Microblogging platforms are one of the most high-tech platforms nowadays. 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. 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.

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 public’s 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.

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.

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. According to (Java et al: 1;2), Twitter’s 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.

According to (enter citation) 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 Islam’s Holy Quran is in Arabic and with Muslim’s high percentage population of the world with 1/6th of the world’s 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 (Neal 2013)(Farid 2013)

**BACKGROUND**

With the population of blogs and social networks, opinion mining and sentiment analysis became a field of interest for many researches. A very broad overview of the ex- isting work was presented in (Pang and Lee, 2008). In their survey, the authors describe existing techniques and approaches for an opinion-oriented information retrieval. However, not many researches in opinion mining consid- ered blogs and even much less addressed microblogging. In (Yang et al., 2007), the authors use web-blogs to con- struct a corpora for sentiment analysis and use emotion icons assigned to blog posts as indicators of users’ mood. The authors applied SVM and CRF learners to classify sen- timents at the sentence level and then investigated several strategies to determine the overall sentiment of the docu- ment. As the result, the winning strategy is defined by con- sidering the sentiment of the last sentence of the document as the sentiment at the document level.

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. Although other works have ignored microblogging when discussing sentiment analysis, (Yang et al., 2007) used web-blogs data where the user’s 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.

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 Naïve Bayes, both known as Emoticons-trained classifiers. (Read, 2005)

J. Read in (Read, 2005) used emoticons such as “:-)” and “:- (” to form a training set for the sentiment classification. For this purpose, the author collected texts containing emoti- cons from Usenet newsgroups. The dataset was divided into “positive” (texts with happy emoticons) and “negative” (texts with sad or angry emoticons) samples. Emoticons- trained classifiers: SVM and Na ̈ıve Bayes, were able to ob- tain up to 70% of an accuracy on the test set.

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 Naïve Bayes classifier. On the other hand, it showed poor performance results when a third class known as “neutral” was introduced. (Go et al., 2009)

In (Go et al., 2009), authors used Twitter to collect train- ing data and then to perform a sentiment search. The ap- proach is similar to (Read, 2005). The authors construct corpora by using emoticons to obtain “positive” and “neg- ative” samples, and then use various classifiers. The best result was obtained by the Na ̈ıve Bayes classifier with a mutual information measure for feature selection. The au- thors were able to obtain up to 81% of accuracy on their test set. However, the method showed a bad performance with three classes (“negative”, “positive” and “neutral”).

**4. Corpus analysis**

First, we checked the distribution of words frequencies in the corpus. A plot of word frequencies is presented in Fig- ure 1. As we can see from the plot, the distribution of word frequencies follows Zipf’s law, which confirms a proper characteristic of the collected corpus.

Next, we used TreeTagger (Schmid, 1994) for English to tag all the posts in the corpus. We are interested in a dif- ference of tags distributions between sets of texts (posi- tive, negative, neutral).

The collected dataset is used to extract features that will be used to train our sentiment classifier. We used the presence of an n-gram as a binary feature, while for general informa- tion retrieval purposes, the frequency of a keyword’s occur- rence is a more suitable feature, since the overall sentiment may not necessarily be indicated through the repeated use of keywords. Pang et al. have obtained better results by using a term presence rather than its frequency (Pang et al., 2002).

We have experimented with unigrams, bigrams, and tri- grams. Pang et al. (Pang et al., 2002) reported that uni- grams outperform bigrams when performing the sentiment classification of movie reviews, and Dave et al. (Dave et al., 2003) have obtained contrary results: bigrams and tri- grams worked better for the product-review polarity classi- fication. 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 “RT”6), 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 “don’t”, “I’ll”, “she’d” 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 ̈ıve Bayes classifier. We also tried SVM (Alpaydin, 2004) and