Media Engineering and Technology Faculty
German University in Cairo

Arabic Sentiment Analysis System

Bachelor Thesis

Author:	Ahmed Abdelghany Mohamed Ibrahim Madi
Supervisors:	Dr. Mohamed Ahmed El-Mahdy
Submission Date:	15 May, 2017

Media Engineering and Technology Faculty
German University in Cairo

Arabic Sentiment Analysis System

Bachelor Thesis

Author: Ahmed Abdelghany Mohamed Ibrahim Madi
Supervisors: Dr. Mohamed Ahmed El-Mahdy
Submission Date: 15 May, 2017

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree

(ii) due acknowlegement has been made in the text to all other material used

Ahmed Abdelghany Mohamed Ibrahim Madi
15 May, 2017

Acknowledgments

I begin by thanking Allah for giving me the strength and perseverance to complete this project till the very end, and for providing me with the knowledge needed to be able to work on it.

I would like to thank Dr. Mohamed El-Mahdy, for his encouragement and guidance throughout the period of this project,without which it wouldnt have turned the way it did.

Furthermore, I must express my very extreme gratitude to my parents and my family for providing me with continuous support and encouragement throughout the progress of my project. This work would not have been possible without them. Words are powerless to express my gratitude. Thank you.

Abstract

Microblogging is one of the official communication platforms for internet users, thus becoming a rich domain for sentiment analysis and opinion mining. We focus on Arabic due to the lack of focus on techniques devoted to Arabic sentiment analysis on data collected from Twitter, the most popular microblogging platform. This paper provides a comprehensive analysis on Arabic sentiment analysis with a data set of over 7000 tweets classified into positive, negative and neutral by sentiment classifiers. The paper goes through the in depth analysis of methods used for data gathering, partitioning and linguistic analysis of the corpus and further explains the outcomes.

Contents

Chapter 1

Introduction

1.1 Motivation

Every day numerous users post profusely, which is one of the signs of their attachment to the platform. Starting with the definition of Microblogging, it is a combination of blogging and instant messaging that makes the communication process much easier than when compared to the old days when physical interactions had to be made or even the traditional blogging. Some of the reasons microblogging has managed to become popular is that it less time is spent developing content providing the user the opportunity for frequent posts due to its character limit which differentiates it from regular blogging. Also, it is the easiest way to share urgent or time-sensitive information and mobile convenient as it is not too hard to write a blog or post using a smartphone or tablet[13, 2]. As the platform becomes more popular, it enables higher number and frequency of posts. This high demand on microblogging allows the introduction of a new term, which is Sentiment Analysis.

1.2 Project Aim

To every project there is a main goal, and a subset of goals that branch from the target achieved. This projects main goal is to build a sentiment analysis system with the following criteria of having different features such as unigram, bigram and trigram as well as achieving the highest rate of accuracy classification possible. Meanwhile, the subsidiary goals are to tackle the obstacles created by pursuing sentiment analysis for Arabic that include the lack of availability of relevant content. Also, another subsidiary goal is to highlight the recent relevance of Arabic sentiment analysis with the growing usage of the Arabic language in microblogging platforms.

