Discovering User Attitudes of Business in Twitter Language Feed

By

Ibrahim Mohammed AbdulNabi

Supervisor
Dr. Samir Tartir

This Thesis was Submitted in Partial Fulfilment of the Requirements for the Master's Degree in Computer Science

Deanship of Academic Research and Graduate Studies
Philadelphia University

2013

جامعة فيلادلفيا

نموذج تفويض

أنا ابراهيم محمد صالح عبدالنبي ،أفوضجامعةفيلادلفيابتزويدنسخمنرسالتيللمكتباتأو المؤسساتأوالهيئاتأوالأشخاصعندطلبها.

التوقيع :

التاريخ :

Philadelphia University
Authorization Form

I am, Ibrahim Mohammed Saleh AbdulNabi we authorize Philadelphia University to supply copies of my thesis to libraries or establishments or individuals upon request.

Signature:

Date:

Discovering User Attitudes of Business in Twitter Language Feed

By

Ibrahim Mohammed AbdulNabi

Supervisor
Dr. Samir Tartir

This Thesis was Submitted in Partial Fulfilment of the Requirements for the Master's Degree in Computer Science

Deanship of Academic Research and Graduate Studies
Philadelphia University

2013

Successfully defended and approved on _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Examination Committee Signature	**Signature**

Dr. Samir Tartir , Chairman. _ _ _ _ _ _ _ _ _ _ _ _ _
Academic Rank:

Dr. Samer Hanna , Member. _ _ _ _ _ _ _ _ _ _ _ _ _
Academic Rank:

Dr. Wael Hadi , External Member._ _ _ _ _ _ _ _ _ _ _ _ _
Academic Rank:

Dedication

This thesis is dedicated to:

The sake of Allah, my Creator and my Master,

My great teacher and messenger, Mohammed (May Allah bless and grant him), who taught us the purpose of life.

Ibrahim AbdulNabi

Acknowledgment

It would not have been possible to write this master thesis without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here.

Above all, I would like to express my thanks and sincere gratitude for the one has guided me through my study and my thesis work; my supervisor *Dr. Samir Tartir*, for giving the wisdom, strength, support and knowledge in exploring things.

I would like to thank my family members; my Mother the symbol of love and giving, brothers, sisters and my daughter Jouri.

My dearest wife, who leads me through the valley of darkness with light of hope and support.

Ibrahim AbdulNabi

Table of Contents

List of Tables

List of Abbreviations

Abbreviation	Full Name
NLP	Natural Language Processing
ML	Machine Learning
CWA	Closed World Assumption
SSA	Social Sentiment Analysis
ANLP	Arabic Natural Language Processing
ASO	Arabic Sentiment Ontology

List of Figures

Abstract

Research done on Arabic sentiment analysis is considered very limited almost in its early steps compared to other languages like English whether at business level or social level.
Twitter now can be considered as an information network instead of a social network; this is because Twitter is a platform for shared experiences and it is a very human network .Twitter, a micro-blogging service, has emerged as a new medium in spotlight through recent happenings.

Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation.

Thesis problems are discovering user attitudes and business insights from Arabic Twitter feed and Enhancing Arabic Natural Language Processing using novel Semantic Web techniques

The Motivation for our work is Social media have swept into every industry and business function and are now an important factor of production, The primary goal of our work is to help companies make better business decisions by enabling semantic web and data mining.

We proposes a new model to discover Arabic business insights from tweets that may affect the decision making process, This model classifies Arabic tweets into positive and negative and tries to enhance the classification precision of tweets by creating Arabic Sentiment Ontology (ASO) for first time and also building our ASO aggregate weights algorithm.

Our Solution will be a starting point for any research that focus on Arabic sentiment analysis and extract information from Arabic social media.

CHAPTER ONE: INTRODUCTION

1.1 Preface

During recent years, online social networks - such as Facebook[1], Twitter[2], and MySpace[3] have become so popular. Millions of internet users have been attracted, many of whom have incorporated these sites into their daily exercises (Benevenutoy and Rodriguesy, 2009).

Through online social media, users communicate with each other, share content-such as videos and audios - and spread information. Many sites supply social links, for example, networks of contacts and professionals (e.g., LinkedIn, MySpace, Facebook) and networks for sharing videos, audios and other content (e.g., YouTube, Flickr), (Benevenutoy and Rodriguesy, 2009).

Online social networks are virtual societies that allow internet users to connect with one another. Social networks supply a diversity of ways for individuals to communicate with their existing family and friends, make new mates, or make “contacts” to build professional network.

These types of social sites can be magnificent ways to reconnect with old colleagues and friend, to share information, photos, videos, audios with friends, and see relevant news readily (Boyd and Ellison, 2008).

Twitter is a free microblogging service founded in 2006 by Jack Dorsey and Biz Stone. At its heart are 140-character bursts of information called tweets, Twitter enables users to send and read "tweets", which are text messages limited to 140 characters. Registered users can read and post tweets,Twitter through the website interface, SMS, or mobile device app

[1] www.facebook.com
[2] www.twitter.com
[3] www.myspace.com

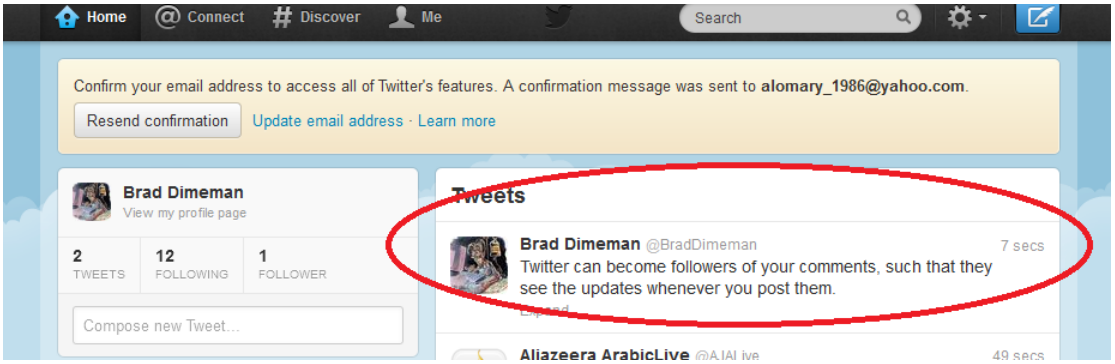

Figure 1.1Twitter status

Twitter now can be considered as a social network - communicate with their existing family and friends –but it's also increasingly being considered as an information network; this is because Twitter is a platform for shared personal information and it is a human network which mean there are interactions between users. Twitter and Facebook, as examples, have extended their platforms and user base markedly.

These days, the total number of Facebook users exceeds 677 million subscribers. In addition, Facebook mobile users have exceeded 250 million users. On the other hand, Twitter users also reached more than 200 million users; Twitter users post about 4 billion comments or tweets a month (Salem and Mourtada , 2011).

Companies now are focus on Twitter and social media in general to make what they called Big data, Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, search, sharing, transfer, analysis, and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, and prevent diseases.

In the Arab world, Twitter users have exceeded 1,311,882 users by March 2012 and 36,900 users in Jordan; thus Arabic language is considered as the fastest rising language on Twitter, among 25 other various languages used to post tweets. To be suitable for this growth, Twitter lately used an Arabic language interface (Salem and Mortada,2012).

1.2 Research Aims

This thesis aims to:

- Use Semantic Web to analyse Arabic text
- Enhance Natural Language Processing (NLP) in Arabic
- Provide A comprehensive framework that applies Arabic NLP and semantic web techniques that extract attitudes trends regarding certain topic from Arabic tweets.
- Apply our framework on Twitter big data.
- Measure the results and suggest enhancements for future work Problem Statement

1.3 Problem Statement

From the above research context, the following challenges may be largely derived,
Twitter is both an inbound and outbound tool that can give you the information you need to execute successful business strategies. Businesses need this information to target their clients and make changes to correct the bad image if a large number of customers have similar opinions. Also, do the same good stuff if a large number of customers agree.

Up to our knowledge, currently in Arabic tweets only simple statistics exist just about politics, like The most popular trending hashtags across the Arab region in the first quarter were #egypt (with 1.4 million mentions in the tweets generated during this period) #jan25 (with 1.2. million mentions), #libya (with 990,000 mentions), #bahrain (640,000 mentions), and #protest (620,000). (Salem and Mortada, 2012)

Our problem statement is discovering user attitudes and business insights from Arabic Twitter feeds focusing on certain Arabic dialect; Jordanian dialect as an example by enhancing Arabic Natural Language Processing (ANLP) and using semantic web to solve the Twitter language problems.

