of data records are labeled as true, while the 728 is false, moreover the filtered creditability training data set is 200 true and 548 is false. But this is was too small and the results of classification was not too good, thus we try to optimize learning throw using the first classification result to feed a second stage learning to be 729 false class and 1023 true class, 574 false class and 389 true class for creditability filter training dataset. That's mean we have initially label part of data manually and use machine learning to get more labeled data to be used in second learning stage.

After labeling the data we start the process of learning, Figure 5.3 show the learning components on rapid miner studio.

First of all we retrieve the train data set from the formatted excel document, and then we set the role to identify each tweet with id, the next is to perform processing for each document as shown in Figure 5.3, the process is conducted by extracting the html content and remove the stopping words and remove documents less than 3 words after the process, then we estimate the valuable words by using TF-IDF for the documents.

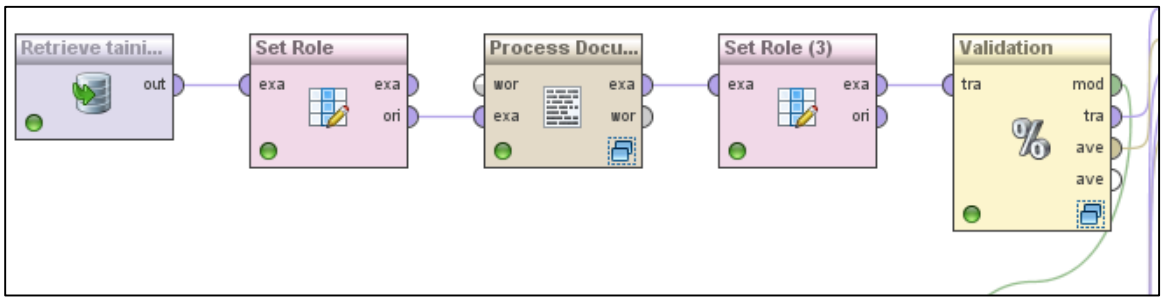

Figure 5.4 Learning Process

Now the documents are ready to submit as input for the machine learning algorithm, we set the role for the classification label (already classified manual) and remove missing attributes then prepare the labeled to be binomial and set the input for the machine learning algorithm (SVM) as shown in Figure 5.4.

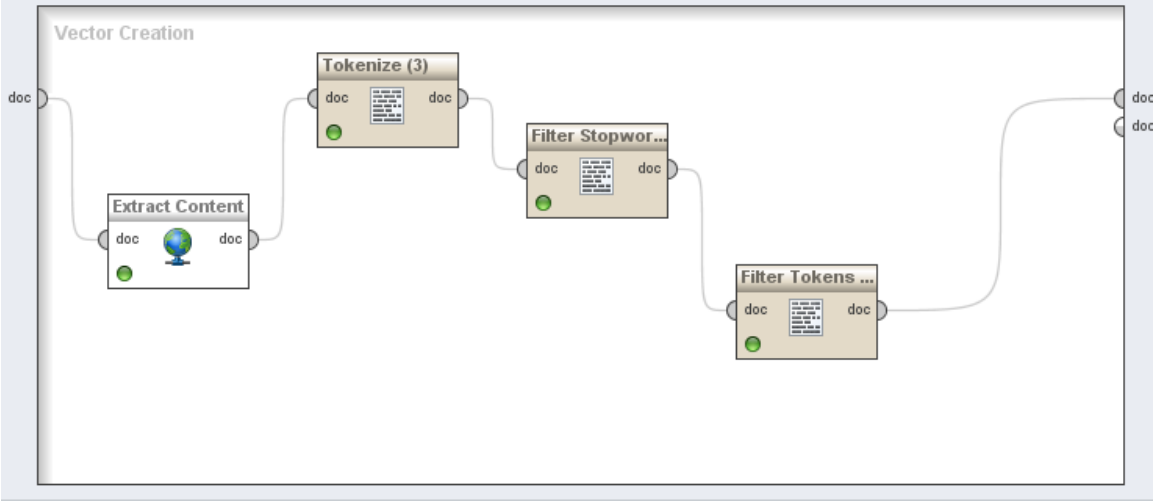

Figure 5.5 Processing Document

In addition we estimate the performance of learning using confusion matrix that previously mentioned in Section 3.6. However the learning process is evaluated and performed.