1.3 Thesis Organization

Chapter 2

Background

2.1 Tools and Technologies

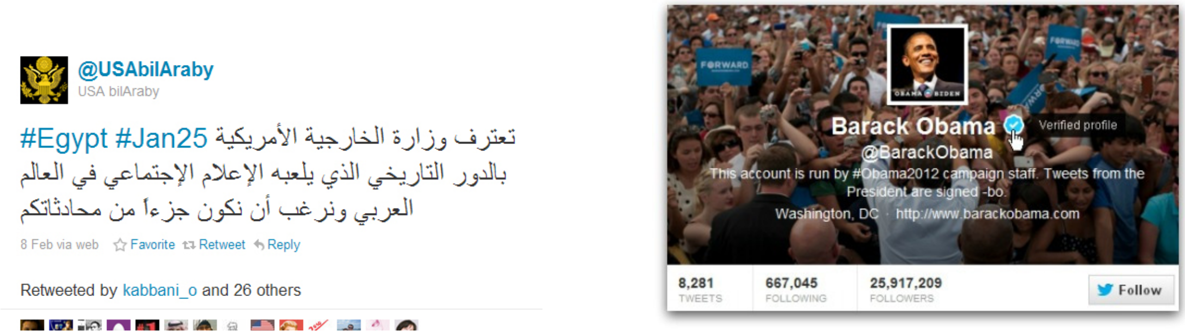
2.1.1 Sentiment analysis

Sentiment analysis - otherwise known as opinion mining - is the process of discovering and determining the emotional meaning or tone behind a series of words. It can help in various ways. Firstly, by helping a company discover the public opinion towards their company or products. Said opinion aids in quality management, tactic and strategy planning as well as any marketing improvements whether through the business, economical or political position of the company and its products. All these changes are made based on the sentiment score provided by the sentiment analysis. Secondly, it can help political parties predict the public opinion towards them or the impact their campaign is having in order to align their goals with the publics interests. Moreover, the entertainment segment can benefit from sentiment analysis through collecting fans feedback on the works of art by celebrities, authors, producers and so on as well as opinions on their public interactions whether through interviews or social media[12].

When comparing the sentiment analysis to manual analysis or surveys, it should be mentioned that the human brain is the most accurate machine on earth. On the other hand, retrieving information in a systematic or computational way is more efficient than manual analysis, surveys and such as it is more time efficient, cost efficient with a good score that can be beneficial in receptive tasks[6].

2.1.2 Twitter

There are multiple microblogging platforms including Jaiku and more recently Pownce. However, the most popular of microblogging platforms is Twitter. Twitter allows you to post statuses or updates that range to a maximum of 140 characters. By 2007, Twitter has managed to become one of the fastest rising microblogging platform with over 94,000 users.Twitters popularity was due to the social communities formed through the mutual interest of its bloggers. The bloggers tended to stay for longer periods when they received positive comments and lasting social friendships through the platform. Also, the diversity of content provided whether through personal blogs or public blogs managed by celebrities in different entertainment segments that include movies, sports, music and art or politicians seeking public approval was enough to capture the attention of the public and engage in the microblogging phenomenon. Moreover, the fact that celebrities have verified accounts on twitter makes its information highly reliable as opposed to other gossip or events cyber venue[1].



@USAbilAraby
USA bilAraby

#Egypt #Jan25 تعترف وزارة الخارجية الأمريكية
بالدور التاريخي الذي يلعبه الإعلام الإجتماعي في العالم
العربي ونرغب أن نكون جزءاً من محادثاتكم

8 Feb via web ☆ Favorite Retweet Reply

Retweeted by kabbani_o and 26 others

Barack Obama Verified profile
@BarackObama
This account is run by #Obama2012 campaign staff. Tweets from the President are signed -bo.
Washington, DC · http://www.barackobama.com

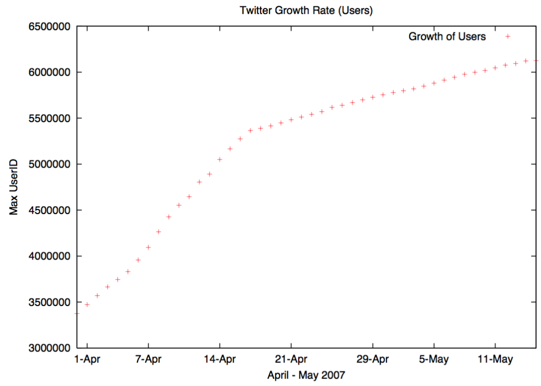
8,281 TWEETS 667,045 FOLLOWING 25,917,209 FOLLOWERS Follow

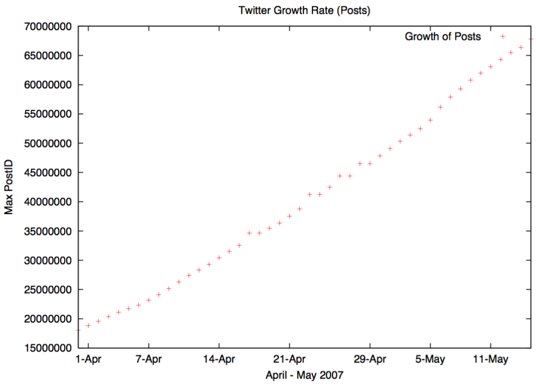
(a) .

(b)

Figure 2.1: T

Citation used Twitters API to show the rate of growth for users and posts for two months in 2007.Although twitter was launched in 2006, it gains popularity after it won the South by SouthWest (SXWS) conference Web Awards6 in march,2007. Despite the number of new users has slowed, the number of posts has increased.