Arabic specifically has many issues when it comes to automatic processing. In order to extract information from Arabic tweets, we will first need to use NLP and Semantic web techniques to understand the tweets. Twitter NLP especially in Arabic has many issues that must be addressed such as very different slangs spoken in different locations, sometimes in

the same country, omitted diacritics, free word-order, long sentences, and highly-inflectional writing (Microsoft, 2012). For examples all ofthe following tweets are equivalent:

“Nizar did not buy new table”
"نزار لم يشتري طاولة جديدة"
"نزار مشتراش طربيزة جديدة"
"نزار مشتراش طاولة جديدة
“Nizar ma 2shtra tawla jdeeda”

1.4 Motivation

Customer’s value is critical interest to companies, because it determines how much it is worth spending to make a particular customer write something about the company or the product. However, traditional measures of customer value ignore the fact that, in addition to buying products himself, a customer may influence others to buy those products, (Domingos, 2001) and from this part the Influencer term is existing in Twitter, influencer have more than 15000 follower, companies would love to have these people sharingopinions about the business in Twitter and this can affect the decision of other and this make twitter more attractive to for business.

Twitter is increasingly becoming valuable for companies; tweets can maximize positive word-of-mouth among customers, Twitter increase’s the influence over consumer word of mouth, also twitter can be used now to discover user attitudes and to extract information about brands and produces, A recent study found that positive word of mouth among customers is by far the best predictor of a company's growth (Reichheld, 2003).

According to our research; little work is being done on extracting knowledge from Arabic tweets and little work is done in ANLP, ANLP has major issues one of them is very different dialects spoken in different locations

1.5 Contributions

This thesis aims at

- Proposing a new model to discover Arabic business insights from tweets that may affect the decision making process, This model classifies Arabic tweets into positive and negative and tries to enhance the precision of tweets
- Tackle dialect problem which is rarely tackle (Ahmed et al,2013).
- Build first ASO for Arabic tweets.
- Build first ASO aggregate weights algorithm.
- Study Twitter influencer weight and how effect on business attitudes.

1.6 Thesis layout

Chapter one: introduction includes the overview, aims, objectives, Problem Statement.
Chapter two: Literature review includes some of the related works about NLP, Twitter and Semantic web.

Chapter Three: Proposed Model that display our ASO aggregate weights algorithm that we use and our tool with some example, build Arabic ontology for classification and comparing result and how successful is it.

Chapter Four: Experimental Result Contain this study result and did its work and to how extent it is useful and if its meet the expectations.

Chapter Five: Implementation Issues, Evaluation and Application Areas.

CHAPTER TWO: LITERETURE REVIEW

This chapter summarizes the Arabic language importance and the previous work that search in the Arabic natural language processing, Twitter, and Sentiment Ontology and how to apply it on Arabic language.

2.1 Overview

Natural Language Processing (NLP) concerned with three scientific fields, artificial intelligence, computer science, and linguistics interested in the interactions between human or natural and computers languages. NLP is concerning to the scope of computer- human interaction. Many challenges in Natural language processing include natural language comprehending and allowing computers to extract meaning from human (natural) language input (Mihalcea et al,2006).NLP transform human language (spoken by human) into formal semantic representations which computer can act on, interpret, and respond with readily understood grammatical sentences. Natural language processing demands analysing underlying lingual relationships and structures, explicit concepts, grammatical rules, implied meanings, and more. Because sentences and individual words overwhelmingly have various meanings, and vice versa, a single idea can be written in considerable different words and forms (Mihalcea et al, 2006).

NLP includes techniques a computer needs to understand, analyse and process human language (typed or spoken by human) and also generate the natural language. As shown in figure 2.1, the Natural Language Processing includes two operations Natural Language understanding and Natural Language generation.

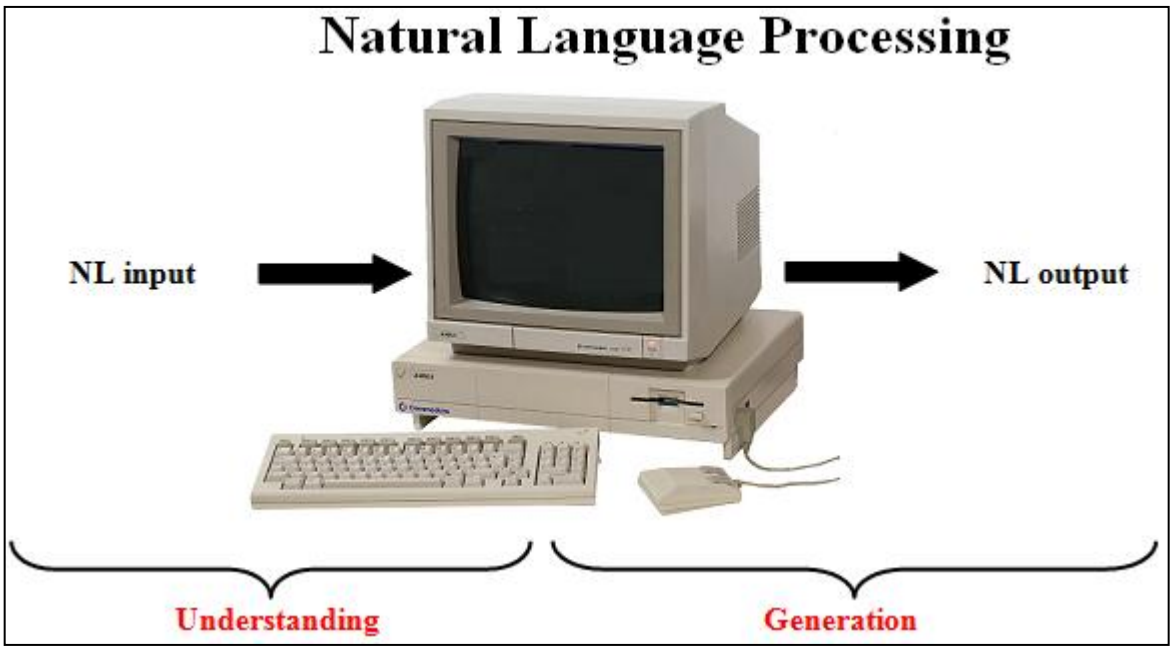

Figure 2.1: Natural Language Processing

2.2 Arabic Language Importance

The Arabic language makes significant challenges to considerable natural language processing applications. Arabic is a highly derived and inflected language (Habash, 2001).

The Arabic language presents developers and researchers of natural language processing applications and frameworks for Arabic speech and text with significant challenges (Farghaly and Shaalan, 2009).The Arabic language is interesting and challenging. It is interesting because of its history, its literary and cultural heritage, and the strategic significance of its people and the territory they occupy. It is also challenging due to its complicated lingual structure. At the historic level, old Arabic has stayed unchanged, functional and understandable for more than fifteen horns. The Arabic language is tightly related with Islam .Strategically, it is the local languages of more than 330 million speakers occupy an important area with massive oil reserves to the world prudence (Farghaly and Shaalan, 2009). Arabic is a supreme language mouthed by more than 330 million speakers as a local language, in an area spreading from the Persian/ Arabian Gulf in the Atlantic

Ocean in the West to the East. ANLP applications must deal with various complex issues relevant to the structure and nature of the Arabic language. Such as, Arabic is the language direction, it is from right to left. Like Chinese, Korean and Japanese, and there is no capitalization in Arabic. Also, Arabic letters alter their shape according to their place in the word.

2.3 Twitter

Currently, online social networks such asFacebook, Twitter, Google+, LinkedIn, and Foursquare have become extremely popular allover the world and play a significant role inpeople's daily lives.

Twitter, a micro-blogging service, has emerged as a new medium in spotlight through recent happenings, such as an American student jailed in Egypt and the US Airways plane crash on the Hudson River. Twitter users follow others or are followed. Unlike on most online social networking sites, such as Facebook or MySpace, the relationship of following and being followed requires no reciprocation. A user can follow any other user, and the user being followed need not follow back. Being a follower on Twitter means that the user receives all the messages (called tweets) from those the user follows. Common practice of responding to a tweet has evolved into well-defined mark-up culture: RT stands for retweet, '@' followed by a user identifier address the user, and '#' followed by a word represents a hashtag. This well-defined mark-up vocabulary combined with a strict limit of 140 characters per posting conveniences users with brevity in expression. The retweet mechanism empowers users to spread information of their choice beyond the reach of the original tweet's followers.

Using NLP (Gwahangno, 2010) and machine learning (ML) techniques to develop a suite of classifiers to differentiate tweets across several dimensions: subjectivity, personal or impersonal style, and linguistic register (formal or informal style). Based on initial analyses of tweet content, they posit tweets that contribute to situational awareness are likely to be written in a style that is objective, impersonal, and formal; therefore, the identification of subjectivity, personal style and formal register could provide useful features for extracting

tweets that contain tactical information. To explore this hypothesis, they have study four mass emergency events: the North American Red River floods of 2009 and 2010, the 2009 Oklahoma grassfires, and the 2010 Haiti earthquake.