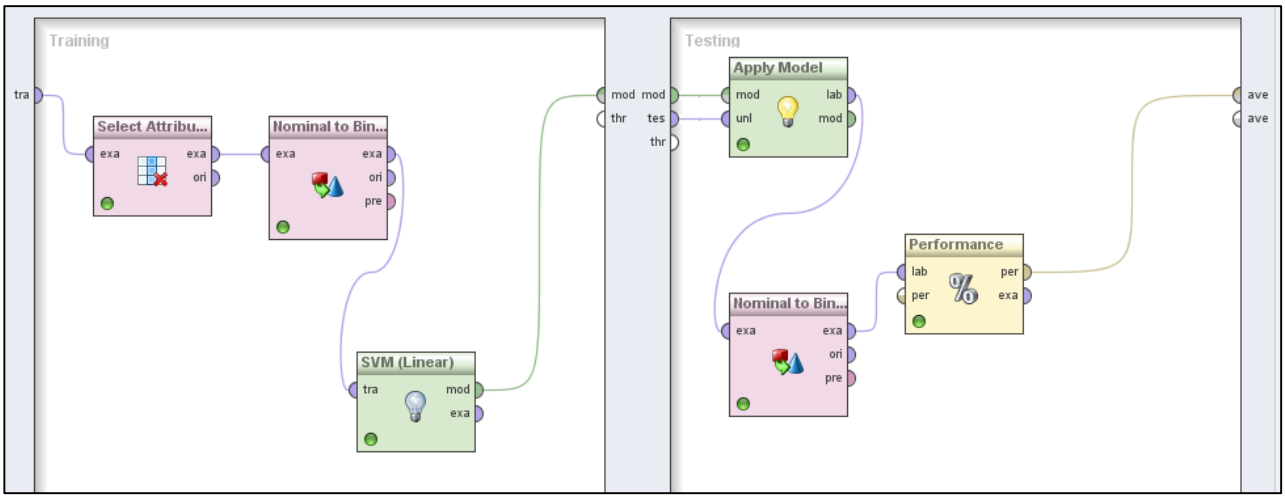

Figure 5.6 Training and Validation Process

5.4 **Classification**

At the end of learning process that prepare the machine learning algorithm to label new data.

Similarly, we have follow the same steps in Figure 5.3 of retrieving data to process the document, the different is in setting the label to be the text, then the data is ready to be classified, therefore we apply the learning model on the new data as showed in Figure 5.5.

We apply the same process for both data set with credibility filter, and data set without creditability filter as showed in Figure 5.6.

Moreover the result are conducted in performance vector for each data set, training and classification views for each data set, Figure 5.6 shows snap of classification result. However the results will discussed in depth in the next section.