2.1.3 Arabic Sentiment Analysis

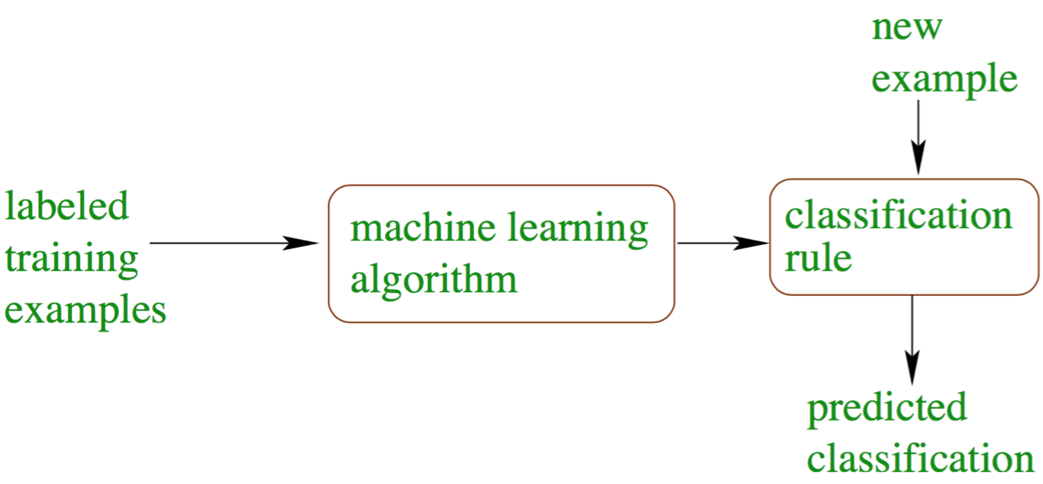
By 2016, Arabic was the fourth most commonly spoken language on Earth. The Arabic language is the official language for most Arabian countries including Egypt, Tunisia, Algeria, Saudi Arabia, Libya, Morocco and multiple gulf countries. It is widely spoken through the middle east and its relevance and importance is highlighted through the geographical and political affairs, as well as multiple works of art by authors such as Khalil Gobran that helped shape the Arab world in terms of culture. Also, the Islams Holy Quran is in Arabic and with Muslims high percentage population of the world with 1/6th of the worlds population, it creates a religious significance to the Arabic language.
Technologies have failed to incorporate the Arabic language in their work throughout the years, however more recently, many applications have used Arabic in search engines and document archiving tools, both known as core work. Furthermore, there has been a recent Arabic epidemic where the usage of the language has increased over the past few years. This increase is clear among social media platforms such as Facebook and for the sake of this topic, particularly Twitter. This further evolves the need for Arabic sentiment analysis and opinion mining[15].

2.1.4 Machine Learning

Machine learning enhances accuracy through allowing computers to modify or adapt their actions. These actions can include making predictions or controlling a robot. Said accuracy is evaluated and measured by how the correct actions are reflected in the chosen actions[10]. Machine learning recognize human reasoning and repeat thought patterns and strategies that add to the decision making process. It does so by producing classifying terms simple enough for humans to comprehend. Machine learning operations are resumed without the interference of humans, however background knowledge may be exploited to allow development of machine learning.[10]

For instance, a simple game of Scrabble can be used as an example. When you play against a computer, you start by winning. As the number of games progresses, the computer starts to win. The computer wins by understanding the pattern of strategies used by us and goes on to use these strategies against us and other players. By not starting from scratch every new player, it creates a form of generalization [10].

Machine Learning aims to generate classifying expressions simple enough to be understood easily by the human. They must mimic human reasoning sufficiently to provide insight into the decision process. Like statistical approaches, background knowledge may be exploited in development, but operation is assumed without human interventio[5].