More in depth analysis of Twitter data is available in other languages, e.g. English. For example, Twitter's analytics infrastructure is built around Hadoop, the open-source implementation of Map Reduce (Dean, 2008), which is a popular framework for large-scale distributed data processing. Rios and Lin use their central data ware-house to build around a large Hadoop cluster. Data arrive in the Hadoop Distributed File System (HDFS) via a number of real-time and batch processes: such as bulk exports from frontend databases, application logs. One of the examples that have been used in the study is FIFA world cup, in the summer of 2010, Twitter users shared their experiences in real-time as they watched games during the FIFA World Cup. During the final match, users from 172 countries tweeted in more than 20 different languages; a record of 3,283 tweets posted in a single second was established during the match between Japan and Denmark (a record unbroken for 6 months). They showed how people tweeted during the games, in a way that highlights both the volume of the conversations but also the competition between different teams. This was captured in a stream graph, where the width of the stream is proportional to the volume of conversation.

Guang Xiang and his colleagues(Sudha, 2011) built a framework that exploits linguistic regularities in profane language via statistical topic modelling on a huge Twitter corpus, and detects offensive tweets using these automatically generated features; their approach performs competitively with a variety of machine learning (ML) algorithms. For instance, the approach achieves a true positive rate (TP) of 75.1% in over 4029 testing tweets using Logistic Regression, significantly outperforming the popular keyword matching baseline, which has a TP of 69.7%, while keeping the false positive rate (FP) at the same level as the baseline at about 77%.

2.4 Natural Language Processing

Natural Language Processing is an interdisciplinary research area at the border between linguistics and artificial intelligence aiming at developing computer programs capable of human-like activities related to understanding or producing texts or speech in a natural language, such as English or Chinese

One of NLP tasks is Natural language understanding which is mean converting chunks of text into more formal representations such as first-order logic structures that are easier for computer programs to manipulate. Natural language understanding involves the identification of the intended semantic from the multiple possible semantics which can be derived from a natural language expression which usually takes the form of organized notations of natural languages concepts. Introduction and creation of language Meta model and ontology are efficient however empirical solutions. An explicit formalization of natural languages semantics without confusions with implicit assumptions such as closed world assumption (CWA) vs. open world assumption, or subjective Yes/No vs. objective True/False is expected for the construction of a basis of semantics formalization (Miguel and Lin, 2012).

Semantics is a more serious problem. The output of the semantic component is the "meaning" of the input. But how can this "meaning" been expressed? Not by anything as simple as a sequence of words. Many different "meaning representation languages" have been developed in an attempt to find a language that has the appropriate expressive power, but there is no uniform semantic representation language that can represent the meaning of every piece of NL (Guang et al, 2012).

In (Aramaki et al, 2011) proposed a new Twitter-based influenza epidemics detection method, which relies on the Natural Language Processing (NLP). Their proposed method could successfully filter out the negative influenza tweets (f-measure=0.76), which are posted by the ones who did not actually catch the influenza. The experiments with the test data empirically demonstrate that the proposed method detects influenza epidemics with high correlation (correlation ratio=0.89), which outperforms the state-of-the-art

Google method. This result shows that Twitter texts precisely reflect the real world, and that the NLP technique can extract the useful information from Twitter streams.

2.5 Semantic Web

A semantic web is a web where the focus is placed on the meaning of words, rather than on the words themselves: information becomes knowledge after semantic analysis is performed. For this reason, a semantic web is a network of knowledge compared with what we have today that can be defined as a network of information (Stefano, 2000).

In his original vision(Mayank et al, 2004)for the World Wide Web, Tim Berners-Lee described two key objectives: (1) to make the Web a collaborative medium; and (2) to make the Web understandable and, thus, process able by machines. During the past decade the first part of this vision has come to pass – today's Web provides a medium for presentation of data/content to humans. Machines are used primarily to retrieve and render information.

Humans are expected to interpret and understand the meaning of the content. Automating anything on the Web (e.g. information retrieval; synthesis) is difficult because interpretation in one form or another is required in order for the Web content to be useful. Current information retrieval technologies are incapable of exploiting the semantic knowledge within documents and, hence, cannot give precise answers to precise questions. (Indeed, since web documents are not designed to be understood by machines, the only real form of search is full-text searching.) (Mayank et al, 2004).

The Semantic Web (Lee, 2001) is an extension of the current web. It aims to give information a well-defined meaning, thereby creating a pathway for machine-to-machine communication and automated services based on descriptions of semantics. Realization of this goal will require mechanisms (i.e., markup languages) that will enable the introduction, coordination, and sharing of the formal semantics of data, as well as an ability to reason

and draw conclusions (i.e., inference) from semantic data obtained by following hyperlinks to definitions of problem domains (i.e., so-called ontologies) (Geroimenko, 2006).

There is integration between Semantic web and NLP, NLP is vital to the success of the semantic web because it is the method of communication between humans and software agents, parsing, knowledge representation.

Information extraction and semantic analysis are used in many semantic web technologies; Ontologies provide a way to add context to information, by specifying which ontology an agent should use the algorithm can eliminate any ambiguities between words.

In (Gruhl et al, 2009) the researchers explore the application of restricted relationship graphs and statistical NLP techniques to improve named entity annotation in challenging Informal English domains. The author validates the approach using on-line forums discussing popular music. Named entity annotation is particularly difficult in this domain because it is characterized by a large number of ambiguous entities, such as the Madonna album “Music” or Lilly Allen’s pop hit “Smile”.

2.6 Twitter and Sentiment Analysis

There are large collections of research around using machine learning techniques for sentiment analysis; sentiment analysis is an automated task where machine learning is used to rapidly determine the sentiment of large amounts of text or speech. Pang et al. (Pang, 2008) they are one of the first to apply sentiment analysis to online movie reviews .Their findings show that ML and especially Support vector machine and Naïve Bayes are good enough to extract the sentiment from movie review when they compared their work to human work.

Research by (Lai, 2011) brought sentiment analysis to the Twitter domain by applying similar machine learning techniques to classifying the sentiment of tweets (Pang, 2008). Their contribution was using emoticons as noisy labels during the training process.

Recently, researchers move from traditional machine elearning techniques to using lexicon-based approaches which used sentiment lexicons to determine word polarity if it positive or negative based on similarity. It should be noted that many of the sentiment lexicons used in these projects are not tailored towards the type of language used in social media.

The research in (Conover, 2012) showed that Twitter does indeed provide a platform for political deliberation. In addition, they use LIWC as a database for sentiment lexicon and their result show that sentiment extraction can produce result similar to traditional election polls, which mean if they expand the lexicon the result will be more efficient, They used the Subjectivity Lexicon from Opinion Finder to classify tweets as positive or negative and correlated the results to hand-measured polls, in the same time The work of (O'Connor, 2010) try to include part-of-speech information and emoticons to enhance sentiment extraction process from tweets and the use of a sentiment lexicon tailored towards text originating from social media, which give more strength to their work by attempting to extract sentiment polarity but also sentiment strength (i.e. strong approval vs.strong disapproval)

The bulk of SSA (Abdul-Mageed et al, 2012) work has focused on movie and product reviews. A number of sentence- and phrase-level classifiers have been built: For example, whereas (Yweet al, 2003) present a system that detects sentiment toward a given subject, (Kim and Hovy's, 2004) system detects sentiment towards a specific, predefined topic. Their work is similar to (Yuand Hatzivassiloglou, 2003) and (Wiebe et al, 1999) in that they use lexical and POS features.

Only few studies have been performed on Arabic. (Abbasi, 2008) use a genetic algorithm for both English and Arabic Web forums sentiment detection on the document level. They exploit both syntactic and stylistic features, but do not use morphological features.
The categorization level task (Soha et al, 2013) depends on the text length.

There are mostly two levels: document and phrase level. At the document level, researchers classify long text (which contains more than one sentence) as subjective or objective then the subjective as positive or negative. Text categorized as documents include forum posts, blog articles, product reviews, and Tweets (Abdul-Mageed et al, 2012).

At the sentence level, short text or even single sentences are classified. Sentence level categorization is also very popular among Arabic sentiment analysis research , perhaps for two reasons: First is the emergence of social media text, which is usually short out of convenience, or even by requirement (e.g., Tweets). Second, sentence level sentiment analysis may be seen as a subtask of document level sentiment analysis and thus any improvement of the subtask will result in an improvement of the main task. Finally, the source and target identification task is concerned with identifying the source of the sentiment. For example, in the sentence "I hate the iPhone", the source of the sentiment is the speaker, while in the sentence "My brother hates it", the source is another person (who is quoted). In both cases, the target of the sentiment is the "iPhone". To the best of our knowledge, target identification was not studied in Arabic sentiment analysis research.