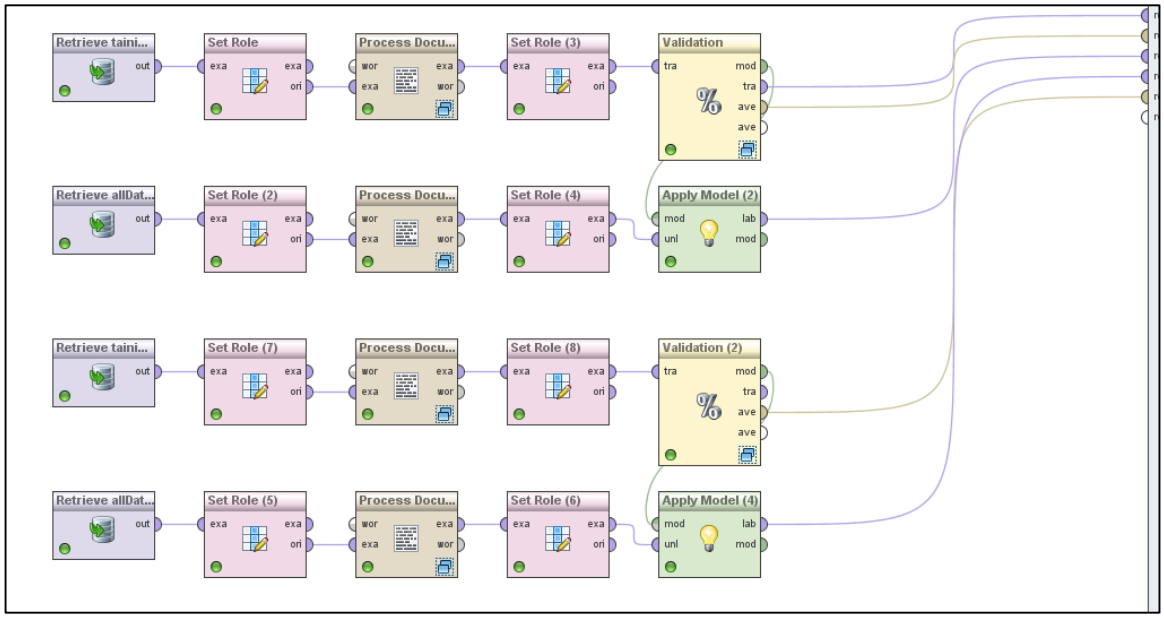

Figure 5.7 Apply Classification

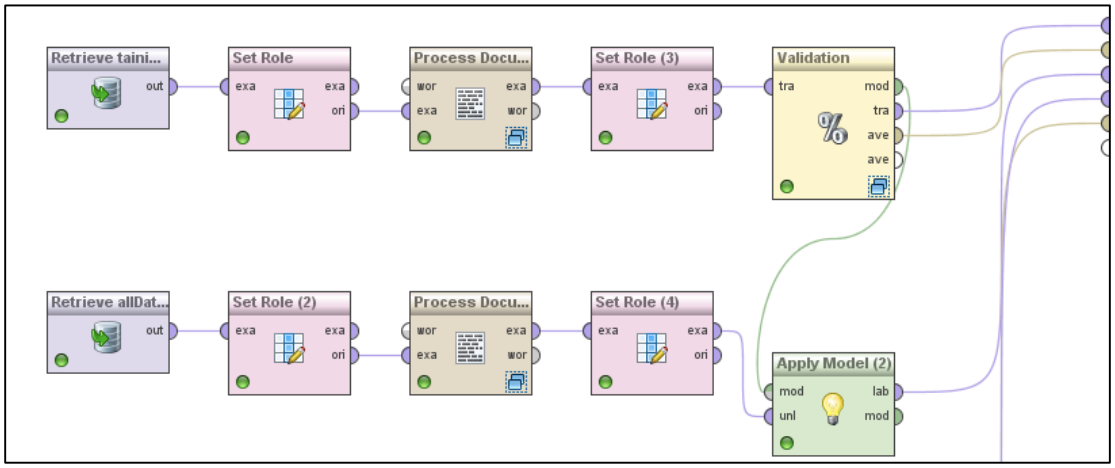

Figure 5.7 Apply Classification for both Data set with Creditability Filter

This chapter presented the implementation and experiments, it take over the phases of the approach describe in Chapter 4, from the learning phase to the classification. In the next chapter we introduce the result of the experiment.

Chapter 6
Results and Evaluation

This chapter presents the results of the experiment and their evaluation with justification. It evaluate two different experiment: 1) classification without creditability filter and 2) classification with creditability filter. Finally it shows some classification samples.

6.1 **Classification Results**

The results are obtained in different perspectives and views, in this section we are going to show, evaluate and justify the results. However we have performed the experiments on two different dataset, the first one without credibility filter and the second with credibility filter.