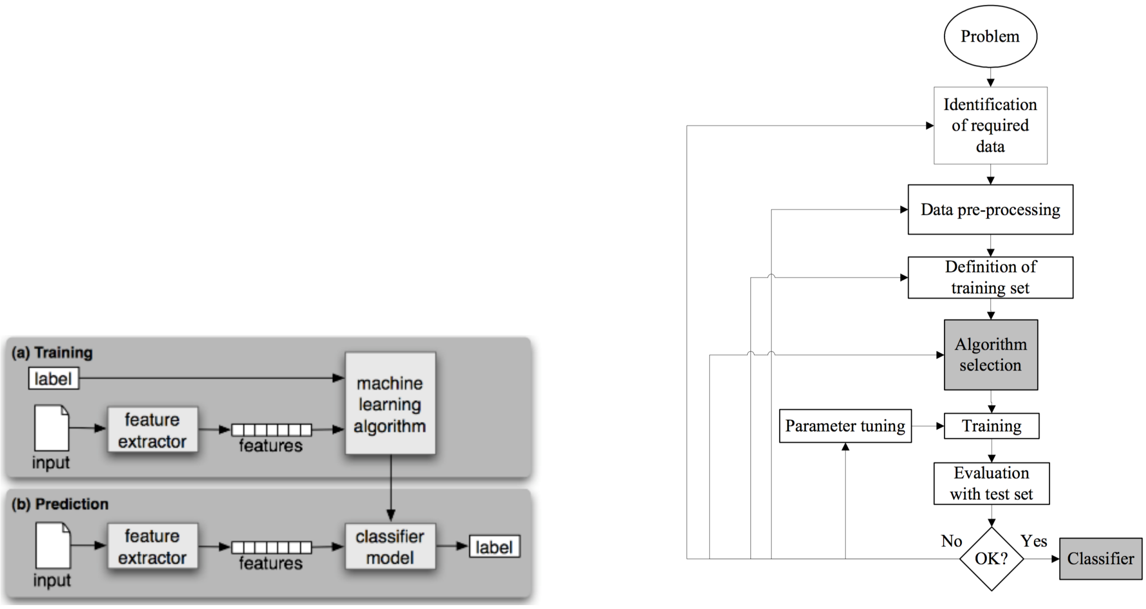


Machine Learning aims to generate classifying expressions simple enough to be understood easily by the human. They must mimic human reasoning sufficiently to provide insight into the decision process. Like statistical approaches, background knowledge may be exploited in development, but operation is assumed without human interventio[5].

2.1.5 Supervised Learning

There are two types of classification, one is known as Unsupervised Learning (or Clustering) and the other is Supervised Learning. Unsupervised learning is the classification of a data set through creating classes or clusters in the data. Meanwhile, Supervised Learning already acknowledges the presence of classes or clusters and aims to establish a rule that classifies newly introduced observations into one the existing classes. The term ”discrimination” also used when talking about Supervised Learning due to correctly constructing a classification rule from formerly classified data. If former data is classified with a high rate of accuracy, the assumption is that someone (The Supervisor) is able classify without error[16].

The function for supervised learning is derived from the training data, which is assembled from training examples.Every example is a pair of input and output, where input is usually a vector and output is the desired result, also known as the supervisory signal. According to (citation article supervised learning) the inferred function is called a classifier. The classifier should be able to generalize the training data in order to for an unknown input to have a correctly predicted output[5].



(a) .

(b)

Figure 2.2: The Supervised Learning Process

2.1.6 Python

According to(Dubois, 2007; Milmann and Avaizis, 2011), One of the most common languages used nowadays in scientific computing is Python[11, 7]. This goes back to its high-level interactive nature and its maturing ecosystem of scientific libraries. It is considered to be the first choice for algorithmic development and exploratory data analysis. Moreover, as a general-purpose language, it is increasingly used not only in academic settings but also in industry settings[8].

2.1.7 Scikit-Learn

scikit learn manages to provide state-of-the-art implementations of multiple popular machine learning algorithms. All the while, maintaining an user friendly interface in line with the Python language. The need for statistical data analysis is satisfied due to the friendly nature of the interface that allows many stakeholders to use it. The stakeholders include not only non-specialists in the software and web industries but also, fields outside of computer science, such as Biology or Physics[8].
There are many features that differentiate Scikit learn from other machine learning toolboxes in Python, some of which include, firstly, the BSD license under which it is distributed. Secondly, when compared to MDP and Pybrain, Scikit learn incorportes compiled code for better efficiency (Zito et al., 2008)[17]. And while Pymvpa has optional dependencies such as R and shogun, and iv, Scikit learn only depends on numpy and

scipy to facilitate easy distribution (Hanke et al., 2009)[9].

Also, another comparison of Pybrain to Scikit learn is the fact that Pybrain uses a data-flow frame work while Scikit learn focuses on imperative programming. While the package is mostly written in Python, it incorporates the C++ libraries LibSVM (Chang and Lin, 2001) and LibLinear (Fan et al., 2008) that provide reference implementations of SVMs and generalized linear models with compatible licenses. Binary packages are available on a rich set of platforms including Windows and any POSIX platforms[3, 14]. Furthermore, thanks to its liberal license, it has been widely distributed as part of major free software distributions such as Ubuntu, Debian, Mandriva, NetBSD and Macports and in commercial distributions such as the Enthought Python Distribution[8].