Various feature (Soha et al, 2013) types have been used for classification purposes in Arabic sentiment analysis.

Table 2.1 A Taxonomy of Sentiment Analysis Classification

Tasks Characteristics	
Category	**Description**
Categorization Classes	Positive/negative sentiment or objective/subjective text
Categorization Levels	Document or sentence/phrase-level classification
Source/Target Identification	Whether source/target of sentiment is known or extracted
Features	
Category	**Description**
Syntactic	Word/POS tag n-grams, phrase patterns, punctuation
Semantic	Polarity tags, appraisal groups, semantic orientation

Stylistic	Lexical and structural measures of style
Techniques	
Category	**Examples**
Machine Learning	Techniques such as SVM, naive Bayes, etc.
Link Analysis	Graph-based sentiment analysis
Similarity Score	Phrase pattern matching, frequency counts, etc.
Input Domain	
Category	**Description**
Reviews	Product, movie, and music reviews
Web Discourse	Web forums and blogs
News Articles	Online news articles and Web pages
Social Media websites	Twitter, Facebook, YouTube

Table 2.1 shows the various feature types have been used for classification purposes in Arabic sentiment analysis. Text is classified as subjective or objective and, secondly, polarity is determined by classifying the subjective text as positive or negative

2.7 Sentiment Analysis in Arabic

Sentiment Analysis for Arabic language still in the early stages and especially for social media, when we try to deal with Arabic content found in social media websites, we are faced with the problem of "diglossia". Diglossia is defined in sociolinguistics as the phenomenon of using linguistic standards in formal media that differs from the ones used in every day-spoken language. For instance, social media sites which are likely to contain opinionated and evaluative content are mostly produced by text written in the informal colloquial variations of Arabic(i.e., dialects) which have no structure and are very difficult to standardize and from this point I've try to focus on Jordanian Dialect.

2.8 Ontology

Ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where Ontology is a systematic account of Existence. For AI systems, what "exists" is that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are

reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, in the context of AI, we can describe the ontology of a program by defining a set of representational terms. In such ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Formally, ontology is the statement of a logical theory.

Ontologies are tools that provide a lot of semantic information. They help to define concepts, relationships, and entities that describe a domain with an unlimited number of terms.

CHAPTER THREE: PROPOSED MODEL

3.1 Introduction

In this section, we describe the method for Twitter data collection, analysis and design.
Our research tries to discover the feasibility of extracting Twitter users' interests by classifying Arabic Tweets and hashtags by building Arabic ontology and using Natural Language Processing (NLP) and semantic web techniques, this research aims to investigate and discover user's attitudes towards business products or companies between Arab Twitter users .

We have chosen Twitter (Piao and Whittle, 2011) for this study for its popularity, variety of usage, real-time feature and public availability. It is being used for a wide range of different purposes, including personal branding, hiring people/job notices, reading news, networking for benefits, taking notes, setting up meetings etc. Quite a few uses are related to people's interests and concerns. In addition, the real-time feature of Twitter is also beneficial for our task. In fact, these features mentioned above make twitter an excellent resource for extracting and tracking users' latest interests, particularly those emerging ones.

Twitter messages have many unique attributes, which differentiates our research from previous research:

- Length The maximum length of a Twitter message is 140 characters. From this training set, we have calculate that the average length of a tweet is 14 words or 78 characters.
- Data availability another difference is the magnitude of data available. With the Twitter API, it is very easy to collect millions of tweets for training.
- Language model Twitter users post messages from many different media, including their cell phones. The frequency of misspellings and slang in tweets is much higher than in other domains.
- Domain Twitter users post short messages about a variety of topics unlike other sites which are tailored to a specific topic.

The majority of the text produced by the social websites is considered to have an unstructured or noisy nature (Soha et al, 2013). This is due to the lack of standardization, spelling mistakes, missing punctuation, nonstandard words, and repetitions. That is why the importance of pre-processing this kind of text is attracting attention these days because of the presence of several websites producing this noisy text.

This research is unique and one of the few research's that focus on Arabic tweets and how Arabic people use twitter for praise or dispraise products or company, we have explore NLP techniques to carry out semantic content analysis of the unstructured Arabic tweets in order to identify the user's interests expressed in natural language.

From our opinion we can say that Arabic sentiment analysis is determining user attitude regarding some product or business or the overall tonality of the text is considered the main task of sentiment analysis, the definition of a positive or negative attitude in sentiment analysis is relatively hard to understand, the degree of Positive and negative varies considerably among people, rendering the labelling process in our task even harder.

As we adopted tweets as the main information source, some unique features of Twitter bring impact, both positive and negative, on our work.

The first concern (Piao and Whittle, 2011) is the short sizes (max 140 characters) of tweets, as users may be unable to write down complete thoughts, failing to provide sufficient information for the interest detection. While the problem is true for individual tweets, if appropriate quantities of tweets can be collected, they can provide a good information source for content analysis. (Kim et al, 2010) found that "even though tweets are brief, they contain enough information to express identifiable characteristics, interests and sentiments". We found that collectively tweets provide a wealth of information about users' thoughts and interests.

As our algorithm is designed to work on tweet collections of individual users, this problem is alleviated.

The second concern is the informal style of the tweets. Most tweets are written informally and often contain numerous typos, abbreviations and have an unstructured or noisy nature, this is due to the lack of standardization.

To avoid or decrease these noises we will pre-process the tweets before use it, so we will pre-process Arabic tweets that we will use during our research. There (Shoukry and Rafea, 2012) are mainly three steps in the pre-processing process: 1) normalization, 2) stop words removal. Normalization is the process of transforming the text in order to be consistent, thus putting it in a common form.

We proposes A new model to discover Arabic business insights from tweets that may affect the decision making process, This model classifies Arabic tweets into positive and negative and tries to enhance the precision of tweets classification through its specific methodology. Enhance the precision of classification by using ASO Aggregate weights that use down to top approach and using similarity score (El Din & Al Taher, 2013).

3.2 Algorithm for Training Data

Our tweet corpus contains the textual messages along with other Meta data such as twitter ID, posting time, etc. The raw tweets in our experiment were collected by using Tweet Archives' website.

Tweets are expressed in an extremely colloquial fashion, with substantial noise and linguistic variation.

3.2.1 Processing Training Data

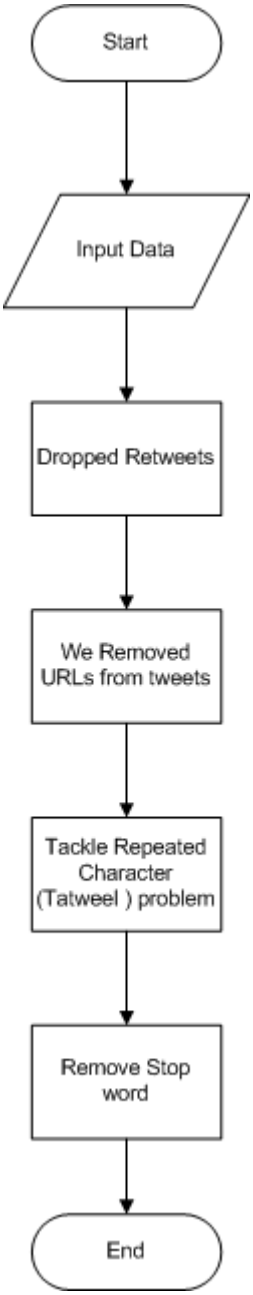

Figure 3.1 Processing Training Data Algorithm

Algorithm for preprocessing of Tweets:

- We intentionally remove retweeted tweets (tweets that have rt), because they unnecessarily magnify the weights of words.
- Remove shortened URL
- Remove stop words
- Fix Elongating (Tatweel) problem like تكبيـــــــــــــــــــــــر Finally we will keep the username and his/her tweet

After the pre-processing we Extract the positive and negative words and combine each word with similar Jordanian dialect words, this create lexicon database for Jordanian dialect.

3.2.2 Arabic Sentiment Ontology

For our work we build the first Arabic Sentiment Ontology for social media and for Jordanian Dialect, our ontology focus on semantic relations, we mean here the relationships between concepts not the words.

We use Subtype relation as a mathematical relations (subset: A is a subset of B), such that every instance in A must also be an instance of B, and Inheritance relation that mean subtypes inherit all properties of their super types.

Below is a sample of our ontology for negative and positive that we use in our classification algorithm.