text	prediction(Classify)	confidence(true)	confidence(false)
نصرين سيار الرماد ومبض ويوشك ضرام الدار بالعودين الحرب مبدؤها كـلام	false	0.280	0.720
يه وندسا فلسطين اليمن مالها انحباسه وطرابلس تشتت حدين والقاهره راحت خلاف الرئاسه خالد غازي الدوسري	false	0.265	0.735
نسخة سفير فلسطين إسلام آباد صاحب الذوق الرفيع حازك لكمية الكحول لصالح السفاره الفلسطينية	true	0.695	0.305
وت بسعر للمليون وحده لتشغيل محطات الكهرباء وتصدر الغاز لمحطات كهرباء إسرائيل بسعر ويتحدثون الدعم	false	0.297	0.703
نريد أصغر رئيسه بلديه فلسطين بشائر عثمان ابده عاما تصلم مهامها كرئيسه لبلديه عالر	false	0.236	0.764
ندوه الكافي بمسجد رابعة العدوية	false	0.353	0.647
نداء أخير أصحاب القلوب الرحيمة والقادمين قريبا الخارج جماعة الخير حواليه وكاله	false	0.298	0.702
نخوض الحرب أرضهم والنصر للثورة السورية قريب باذن الله انتهي الشكر	false	0.280	0.720
نجران حماس بشعله الأعلامي الصغير	false	0.289	0.711
نتحرر ارجع فلسطين وامارس حقوقي الطبيعيه اجرب معلى عندي	false	0.236	0.764
ناقصها برده الدنيا مصممه نفكردا بهزات الحرب	false	0.238	0.762
ناجعا بوقف الحرب اتفاق سياسي الأطراف المتقاتله الداخل تغليب مصلحة الوطن	false	0.287	0.713
نائب الرئيس اليمني يدفي علمه بموعد عوده صالح صنعاء بتاريخ الساعة بتوقيت القدس	false	0.263	0.737
ميداني الطيران يستهدف جنوب قطاع	true	0.508	0.492

Figure 6.1 Snapshot of the Classification Result

6.1.1 **Results without Creditability Filter**

Data set size 6408 record and the 1752 for training 1024 labeled as true and 729 labeled as false, the result from machine learning classification the data set labeled 3120 true and 3288 false. The confusion matrix to evaluate our classifier model performance is produced as the following:

Table 6.1 Confusion Matrix Results for Data set without Creditability Filter

True Class		
Negative	Positive	
24 (FP)	1000 (TP)	Positive
530 (TN)	198 (FN)	Negative

Thus the confusion matrix is constructed, we can perform the following evaluation measures:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [4]$$

$$\frac{1000 + 530}{1000 + 530 + 198 + 24} \times 100 = \ 87.33\%$$

This sign the proportion of the total number of predictions that were correct is 87.33%, the result considered very good because the number of false prediction for both negative and positive are too small compared to the positive predications.

$$Precision = \frac{TP}{TP + FP} \quad [1]$$

$$= \frac{1000}{1000 + 24} \times 100 = \ 97.65\ \%$$

The proportion of the predicted positive cases that were correct is 97.65%, the precision is excellent probably because the learning have many positive cases example that support the algorithm in truly classify positive cases.

$$Recall = \frac{TP}{TP + FN} \ [2]$$

$$= \frac{1000}{1000 + 198} \times 100 = \ 83.47\ \%$$

The proportion of positive cases that were correctly identified is 83.47%, the recall is very good this is probably because the data set is considered large, therefore the ability to detect tweets correctly is decreased when the data become larger.

$$F = 2 \times \frac{Precision\,.Recall}{Precision + \ Recall} \ [3]$$

$$= 2 \times \frac{0.9765 \times 0.8347}{0.9765 + 0.8347} = \ 90.0\ \%$$

F-measure is an indicator that shows the weight of both precision and recall if our system doesn't have a clear goal weather the precision is the important or the recall, therefore in our case we believe that the precision is more important than the recall.

6.1.2 **Results with Creditability Filter**

Data set size 2776 record and the 963 for training 574 labeled as true and 389 labeled as false, the result from machine learning classification the data set labeled 484 true and 2292 false. The confusion matrix is produced as the follows:

Table 6.2 Confusion Matrix Results for Data set with Creditability Filter

True Class		
Negative	Positive	
77 (FP)	312 (TP)	Positive
537 (TN)	9 (FN)	Negative

Thus the confusion matrix is constructed, we can perform the following evaluation measures:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \text{ [4]}$$

$$= \frac{312 + 537}{312 + 537 + 77 + 9} \times 100 = \ 90.8\,\%$$

This sign the proportion of the total number of predictions that were correct is 90.8%, the result considered excellent because the number of false prediction for both negative and positive are too small comparing to the positive predications.

$$Precision = \frac{TP}{TP + FP} \text{ [1]}$$

$$= \frac{312}{312 + 77} \times 100 = \ 80.20\%$$

The proportion of the predicted positive cases that were correct is 80.20%, the precision is very good as it's affected from the number of false positive predication in comparing to the true predication probably because the learning data for true positive cases are limited in compared to the original data set as in the previous result.