2.2 Related Work

As it was mentioned in 2.1.1, According to (Pang and Lee, 2008) the increase of blog and social network usage is directly proportional to the interest of researchers in opinion mining and sentiment analysis. They continue to discuss several techniques and strategy approaches for opinion-oriented information collection[12]. Although other works have ignored microblogging when discussing sentiment analysis, (Yang et al., 2007) used weblogs data where the users mood was determined through emotion icons assigned to blog posts and related comments. Next, for their research, sentiments were classified at sentence level through SVM and CRF learners. Various strategies were then discussed and analysed to set the overall sentiment of the document, when said sentiment matches the sentiment of the last sentence of the document, the strategy used is announced as the chosen strategy[4]

Another article states that emoticon included texts were collected with the purpose of forming a training set for sentiment classification. The data was collected from Usenet newsgroups. The dataset that was divided into positive subset and negative subset where the positive sample included texts with happy emoticons, while the negative samples included sad or angry emoticons. A 70% accuracy rate was achieved when SVM and Nave Bayes, both known as Emoticons-trained classifiers, were used. (Read, 2005)
Another approach used, similar to the one performed by Read 2005, where the data set was classified similarly with an 81% rate of accuracy through the Nave Bayes classifier. On the other hand, it showed poor performance results when a third class known as neutral was introduced. (Go et al., 2009)
(Schmid 1994) states that another method whose main interest is a difference of tags distribution between sets of text in English. It starts by checking the distribution of word frequency. Next, TreeTagger is used for English to tag all posts. This is a highlighted feature of the method due to the particular interest in difference of tags distribution between texts (positive, negative and neutral). Then, in order to train the sentiment classifier,

features are extracted from the dataset acquired. Nevertheless, (Pang el al 2002) have achieved better results when they used term presence instead of term frequency.

2.3 Previous Work

Others recently working on sentiment analysis for Arabic include (Nabil, Aly, Atiya, 2015). With over 84,000 tweets collected, leaving 54,176 Arabic tweets after filtration, they have divided their dataset into 4 data sets with positive, negative, subjective and subjective mixed categories. All the while, performing a standard partitioning to the data set using a wide range of standard classifiers to perform 4 way sentiment classification to avoid sentiment polarity classification problems.

Although the detection of user sentiment in text is recent and even harder to find in Arabic, there are still some articles that discuss sentiment analysis for Arabic. There are various processes for data collection and performing sentiment analysis, some of which include the SAMAR system, where multi domain datasets are exposed to subjectivity and sentiment analysis for Arabic social media. Said datasets are collected from Wikipedia TalkPages, Twitter and Arabic Forums. (Abdul Mageed et al., 2014). Similar to Nabil, Aly, Attiya, 2013, (El Sahar and El Beltagy, 2015) have also classified data into four datasets. This was intended for the purpose of building a multidomain Arabic resource also known as sentiment lexicon. All the while, (Nabil et al., 2014) and (El Sahar and El Baltagy, 2015) introduced a semi-supervised approach for building a sentiment lexicon that when used appropriately can be effective and efficient in sentiment analysis.

Another article by (Aly and Atiya, 2013) presented LABR for a book reviews dataset collected from GoodReads. Another article dataset related to entertainment was discussed by (Rushdi-Saleh et al., 2011) where 500 movie reviews were collected from multiple webpages and used as datasets for sentiment analysis. Moving on to works relevant to microblogging, (Refaee and Rieser, 2014) have managed to collect a dataset of 8,868 tweets that was manually annotated Arabic social corpus and they continue to review their method of collecting and annotating the corpus.

We tried to determine the best settings for the mi- croblogging data. On one hand high-order n-grams, such as trigrams, should better capture patterns of sentiments expressions. On the other hand, unigrams should provide a good coverage of the data. The process of obtaining n- grams from a Twitter post is as follows: 1. Filtering we remove URL links (e.g. http://example.com), Twitter user names (e.g. @alex with symbol @ indicating a user name), Twitter special words (such as RT6), and emoticons. - Tokenization we segment text by splitting it by spaces and punctuation marks, and form a bag of words. However, we make sure that short forms such as dont, Ill, shed will remain as one word.

- Removing stopwords we remove articles (a, an, the) from the bag of words.

- Constructing n-grams we make a set of n-grams out of consecutive words. A negation (such as no and not) is attached to a word which precedes it or fol- lows it. For example, a sentence I do not like fish will form two bigrams: I do+not, do+not like, not+like fish. Such a procedure allows to improve the accuracy of the classification since the negation plays a special role in an opinion and sentiment ex- pression(Wilson et al., 2005).

- We build a sentiment classifier using the multinomial Na ve Bayes classifier. We also tried SVM (Alpaydin, 2004) .