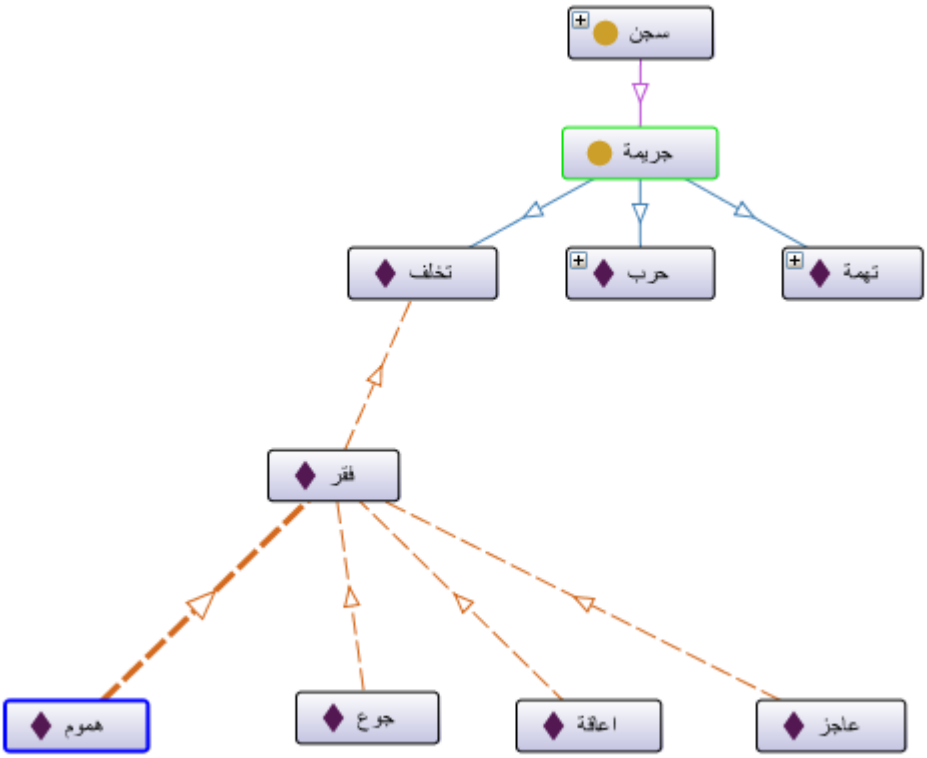

Figure 3.2 Negative Ontology

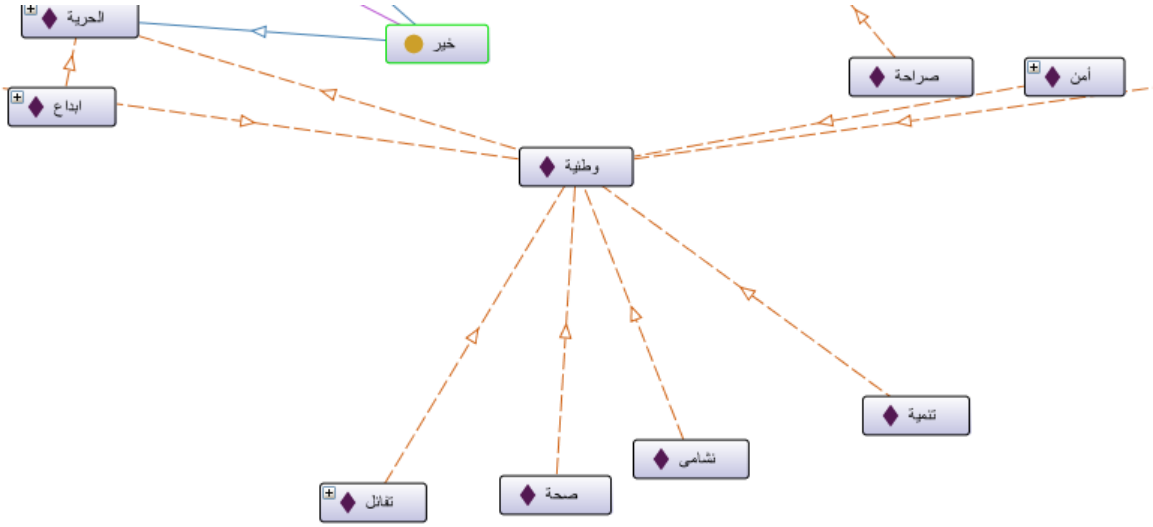

Figure 3.3 Positive Ontology

3.2.3 Sentiment Classification

We built a classifier that judges whether a given tweet is positive or negative. This task setting is similar to a sentence classification (such as spam e-mail filtering, sentiment analysis, and so on).

Table 3.1 Tweets Sentiment Classification

Positive	Positive indicator on topic
Negative	Negative indicator on topic
Neutral	● Neither positive nor negative indicators ● Mixed positive and negative indicators ● On topic, but indicator undeterminable ● Questions with no strong emotions Indicated

3.2.4 Algorithm for Tweets Classification

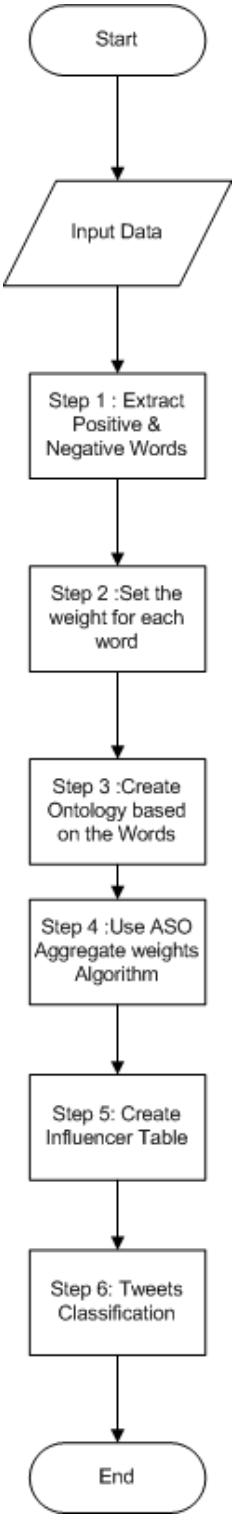

Figure 3.4 Tweet Classification Algorithm

- Extract positive & negative words, combine each word with similar Jordanian dialect word.
- Each word have a weight, if it is positive the weight scale will be from 1-5 and 5 is the highest, the negative scale is from -1 to -5 and -5 is highest.

- These weights come from Survey for 20 people to get the right weight and avoid the bias from our side.
- Each positive or negative word is linked to another word (father) based on our Arabic Sentiment Ontology, this relationship will make hierarchy
- Each father to the root have also weight, our ASO Aggregate Weight will be based on the weight of the word and also the total sum of fathers, for example: (3-) مشاجرة (4-) مرتبطة بكلمة قتل وزنها (2-) مرتبطة بكلمة جريمة وزنهاso the total will be -9, if we have positive words we will treat it in the same way, then sum the negative to positive and get the result if this tweet positive or negative.

- We have table for some of users that work for our sample data for example (users that work for Ro'ya TV) we put them in a table and if we have tweets from these users about Roy'a TV then we will give them special weight, -3 for the users that work in the Ro'ya TVand add the -3 to the total.

CHAPTER FOUR: EXPERIMENTAL RESULT

4.1 Tweets Collection

NLP classifiers require "training" data in the form of annotated or coded text, which they use to "learn" how to distinguish between positive and negative.

Users on Twitter generate over 400 million Tweets every day (Kumar and Morstatter, 2013) some of these Tweets are available to researchers and practitioners through public Application Programming Interface (API) at no cost, but unfortunately twitter stop using the API for retrieving tweets for more than one week, so we use another tool to get the collection of tweets.

We use Tweet Archivist[4]to collect tweets about our topics, Tweet Archivist is a Twitter analytics tool to search, archive, analyse, visualize, save and export tweets based on a search term or hash tag that is easy to crawl and collect data.

On Twitter, users create profiles to describe themselves to other users on Twitter. A user's profile is a rich source of information about him. We crawled and collected profiles of users on Twitter. Additionally, we collected profiles of users that work for one of the companies that we use in our thesis, we collect these profile to use it in our ASO Aggregate Weight algorithm.

For our work we collect 1100 tweets (then it's become 1000 after removing English tweets) for three major companies in Jordan

1) Ro'ya TV: is a satellite channel launched from Amman to join the media package of al Sayegh group.

[4]http://www.tweetarchivist.com/

2) Jamalon is Arabic bookstore based in Amman and Jamalon now consider as the main source for Arabic books in Arab world.

3) Khaberni is Jordanian website that cares mainly about local issues and other worldwide issues.

The 1000 tweets were extracted from three hashtags (#TVRo'ya, #Jamalon, and #Khaberni) and 3 mentions (@tvro'ya, @jamalon, and @khaberni).