$$Recall = \frac{TP}{TP + FN} \quad [2]$$

$$= \frac{312}{312 + 9} \times 100 = \ 97.19\ \%$$

The proportion of positive cases that were correctly identified is 97.19%. The recall is excellent and this because the creditability filter support the algorithm to detect the more tweets related to classification area.

$$F = 2\ \times \frac{Precision\ .Recall}{Precision + \ Recall} \quad [3]$$

$$= 2\ \times \frac{0.8020\ \times 0.9719}{0.8020\ + 0.9719} = \ 87.88\ \%$$

F-measure is an indicator that shows the weight of both precision and recall if our system doesn't have a clear goal weather the precision is the important or the recall, therefore in our case we believe that the precision is more important than the recall.

We can find that the results have look different in the two data set, therefore the accuracy in the data without creditability filter is less than the accuracy with the filter that indicates the correct classification in general in the data with filter is better than the one without filter.

The recall in the first classification is less than the recall with creditability filter, thus the precision is greater than creditability filter precision. However creditability filter support ignore any fake tweets that could deceive the classification algorithm.

Generally, our model has good result of recall and precision to identify the Palestinian political tweets. However we have made perform optimization effort to get this results specially in learning phase. In the next section we introduce sample of tweets classification.

6.2 Samples of Classified Tweets

In this section we presents snaps of tweets, with different cases, the sample represent the original tweet and related properties and how it looks after processing the text and cleaning, moreover the classification result for both creditability filter classification if it's a creditable tweet and the original classification, each tweet describe in a table as below.

Table 6.3 Sample1

Original Tweet	**صور جميلة لغزة قبل الحرب: صورة: http://pms.panet.co.il/online/images/articles/2008/09/15-09-08/3_22.jpg**
Verified User	**False**
Followers	**0**
Tweets Count	**0**
Favorited	**0**
Creditable Tweet	**False**
Tweet Cleaned	**صور جميلة لغزة قبل الحرب: صورة:**
Text Processed	**جميلة لغزة الحرب صورة صورة**
Confidence false	**0.625226**
Confidence true	**0.374774**
Classification	**False**
Creditable Class	**Null**
Justification	**Tweet submitted from not a verified source as the user are not verified and the numbers looks less than credit tweets, the content contain words related to false classification more than true therefore the confidence of false are more than the true and the tweet classified as false (not a Arabic Palestinian political tweet)**

Table 6.4 Sample2

Original Tweet	**RT @AlMayadeenLive: فلسطين: مراسل الميادين: شهيد وعدد من الإصابات في استهداف سيارة من قبل الاحتلال على الطريق الساحلي لمخيم الشاطىء في غزة**
Verified User	**False**
Followers	**87**
Tweets Count	**2767**
Favorited	**2961**
Creditable Tweet	**False**
Tweet Cleaned	**فلسطين: مراسل الميادين: شهيد وعدد من الإصابات في استهداف سيارة من قبل : الاحتلال على الطريق الساحلي لمخيم الشاطىء في غزة**
Text Processed	**فلسطين مراسل الميادين شهيد وعدد الإصابات استهداف سيارة الاحتلال الطريق الساحلي لمخيم الشاطىء**
Confidence false	**0.294243812086325**
Confidence true	**0.705756187913674**
Classification	**True**
Creditable Class	**False**
Justification	**Tweet submitted from not a verified source as the user are not verified and the numbers looks less for credit tweets, the content contain words related to true classification more than false therefore the confidence of true are more than the false and the tweet classified as true.**

Table 6.5 Sample3

Original Tweet	**إسرائيل تتوعد الصحفيين بأسطول الحرية - هددت إسرائيل الأحد الصحفيين الأجانب والمؤسسات الإعلام http://ow.ly/1dt6E3**
Verified User	**False**
Followers	**145**
Tweets Count	**10500**
Favorited	**0**
Creditable Tweet	**True**
Tweet Cleaned	**إسرائيل تتوعد الصحفيين بأسطول الحرية - هددت إسرائيل الأحد الصحفيين الأجانب والمؤسسات الإعلام**
Text Processed	**إسرائيل تتوعد الصحفيين بأسطول الحرية إسرائيل تتوعد الصحفيين بأسطول الحرية هددت إسرائيل الأحد**
Confidence false	**0.288823634364719**
Confidence true	**0.71117636563528**
Classification	**True**
Confidence false (Creditability Classification)	**0.411144817890993**
Confidence true (Creditability Classification)	**0.588855182109006**
Creditable Class	**True**
Justification	**Tweet submitted from a confident source as the user are not verified but the numbers looks good for credit tweets, in both normal classification and creditability classification the content contain words related to true classification more than false therefore the confidence of true are more than the false and the tweet classified as true.**