Chapter 3

Implementation

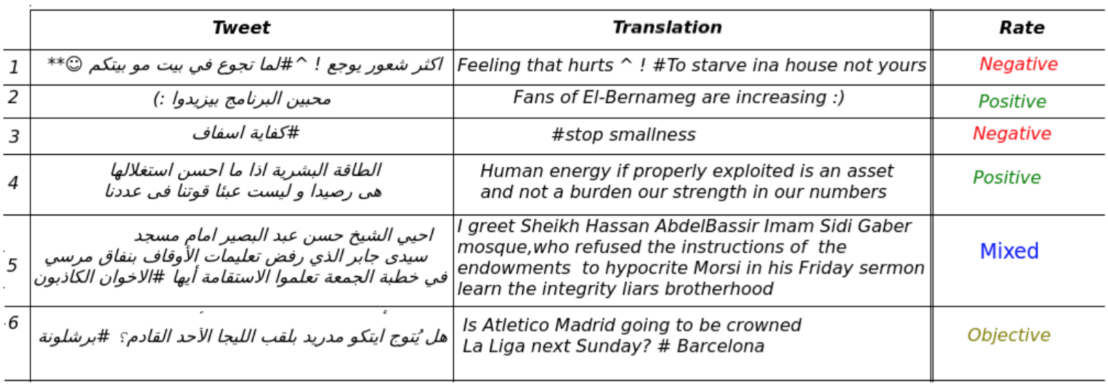
3.1 Dataset

3.1.1 Dataset Collection

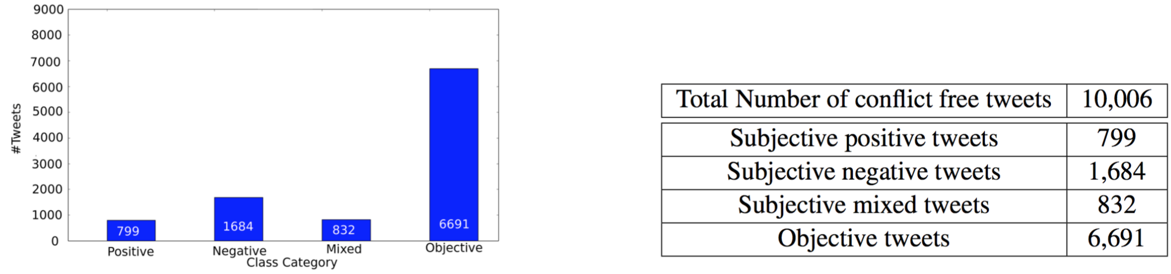
ASTD

We use the same dataset that (citation) have used, They divided their work to collect those tweets into three stages.Also, they have used high-tech technologies to collect those tweets. Starting with the first stage, which aimed to determine the most active Egyptian Twitter accounts with their recent tweets till November 2013.They ended up with almost 30 accounts and about 36,00 tweets. Proceeding to the second stage, they have crawled EgyptTrends 3, which is a Twitter page for the top trending hashtags in Egypt. They have got a list of 2500 unique hashtags which have used again to download more tweets.In this process, the number of tweets has increased to reach 84,000 tweets. Finally, they filtered those tweets and removed the non-Arabic tweets and redundant expressions like HTML, they ended up with 54,716 Arabic tweets.

After collecting those tweets, they moved on to another process which is annotating those tweets.Annotation is the process of adding information to already existing data. Whenever you have a large number of semi or unstructured data, Annotation is the best method to be use to convert those unstructured data into a well structured data [cite]. Also, another high-tech is used to automate this process. Amazon Mechanical Turk (AMT) service provides an API called Boto 4 used to annotate those tweets. They have used four categories through this operation, which are: Objective, Subjective Positive, Subjective Negative and Subjective mixed. Another filtration has used in this process when a tweet has assigned to three or more labels, it has considered being unuseful and has ignored. Another conflict has accepted for further processing and has considered being conflict-free has occurred when a tweet is assigned to two labels. Finally, they have ended up with 10k labeled tweets.