We use 400 tweets for Ro'ya TV, 400 tweets for Jamalon and 200 tweets for Khaberni, all these tweets were in September 2013 using Tweet Archivist website in Excel format, that contain user id, tweets, location and time.

4.2 Proposed Tweets Sentiment Classification

Our Proposed system use our ASO Aggregate Weight Algorithm with our Arabic Sentiment Ontology (ASO), to get best result using, the system components include pre-processing that improved the classification accuracy, ASO aggregate similarity score, below is our proposed architecture

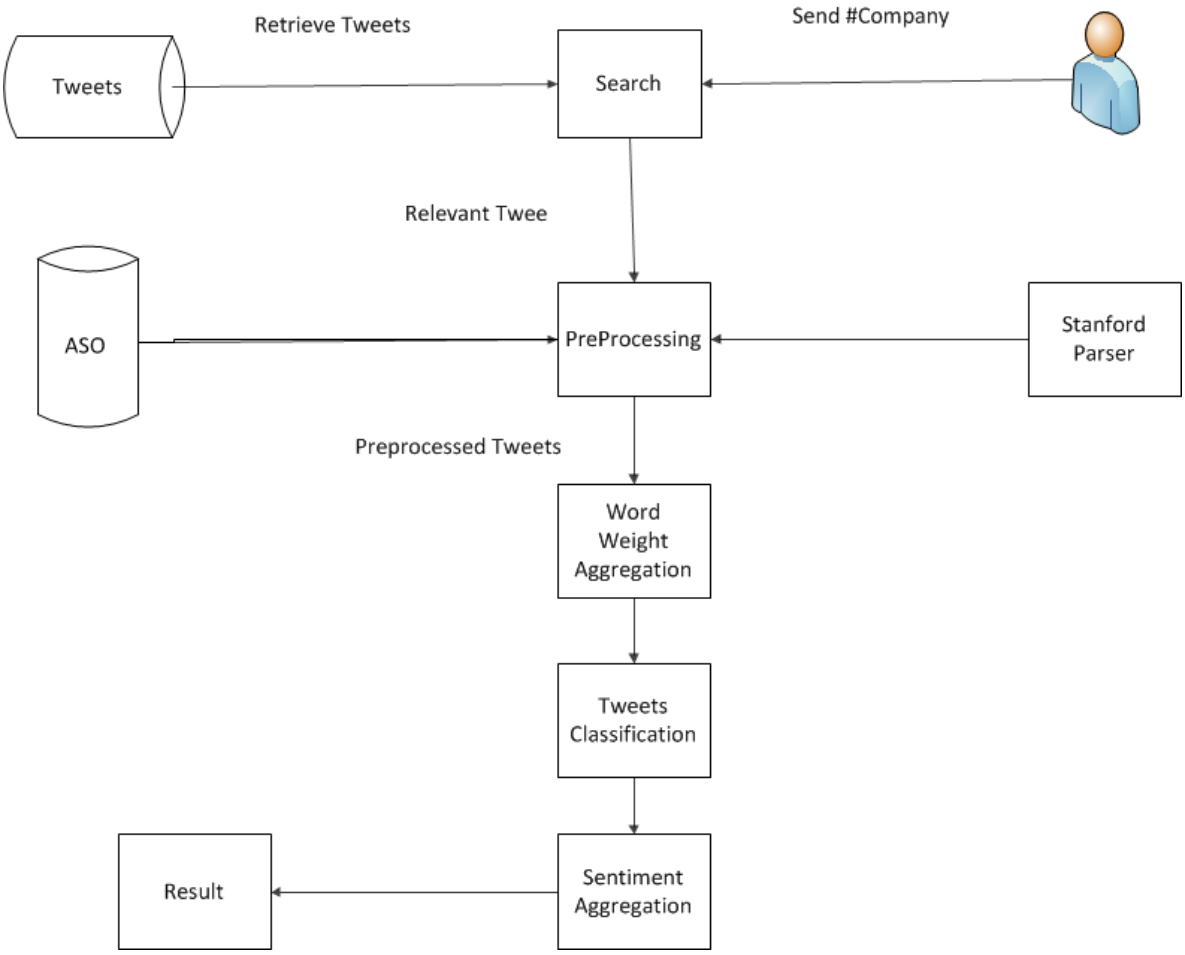

Figure 4.1 Proposed Architecture.

We will explain more the proposed architecture:

- Step 1: Raw Data (Extracted tweets from Tweet Archivist)
- Step 2:Pre-processing our data includes
    - Remove English tweets
    - Remove stop word
- Step 3: Building Ontology & Annotators Classification
    - Extract Jordanian dialect Positive & Negative words
    - Make a survey for words weight, and give each word specific weight
    - 3 Annotators to classify tweets manually
    - Build Arabic Sentiment Ontology based on extracted words
    - Build table Twitter influencer weight and how effect on business attitudes

- Step 4: Building ASO Aggregate Weights algorithm, this algorithm sum all the weights for the word in the ontology that the same word in the tweet till the root.
- Step 5: Compare the results between our approach and the results of annotators

Our classification approach is to classify Jordanian tweets to natural tweets, positive tweets and negative tweets, We extracted the positive and negative words from the tweets and then make a survey for 20 person's to average weight for each word, below is a sample for positive weight table

Table 4.1 Sample of Positive weight words

Word	**Weight**
كويس	4
نشامى	3
حلو	4
صراحة	4
مبروك	3

صح	2
مليح	3
فخور	4
عطف	3
تنمية	1
محترم	3
عظيم	3

And we did the survey for the negative word that had been extracted from tweets collection and below is a sample of negative weight words.

Table 4.2 Sample of Negative weight words

Word	**Weight**
أسعار	- 1
طوشة	-3
شكوى	-4
قرف	-2
هبل	-1
وساخة	-4
فقر	-3
ضياع	-4
ممل	-4
تأخير	-4
عيب	-3

4.3Annotation Process

We asked three annotators, native Arabic language speakers, to annotate 1000 Tweets from each of the three domains (Khaberni, Jamalon and Ro'ya TV)to compare the human annotate process to our classification process.We choose the way of using three annotators to classify the tweets because unfortunately no publically available Tweet corpus for evaluating target-dependent Twitter Arabic text sentiment classification.We asked them to manually classify each Tweet as positive, negative, or neutral towards the target, or spam. Table 4.3 presents a summary of the data set (size and numbers of positive, negative and neutral Tweets) by using the following formula

- The distribution % for annotators for positive tweets = number of positive tweets/ size of data set.

- The distribution % for annotators for negative tweets = number of negative tweets/ size of data set.

- The distribution % for annotators for Neutral tweets = number of Neutral tweets/ size of data set

Table 4.3 the Dataset Annotators Statistics

Data Set	Jamalon	distribution %	Khaberni	distribution %	Tv Ro'ya	distribution %
Size	400		200		400	
Positive	225	56.3 %	85	42.5 %	310	77.5 %
Negative	109	27.2 %	60	30 %	60	15 %
Neutral	66	16.5 %	55	27.5 %	30	7.5 %

These three annotators classify the tweets manually base on how they understand the tweets, these results will be our baseline to compare the results from that appear from our classification algorithm that build using our Arabic Sentiment Ontology

4.4 Classification Results

After building our ASO Aggregate Weight that use the similarity score based on the survey that we had done before, and our Arabic Sentiment Ontology, We run this approach on the same data set that used in Annotation Process, and We got the following results that we will compare in the next section with Annotation Process to get the precision and recall for our algorithm.