Table 6.6 Sample4

Original Tweet	**المستفيد الأكبر من إحداث القلاقل والصراعات في البلدان العربية هي إسرائيل وإيران** **http://t.co/qGnkrItHoa ...لن يكون أي مسلم مخلص ولا عربي حر**
Verified User	**False**
Followers	**145**
Tweets Count	**10500**
Favorited	**0**
Creditable Tweet	**False**
Tweet Cleaned	**المستفيد الأكبر من إحداث القلاقل والصراعات في البلدان العربية هي إسرائيل وإيران** **...لن يكون أي مسلم مخلص ولا عربي حر**
Text Processed	**المستفيد الأكبر إحداث القلاقل والصراعات البلدان العربية إسرائيل وإيران مسلم مخلص عربي**
Confidence false	**0.584554980133149**
Confidence true	**0.41544501986685**
Classification	**True**
Confidence false (Creditability Classification)	**0.695281602216227**
Confidence true (Creditability Classification)	**0.304718397783772**
Creditable Class	**False**
Justification	**Tweet submitted from a confident source as the user are not verified but the numbers looks good for credit tweets, in both normal classification and creditability classification the content contain words related to false classification more than false therefore the confidence of true are more than the false and the tweet classified as true.**

Chapter 7
Conclusion and Future Work

In this research we have designed an approach to identify Arabic Palestinian tweets based on text content, the approach consists of two main phases: the first one is to collect and prepare the data set, and the second one to classify and evaluate the results.

We create TwitterPalPol simple web application that collect the data from different twitter api's with specified criteria and cleaning the collected tweets, then format the tweets with its properties into proper format.

Thus the data formatted, we split them into training and testing dataset, therefore we label the training data manually and then we apply the SVM to classify the testing data set, we choose SVM because many studies recommend it as machine learning algorithm to be applied in text classification and it has a great reputation of performance in two classes cases, thus our study is two class classification Palestinian political Arabic or not SVM is suitable and we get promising results in our research.

We have performed the learning and classification on two different datasets, the first was the original dataset while the second was filtered by creditability filter to remove fake tweets depending on the content and the user profile properties. After performing the classification the results was not unsatisfactory as the recall was too low, therefore we try to improve results by making another stage of learning through getting the first classification result feed into learning to maximize training data especially with positive class and adding new data to classify.

However we got the results for the classification of both data sets. We classifying 6048 records is the data set labeled 3120 true and 3288 false and we got the following measures 1) accuracy is 87.33% of total number of our system predictions were correct 2) precision is 97.65 of predicted positive cases were correct 3) Recall is 83.47% of positive cases were correctly identified 4) F-Measure is 90% as weight for both recall and precision.

Additionally we performed the classification on the data set with creditability filter with size 2776 records and the result from machine learning classification the data set labeled 484 true and 2292 false and got the following measures 1) accuracy is 90.8% of total number of our system predictions were correct 2) precision is 80.20 of predicted positive cases were correct 3) Recall is 97.19 % of positive cases were correctly identified 4) F-Measure is 88% as weight for both recall and precision.

Overall, results in the two data sets look close, thus it can be used to classify political Palestinian Arabic tweets because it has sufficient recall, precision and accuracy in general.

As a future work, the research approach can be updated through adding the following:

- Add more classes for classification like sport, health…etc.
- Enable the process to be online as to connect the application and process classification directly.
- Improve the results through add more learn data that helps the machine learning algorithm to increase the results of recall and to classify new data especially when connecting the classification online.

References

Ahmed Ibrahim, M. & Salim, N., 2013. Opinion analysis for twitter and Arabic tweets: A systematic literature review. *Journal of Theoretical and Applied Information Technology*, 56(3), pp.338–348.

Al-Eidan, R.M.B., Al-Khalifa, R.S. & Al-Salman, A.S., 2010. Measuring the credibility of Arabic text content in twitter. In *2010 5th International Conference on Digital Information Management, ICDIM 2010*. pp. 285–291.