	Tweet	Translation	Rate
1	اكثر شعور يوجع ! ^#لما تجوع في بيت مو بيتكم ☺**	Feeling that hurts ^ ! #To starve ina house not yours	Negative
2	محبين البرنامج بيزيدوا :)	Fans of El-Bernameg are increasing :)	Positive
3	#كفاية اسفاف	#stop smallness	Negative
4	الطاقة البشرية اذا ما احسن استغلالها هى رصيدا و ليست عبئا قوتنا فى عددنا	Human energy if properly exploited is an asset and not a burden our strength in our numbers	Positive
5	احيي الشيخ حسن عبد البصير امام مسجد سيدى جابر الذي رفض تعليمات الأوقاف بنفاق مرسي في خطبة الجمعة تعلموا الاستقامة أيها #الاخوان الكاذبون	I greet Sheikh Hassan AbdelBassir Imam Sidi Gaber mosque,who refused the instructions of the endowments to hypocrite Morsi in his Friday sermon learn the integrity liars brotherhood	Mixed
6	هل يُتوج ايتكو مدريد بلقب الليجا الآحد القادم؟ #برشلونة	Is Atletico Madrid going to be crowned La Liga next Sunday? # Barcelona	Objective



(a) .

Total Number of conflict free tweets	10,006
Subjective positive tweets	799
Subjective negative tweets	1,684
Subjective mixed tweets	832
Objective tweets	6,691

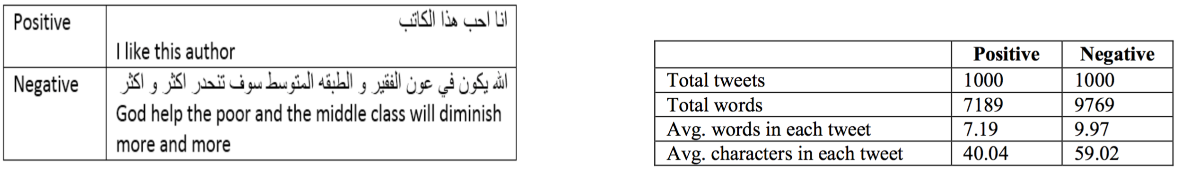
(b)

Figure 3.1: The Supervised Learning Process

The Tweets dataset

(citraion) worked on the same topic and used a tweet crawler. Unfortunately, they did not mention it. They have collected 2000 labeled tweets (1000 positive and 1000 negative ones) which were written in both Modern Standard Arabic (MSA) and the Jordanian dialect. They did not stick to one topic, they tried to cover various topics such as politics, arts and sports. Also, they did not use any automated system to annotate the data. Tweets were annotated using human experts, Exactly two human experts were used to label those tweets. During the process of categorizing them, if they agreed on the same label of a tweet, it was assigned to directly to its class. If a conflict has occurred, a third expert was used to label this tweet.

After collecting the data, they moved to another procedure which is dataset preparation. They have started with separating each tweet in a single file and split those tweets to be in a positive or negative folder. Then, they have used a naive algorithm to remove the repeated letter (e.g., muuuuuuch =) and the MS Word dictionary as a reference to correct misspelled words.The naive algorithm worked as follow, it started with counting the number of letter in each word and checked whether this number exceeded five or not. If a letter's occurrences exceeded five, it eliminated the repeated letters and looked it up in the MS word dictionary.



Positive	انا احب هذا الكاتب I like this author
Negative	الله يكون في عون الفقير و الطبقه المتوسط سوف تنحدر اكثر و اكثر God help the poor and the middle class will diminish more and more

(a) .

	Positive	Negative
Total tweets	1000	1000
Total words	7189	9769
Avg. words in each tweet	7.19	9.97
Avg. characters in each tweet	40.04	59.02

(b)

Figure 3.2: The Supervised Learning Process

Finally, they tried to enhance the dataset's quality by removing stop-words. A combined list of stop-words, which was obtained from the Khoja stemmer tool and added manually from different Arabic dialects was used to remove that stop-words from the whole tweets. Another enhancement was used, which was normalized some letters to one shape.

3.1.2 Reading Dataset

Those were the datasets that I worked on.Now, moving on to the first step of my implementation, which was reading those datasets correctly and modify them.

ASTD

ASTD was represented as a file contains the whole tweets with its labels.Each tweet is represented by one line with its corresponding label. The first part of each line is the tweet itself and the rest is the label of its tweet. The aim of this part was to read the tweets and its labels without any redundant expressions(e.g,the new line and tab epressions) and combine the objective and subjective mixed classes into one class and considered them as a ”Neutral” class.

The Tweets Dataset

The Tweets dataset was represented as folder contains two files, the first one is positive which contains the positive tweets and the another one is negative which contains the negative tweets as it was mentioned in ().The procedure went the same way as ASTD one, except that there are only two classes. There is no need to combine two classes into one class.Yet, there were some bad encoding of some tweets which was ignored during the reading process.

3.2 Feature Extraction

Feature extraction is considered to be the most difficult part in sentiment analysis. As it is the major procedure that every procedure depends on. Feature extraction is the process of transforming the input data into a set of features. Features are set of distinctive properties of the input that help in differentiating between the categories of input patterns. In this section,different approaches of feature extraction will be discussed .In addition to some usful datasets' prepration tehhniques.