We build our own application that implement the ASO Aggregate Weight, similarity score and looking up to the ontology

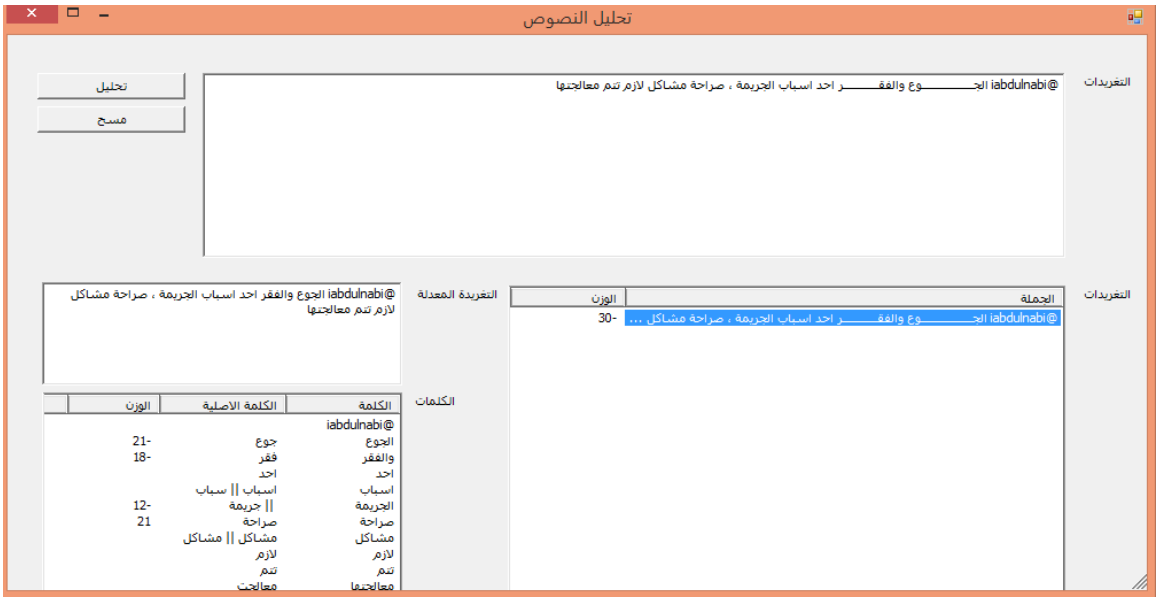

Figure 4.2 Arabic Sentiment Ontology Applications

Based on our application we got the following results

Table 4.4 the Data Set Statistics based on our Algorithm

Data Set	Jamalon	% of classification	Khaberni	% of classification	Tv Ro'ya	% of classification
Size	400		200		400	
Positive	203	50.5 %	73	36.5%	292	73 %
Negative	98	24 .5%	70	35 %	55	13.7 %
Neutral	99	24.7 %	57	28.5 %	53	13.3 %

The above table shows the statistics for the tweets based on our algorithm and the % of classification.

4.5 Results Comparison

In this section we will compare Annotation Process result and the result of our ASO Aggregate Weight Algorithm, in this comparison we will calculate the precision and recall for positive, negative and neutral categories for three domains (Khaberni, Jamalon and TV Ro'ya) based on the following Formulas:

- Precision (Pos) = $\frac{|Positive\ Retrieved| \cap |Positive\ Annotated|}{|Retrieved|}$

- Recall (Pos) = $\frac{|Positive\ Retrieved| \cap |Positive\ Annotated|}{|Positive\ Annotated|}$

And the same formula goes for negative and neutral for all data set.
And based on above formulas we got the following results:

Table 4.5 Precision and Recall for Jamalon

	Jamalon	
	Precision	Recall
Positive	180/225= 80%	180/203= 88%
Negative	60/109= 55%	60/70 = 85%
Neutral	50/66= 75.7 %	50/99 = 50 %
Average	70%	74.3 %

The above table shows that the precision for Jamalon is for positive 80%, Negative 55% and Neutral is 75%, the recall is for positive is 88%, Negative is 85% and Neutral is 50%.

Table 4.6 Precision and Recall for Khaberni

	Khaberni	
	Precision	Recall
Positive	67/85= 78%	67/73= 91%
Negative	50/60= 83%	50/70 = 71%
Neutral	40/55= 72%	40/57 = 70 %
Average	77.6%	77.3 %

The above table shows that the precision for Khaberni is for positive 78%, Negative 83% and Neutral is 72%, the recall is for positive is 91%, Negative is 71% and Neutral is 70%.

Table 4.7 Precision and Recall for Ro'ya TV

	Precision	Recall
Positive	280/310= 90%	280/292= 95%
Negative	43/60=71%	43/55 = 78%
Neutral	19/30= 63%	19/53 = 35 %
Average	74%	69.3 %

The above table shows that the precision for Ro'ya T.V is for positive 90%, Negative 71% and Neutral is 63%, the recall is for positive is 95%, Negative is 78% and Neutral is 35%. And the final result for our work can be summarized for three domains in below table

Table 4.8 Summery Result

Summery Result		
	Precision	Recall
Jamalon	70%	74 %
Khaberni	77.6%	77.3 %
T.V Ro'ya	74%	69.3 %
Average	73.6 %	64.8%

The above table shows the summery result for our approach and the average precision and recall, the average precision for our work is 73.6 % and the average for recall is 70.8%

CHAPTER FIVE: IMPLEMENTATION ISSUES, EVALUATION AND APPLICATION AREAS

5.1 Introduction

In this chapter, we will show the implementation issues for our work. We will discuss the application issues and areas where our classification algorithm can be used. We will evaluate our work by presenting the concepts that we modify in conventional approaches and define our own concepts, and compare the work and results we obtained with other works.We will end this chapter by a conclusion describing the perspectives and future works.

5.2 Implementation Issues

The implementation environment of this methodology requires a strongly typed and an object-oriented programming language. The checking process should guarantee the right pre-processing, execute the ASO Aggregate Weight based on the ontology and the similarity score based on the weights.

Recognizing opinion holders from text is a challenging task it would be very hard for machines to recognize that the stated opinion is from another person and not that of the author of the sentence, but we try to enhance the classification for tweets using ASO, and we should make sure that if we increase the size of this ontology and the size of positive and negative words this will mean better results, as we mention before that there is no publically available Tweet corpus for evaluating target-dependent Twitter Arabic text sentiment classification.

5.3 Application Areas

Discovering business attitudes for social media especially for Twitter will be strengthened by adding our algorithm to its classification tweets application, since it is more natural than current conventional approaches in presenting tweets classification and focusing on Arabic Dialect.

Our approach in highly recommended to be used in any business intelligence area like analysing social media, search for opinions, archive social media and visualize the market result.

5.4 Evaluation

In the following, we will compare the power of our concepts and compare our contributions with similar works.

Comparison with similar works: while reading the literature (Ahmed, 2013; Alaa El Din 2013; Abdul-Mageed 2012; El-Beltagy 2013), we found that Arabic sentiment approaches can be compared based on several criteria. We selected the most recent and closest researches to our work, and we choose the most important (from our vision) comparison criteria to be:

1. Using Ontology.
2. Algorithm that tackle Arabic dialect.
3. Precision measure.
4. Supporting methodology

And the papers used in this comparison are:

A. Soha Ahmed, Michel Pasquier, Ghassan Qadah; 2013;Key Issues in Conducting Sentiment Analysis on Arabic Social Media Text;IEEE Innovations in Information Technology (IIT).

B. Amira Shoukry, Ahmed Rafea; 2012; Pre-processing Egyptian Dialect Tweets for Sentiment Mining; The Fourth Workshop onComputational Approaches to Arabic Script-based Languages.

C. Samhaa R. El-Beltagy, Ahmed Alwe; 2013; Open Issues in the Sentiment Analysis of Arabic Social Media: A Case Study; IEEE Innovations in Information Technology (IIT).

D. Alaa El Din & Fatima Al Taher; 2013; Sentiment Analyzer for Arabic Comments System, International Journal of Advanced Computer Science & Applications.

This comparison is summarized in Table 5.1 Comparisons with other's work. The papers used for the comparison are from A-D.✓Symbol means strongly supported. ✖ Symbol means not supported, and ◆symbol means weak supporting.

Table 5.1 Comparisons with other's work

Paper/ Criterion	**A**	**B**	**C**	**D**	**Our approach**
Using Ontology	X	X	X	X	✓
Algorithm that tackle Arabic dialect.	X	✓	◆	X	✓
Precision	71%	74%	77 %	77 % using SVM 73% using Decision tree	73.6 %
Supporting Methodology	using SVM	using SVM	weighting system	SVM &Decision tree	using ASO Aggregate Weight

5.5 Conclusion: perspectives and future works

Through our study about Sentiment Analysis and Twitter Classification, we found that Arabic sentiment still need more and more work from research. We found that linking between discovering Arab user attitudes and extracting the business insights from social media is very weak, and there is a lack application that implement sentiment classification algorithm for Arabic.

In section 1.5, we proposed 4 contributions to be done during this thesis. The first try to enhance the precision of tweets classification through its specific methodology: Enhance the precision classifier by using down to top approach (ASO Aggregate Weight) and using similarity score (Alaa El Din & Fatima Al Taher, 2013) and we enhance the precision based on our result.

Tackle dialect problem which is rarely tackle (Ahmed et al, 2013) and we take Jordanian dialect as an example for our work. And we build first ASO for Arabic tweets.
And Tackle Twitter influencer weight and how effect on business attitudes

Our work can be extended and developed in future to:

- Expand the ontology to have more positive and negative words.
- Build Framework for all Arabic dialect.
- Enhance the precision and accuracy of results.
- Using other algorithm rather than ASO Aggregate Weight like Naïve Bayes, SVM, etc.

References

Abbasi, A, Chen, A; 2008; Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums; ACM Transactions on Information Systems, 26:1–34.

Abdul-Mageed, M, Kubler, S, Diab, M; 2012; SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media; ACM WASSA 12 Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis.