At, D. of C.S. at the U. of I., Supervised Learning. Available at: https://www.cs.uic.edu/.

Basu, a, Watters, C. & Shepherd, M., 2002. Support Vector Machines for Text Categorization. *Sciences-New York*, pp.1–7.

Bekkali, M. & Lachkar, A., 2014. ARABIC TWEETS CATEGORIZATION BASED ON ROUGH SET THEORY. *International Journal of Computer Science & Information Technology*, 6(6).

Cheng, Z., Caverlee, J. & Lee, K., You Are Where You Tweet : A Content-Based Approach to Geo-locating Twitter Users.

Ciotec, S., Dascalu, M. & Trausan-Matu, S., 2014. A comprehensive study of Twitter social networks. In *RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference, 2014*. pp. 1–7.

Corresponding, R.E., Rezaei, A. & Minaei-bidgoli, B., 2009. Comparison of Classification Methods Based on the Type of Attributes and Sample Size. *Journal of Convergence Information Technology 4.3*, 94-102. Available at: http://www4.ncsu.edu/~arezaei2/paper/JCIT4-184028_Camera Ready.pdf.

Duwairi, R.M. & Qarqaz, I., Arabic Sentiment Analysis using Supervised Classification.

El-Halees, A.M., 2007. Arabic Text Classification Using Maximum Entropy. *The Islamic University Journal*, 15(1), pp.157–167.

Huang, J. & Ling, C.X., 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3), pp.299–310.

Jazra, K., 2014. 15 stats about social media in the Middle East that you need to know. *social4ce*. Available at: http://social4ce.com/blog/2014/07/01/15-stats-you-need-to-know-about-social-media-in-the-middle-east.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, 1398(2), pp.137 – 142.

Kim, D. et al., 2010. Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users. *Search*, p.4. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.163.7391.

Kourdi, M.E.L. et al., 2004. Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. *Proceedings of COLING 20th Workshop on Computational Approaches to Arabic Script-based Language*, pp.51–58.

Lee, K. et al., 2011. Twitter trending topic classification. In *Proceedings - IEEE International Conference on Data Mining, ICDM*. pp. 251–258.

O'Donnell, C., 2011. New study quantifies use of social media in Arab Spring. *University of washington news blog*. Available at: http://www.washington.edu/news/2011/09/12/new-study-quantifies-use-of-social-media-in-arab-spring/.

Pennacchiotti, M., Democrats , Republicans and Starbucks Afficionados : User Classification in Twitter. , pp.430–438.

Pennacchiotti, M. & Popescu, A.-M., 2011. A Machine Learning Approach to Twitter User Classification. In *ICWSM*. pp. 281–288. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2886/3262.

Piao, S. & Whittle, J., 2011. A feasibility study on extracting twitter users'interests using NLP tools for serendipitous connections. In *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*. pp. 910–915.

Pilászy, I., 2005. Text categorization and support vector machines. *The Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence*, 1.

Sakaguchi, T. et al., 2010. Recommendations in twitter using conceptual fuzzy sets. In *Annual Conference of the North American Fuzzy Information Processing Society - NAFIPS*.

Topsy, 2015. help with topsy.com social search. Available at: http://about.topsy.com/support/search/.

Tsukayama, H., 2013. Users send over 400 million tweets per day. *Washington*. Available at: http://www.washingtonpost.com/business/technology/twitter-turns-7-users-send-over-400-million-tweets-per-day/2013/03/21/2925ef60-9222-11e2-bdea-e32ad90da239_story.html.

Twitter, 2014a. Following people on Twitter. *Twitter*. Available at:

https://support.twitter.com/groups/52-notifications/topics/213-following/articles/162981-following-people-on-twitter.

Twitter, 2014b. New user FAQs. Available at: https://support.twitter.com/articles/13920#.

Twitter, 2015. OAuth. Available at: https://dev.twitter.com/oauth.

Twitter, REST APIs. Available at: https://dev.twitter.com/rest/public.

Twitter, 2014c. Using hashtags on Twitter. Available at: https://support.twitter.com/articles/49309.

Weber, I. & Garimella, K., 2013. #Egypt: Visualizing Islamist vs. secular tension on Twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*. pp. 1100–1101.