3.2.1 Removing Hashtags

Hashtags is a major issue that is faced when dealing with Tweets' datasets. The hashtag symbol is commonly used by people before a relevant keyword or phrase in their Tweet to categorize those Tweets or to search for related topics or Tweets to this hashtag.Hashtagged words that become very popular are often Trending Topics.For the sake of our topic Sentiment Analysis, Hashtags are redundant expressions that could be useless or might decrease the accuracy of the models.As it can be included in most of those Tweets.So, removing them is considered to be a good start for feature extraction.This work is done by the "remove-hashtags" function which takes all the words in those Tweets as an input, then it proceeds with removing any hashtag symbol that comes before any word.

3.2.2 Removing Punctuation

There is no doubt that punctuation is used to create sense, clarity and stress in sentences. Punctuation marks are used to structure and organize what you want to write.Those signs and symbols are given to the reader to show how a sentence is constructed, how it should be read and make the meaning clear. On the other hand, sentiment analysis does not take into account the effect of those marks.It may weaken the system and without those marks, the system performance will be more efficient.The remove-punctuation function handles those marks, which convey no meaningful sentiment and removes them wherever they are found.

Chapter 4

Experimental Results

Chapter 5

Conclusion

Conclusion

Chapter 6

Future Work

Text

Appendix

Appendix A

Lists

List of Figures

Bibliography

[1] Tim Finin Belle Tseng Akshay Java, Xiaodan Song. Why We Twitter: Understanding Microblogging Usage and Communities. *Proceedings of the Ninth WebKDD and First SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07) (pp. 5665). New York: ACM Press.*

[2] Patrick Paroubek Alexander Pak. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Universite de Paris-Sud, Laboratoire LIMSI-CNRS,Laboratoire LIMSI-CNRS,Orsay Cedex, France.*

[3] C.C. Chang and C.J. Lin. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001.

[4] Kevin Hsin-Yih Lin Changhua Yang and Hsin-Hsi Chen. Emotion classification using web blog corpora. In WI 07: Proceedings of the IEEE/WIC/ACM Interna- tional Conference on Web Intelligence. *Washington, DC, USA. IEEE Computer Society.*, pages 275–278, 2007.

[5] C.C. Taylor D. Michie, D.J. Spiegelhalter. *Machine Learning, Neural and Statistical Classification.* 1994.

[6] Kalina Bontcheva Diana Maynard. Challenges of Evaluating Sentiment Analysis Tools on Social Media. *University of Sheffield, Department of Computer Science Regent Court, 211 Portobello, Sheffield, S1 4DP, UK.*

[7] P.F. Dubois. Python: Batteries Included. *IEEE/AIP*, 9, 2007.

[8] Alexandre Gramfort Vincent Michel Bertrand Thirion Olivier Grisel Mathieu Blondel Peter Prettenhofer Ron Weiss Vincent Dubourg Jake Vanderplas Alexandre Passos David Cournapeau Matthieu Brucher Matthieu Perrot Fabian Pedregosa, Gae l Varoquaux and E douard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2012.

[9] P.B. Sederberg S.J. Hanson J.V. Haxby M. Hanke, Y.O. Halchenko and S. Pollmann. PyMVPA: A Python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, 7(1):37–53, 2009.

[10] Stephen Marsland. *MACHINE LEARNING An Algorithmic Perspective.*

[11] K.J. Milmann and M. Avaizis. Scientific Python. *IEEE/AIP*, 11, 2011.

[12] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(12):1–135, 2008.

[13] Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau. Sentiment Analysis of Twitter Data. *Department of Computer Science,Columbia University,New York,USA*.

[14] C.J. Hsieh X.R. Wang R.E. Fan, K.W. Chang and C.J. Lin. LIBLINEAR: a library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[15] Ahmed Ali Samhaa R. El-Beltagy. Open issues in the sentiment analysis of Arabic social media: A case study. *IEEE Innovations in Information Technology (IIT), 2013 9th International Conference*, 2013.

[16] Reddi Sanjeev Kumar Shirdi Wazeed Baba1. Data Mining: Text Classification System for Classifying Abstracts of Research Papers. *International Journal of advance Research,Ideas and Innovations In Technology*, 2, 2016.

[17] L. Wiskott T. Zito, N. Wilbert and P. Berkes. Modular toolkit for data processing (MDP): A Python data processing framework. *Frontiers in Neuroinformatics*, 2, 2008.