Ahmed, S, Pasquier, M, Qadah, G; 2013; Key Issues in Conducting Sentiment Analysis on Arabic Social Media Text; IEEE Innovations in Information Technology (IIT).

Amira, S, Rafea, A; 2012; Pre-processing Egyptian Dialect Tweets for Sentiment Mining; The Fourth Workshop on Computational Approaches to Arabic Script-based Languages, San Diego, 2012.

Aramak, E, Maskawa, S, Morita, M; 2011; Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter; ACM EMNLP 11 Proceedings of the Conference on Empirical Methods in Natural Language Processing; Stroudsburg, PA, USA.

Benevenutoy, R, Cha, F, Al Meiday, M.; 2009; Characterizing User Behavior in Online Social Networks; Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference; Berlin, Germany; pp. 49-62.

Boyd, D, Ellison, N; 2008; Social Network Sites: Definition, History, and Scholarship, Journal of Computer-Mediated Communication; The Oxford Handbook Of Internet Studies .Oxford: Oxford University Press, pp. 151---172.

Conover, D, Ratkiewicz, M, Francisco, B, Flammini, F, Menczer; 2012; Political Polarization on Twitter, Fifth International AAAI Conference on Weblogs and Social Media,; Barcelona.

Daniel, G, Nagarajan, M, Pieper, J, Robson, C, Sheth, A; 2009; Context and Domain Knowledge Enhanced Entity Spotting In Informal Text; The Semantic Web – ISWC; Chantilly, VA, USA.

Dean, J, Ghemawatl; 2008; S. MapReduce: Simplified data processing on large clusters. Communications of the ACM - 50th anniversary issue: 1958, Vol. 51, Issue 1, January, 2008, pp. 107-113.

Domingos, P; 2005; Mining Social Networks for Viral Marketing; IEEE Intelligent Systems.

El Beltagy, S, Ahmed, A; 2013; Open Issues in the Sentiment Analysis of Arabic Social Media: A Case Study; IEEE Innovations in Information Technology (IIT).

El Din, A, Al Taher, F; 2013; Sentiment Analyzer for Arabic Comments System; International Journal of Advanced Computer Science & Applications, Vol. 4, Issue 3, pp. 99-103. 5p.

Fadi , S, Mourtada, R; Jul. 2012; 'Social Media in the Arab World: Influencing Societal and Cultural Change?; Journal name: The Arab Social Media Report; Publisher: Governance and Innovation Program, Dubai school of government.

Gwahangno, D, Yuseong; 2010; What is Twitter, a Social Network or a News Media?; ACM, Proceedings of the 19th international conference on World wide web,; Raleigh, NC, USA; pp. 591-600.

Habash, N; 2001; Introduction to Arabic natural language processing; Synthesis Lectures on Human Language Technologies; Morgan & Claypool Publishers; 1-87.

Kim, D, Jo, Y, Moon, O, Oh, A; 2010; Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users; Proc. Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems (CHWE2010), Atlanta, Georgia.

Lin, L, Riosand, M; 2012; Distilling Massive Amounts of Data into Simple Visualizations: Twitter Case Studies; Twitter, Inc.

Mourtada, R, Salem, F; 2011; Civil Movements: The Impact of Facebook and Twitter, Journal name: The Arab Social Media Report; Publisher: Governance and Innovation Program, Dubai School of Government.

Microsoft, Arabic NLP Toolkit (ATK) 11/2012 حزمة أدوات اللغة العربية For Academia in the Arab World, Power Point Presentation.

Mihalcea, R, Liu, H, Lieberman, H; 2006; Nlp (natural language processing) for nlp (natural language programming) ; Springer, Verlag Berlin Heidelberg, LNCS 3878, pp. 319–330.

O'Connory, B; 2010; From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series; In Proceedings of the International AAAI Conference on Weblogs and Social Media Key: citeulike: 7044833.

Pang, P, Lee, L; 2008; Opinion Mining and Sentiment Analysis; Foundations and Trends in Information Retrieval archive Vol. 2, Issue 1-2, January 2008, pp. 1-135.

Patrick, L; 2011; Extracting Strong Sentiment Trends from Twitter; Springer, Computer Science Department Stanford University.

Piao, S, Whittle, J; A Feasibility Study on Extracting Twitter Users' Interests using NLP Tools for Serendipitous Connections; Published in: Privacy, security, risk and trust IEEE third international conference on and IEEE third international conference on social computing.

Reichheld. F; 2003; The one number you need to grow. Harvard Business Review, 81 (12):47 -54.

Shaalan, K, Farghaly, A; 2009; 'Arabic Natural Language Processing: challenges and Solutions', ACM Transactions on Asian Language Information Processing.

Shamanth, K, Morstatter, F, Liu, L; 2013; Twitter Data Analytics; Springer page 1. 123 briefs in computer science.

Stefano, M; 2000; Toward the semantic web a view of XML from outer space, Apache CON Europe - London, UK.

Sudha, V, Vieweg, S, Corvey, W, Palen, L, Martin, H, Palmer, M, Schram, A, Kennet, M; 2011; Natural Language Processing to the Rescue? "Extracting "Situational Awareness" Tweets During Mass Emergency; Fifth International AAAI Conference on Weblogs and Social Media, Barcelona.

Twitter - http://blog.twitter.com, 2012.

Vimal, M, Kositsyna, N, Austin, M; 2004; Requirements Engineering and the Semantic Web: Part II. Representation, Management, and Validation of Requirements and System-Level Architectures, ISR Technical Report.

Xiang, G, Fan, B, Wang, L, Hong, J, Rose, C; 2012; Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus; Proceedings of the 21st ACM international conference on Information and knowledge management; Hawii.

ملخص

عملية اكتشاف وفهم المستخدمين العرب للإعلام المجتمعي مهم جدا، حيث ان عملية فهم المستخدمين العرب تساعد اصحاب القرار باتخاذ القرار الصحيح بناءا على تطلعات واراء المستخدمين.

الاعلام المجتمعي مثل الفيس بوك وتويتر أصبح منتشر جدا بالعالم العربي وهنالك اكثر من اربعة مليون مستخدم على الاقل يقومون يوميا بكتابة تغريدات على تويتر متعلقة بالكثير من الامور مثل السياسة والاقتصاد والرياضة.

فكرة الرسالة قائمة على تحليل وتصنيف التغريدات العربية والمحكية باللغة العامية الاردنية، وهذا الموضوع لم يأخذ حقه بالنقاش وما زال يحتاج الى عمل كثير والتعمق به.

قمنا بهذه الرسالة بجمع ألف تغريدة عربية تتكلم عن ثلاث مجالات اساسية ومشهورة بالاردن وهيا موقع جملون الموقع الخاص بالكتب وتوصيلها، وموقع خبرني الموقع الخاص بالاخبار الاردنية والعالمية، وتلفزيون رؤيا الذي اصبح التلفزيون المحلي الاكثر مشاهدة في الاردن، الألف تغريدة كانت مقسمة على هذه 3 مجالات، تم استخراج كلمات سلبية وايجابية منها وعمل استبيان لعشرين شخص لوضع الاوزان المناسبة لكل كلمة مستخرجة، ومن ثم بمعاونة 3 أشخاص عرب تم تصنيف التغريدات لأيجابية وسلبية حتى يكون هذا التصنيف مقياس للمقارنة مع النتائج التي سنحصل عليها لاحقا من جراء استخدام الخوارزمية الخاصة بنا.

في هذه الرسالة تم بناء أول اونتولوجي عربي مختص بالاعلام المجتمعي العربي لجميع الكلمات العربية والسلبية وتم ربط الكلمات مع كلمات اخرى لها علاقة بها، وبناءا عليه تم عمل خوارزمية خاصة تأخذ بعين الاعتبار الاونتولوجي والاوزان لكل كلمة واوزان الكلمات المتعلقة بها.

في هذه الرسالة تم تطبيق وتجريب العمل على التغريدات وخرجنا بنتائج مبشرة وجيدة جدا وتمت المقارنة مع اعمال سابقة بهدف مقارنة نتائجنا مع اعمال اخرى.

هذا العمل يعتبر مقدمة ويمهد الطريق لأعمال أخرى مهتمة بمجال ربط السوق مع الاعلام المجتمعي العربي لمعرفة توجهات السوق الحقيقية.

فهم واكتشاف توجهات المستخدمين العرب لتويتر وربطها مع الأونتولوجي

بواسـطة
ابراهيم محمد عبد النبي

بإشـراف
د. سمير الترتير

قدمت هذه الرسالة استكمالاً لـمتطلبات الحصول على درجة
الـمـاجـسـتير في علـم الـحـاسـوب

عـمـادة البحث العلمي والدراسات العليا
جامعة فيلادلفيا

كانون الأول 2013