Journal of Theoretical and Applied Information Technology
15th October 2017. Vol.95. No 19
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COMPARISON STUDY OF AUTOMATIC CLASSIFIERS PERFORMANCE IN EMOTION RECOGNITION OF ARABIC SOCIAL MEDIA USERS

[1] ABDULLAH DAOOD, [2] ISSAM SALMAN, [3] NADA GHNEIM

[1]Department of Web Sciences, Syrian Virtual University, Syria
[2]Czech Technical University in Prague, Czech Republic
[3]Department of Web Sciences, Syrian Virtual University, Syria

Email: [1]Abdullah_64730@svuonline.org,[2]issam.salman@fjfi.cvut.cz,[3]t_nghneim@svuonline.org

ABSTRACT

Emotion recognition from text gained a lot of interest in the last years, but some languages such as Arabic (with its different spoken dialects) have not been given such attention. In this paper, we present our work in the Emotion detection of Arabic texts, with a focus on Levantine Twitter Messages. We have constructed a corpus of Arabic Levantine tweets, and annotated it with correspondent emotions. We implemented different methods to automatically classify text messages of individuals to infer their emotional states. We compared the results of different machine learning algorithms, and the inclusion of different features, to determine the best configuration of the emotion recognition system.

Keywords: *Emotional Analysis, Data Mining, Emotion Detection From Arabic Text, Twitter, Syrian Dialects.*

1. INTRODUCTION

The field of "sentiment analysis" and "Emotion Analysis" from text is one of the promising fields in “Text Mining" for its services and great benefits to decision makers in political, social, financial and marketing institutions. This field has doubled its importance with the growing of social networks and the feedbacks they provide to users, through which they express their opinions and evaluation of the products they buy, the books they read and the services they receive. Publications and events on Social networks is prone to judgments and polar views. The attempt to attract the wide electronic public is one of the first tasks of the marketing companies and the various institutions that are trying to build a good reputation to win more customers and achieve more profits and benefits. This is difficult to achieve without effective tools to understand the opinions of customers towards any service or product or performance carried out by these institutions. Therefore, the field "Sentiment Analysis" and "Emotion Analysis" from text has gained great importance by researchers and scholars to build tools and techniques that help business organizations to build their reputation based on a deep understanding of the directions and tendencies of their customers.

In the matter of English language, there are various tools for word processing, these tools are able to determine the beginning and the end of each sentence in the text, as well as analyse the sentence to its components (verb, subject and object). In addition to these tools, there are dictionaries and applications, which are often available in English. This is very helpful for research activity in this field.

As for Arabic language and in particular the spoken dialects, research has been limited to few studies on specific subjects, which often remain in the scope of characterization, comparisons and suggestions without constituting a practical basis for building real and effective applications in the treatment of this language. The reason for this is the lack of appropriate tools that help researchers and the difficulty of Arabic grammar in addition to the lack of familiarity of most researchers with this grammar.

In general, we can say that the main task in sentiment analysis of text is to categorize the sentiment significance of a text in a document or sentence if the expressed opinion is positive, negative or neutral. It should be noted that the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

determination of the sentiment at sentence level suffers from the difficulty of relying heavily on contextual words, but when working at document level the difficulty lies in the fact that one document may contain a set of contradictory views on the same objective.

In particular, the advanced task of recognizing the general sentiment of the text is to specialize in emotion analysis to reveal deeper human emotional connotations such as "anger, disgust, trust, sadness, joy, surprise". In this paper, we will present our approach to identify the emotional category of texts written in Arabic (Syrian dialect), using natural language processing tools, mathematical science, methods and methodologies of classification.

We focused in our research on studying user's tweets on Twitter. We used more than 1320 tweets and comments in Arabic Language- Syrian dialect. This tweets and comments were automatically collected and manually classified to build an emotional classification model that achieved 66.9% F-measure while using 10 Folds Cross-Validation**.**

2. CHALLENGES AND IMPORTANCE OF RESEARCH

The main challenge in Sentiment Analysis from Arabic text is its complexity either from syntactical level including the convoluted Arabic grammar, or from the emotional level. There are different words derivations and there is an impact of the diacritics marks on the word's meaning. Besides, there are other challenges related to the usage of the stem, in which a word's stem provides words with different meaning: "The same stem could result in some new different words when adding (prefixes, suffixes, or infixes)". Moreover, there is a lack of emotional state analysers, Arabic dictionaries, accurate morphological analysers, Part of Speech Taggers and syntactic parsers that can specify word tags such as, verbs, subjects, objects...etc.

3. RELATED WORK

Current researches in the domain of the automatic emotion detection of a text differ according to the following points:

a. The expected categories: Emotions (sadness, joy, surprising, disgust, fear, anger, …), Sentiment or Opinion (positive, negative, neutral).

b. Classification levels (sentence, clause, text).

c. Features that should be considered while classifying and the classification method adopted.

c-1. Semantic: considering the meaning of the word.

c-2. Syntactic: considering the word's structure and the sentence structure. (Stem's frequency, N-gram word's frequency, Punctuations).

d. Stylistic: where symbolic meaning, rhythm and word's weighting should co considering.

A. Salem & Chain [1] used many semantic, syntactic and stylistic features to classify English and Arabic sentiment on the official web "Official language, basic not spoken". Data set was 1300 samples. Instead of using semantic features, they used Entropy and achieved a 90% accuracy using SVM and merging syntactical and semantic features.

R. Obeidat & Ral-Shalabi [2] applied "KNN K-Nearest Neighbour" classification to reveal the category that a document belongs to according to the written text. They trained this classification on training set (1445 tweets in official Arabic language). This research concluded that using N-grams is 6-7% better than using traditional indexing within their categories.

Zreik &Hajjar [3] advised different kinds of morphological analysis for Arabic words in the case of information mining in Arabic language. They compared N-gram method with indexing method that deal with stems, then they provided a hybrid method from them.

On the other hand, (Taner Danis man & Adil Apkocak [4]) and (Hyo Jin Do & Ho-Jin Choi [5]) offered two different researches to detect "feelings" from a text. Training data had been represented as a vector space model depending on dictionaries built directly from sample data. In every research, weighting method was used to arrange the sample data in many different ways. In [4] the weighting method was applied using an equation like TF-IDF that takes in consideration every word's frequency within a document and its frequency within each category, in addition to the number of documents within every sentiment category. The weighting method in [5] is the same as [4] except that it did not take word's frequency within every category in consideration.

Journal of Theoretical and Applied Information Technology
15th October 2017. Vol.95. No 19
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In [4] they used data set (801 samples in English) for training “SVM and Naïve Bayes” classifiers using 10-Folds cross validation. In [5] they used data set (801 samples in Korean) for training SVM classifier using 10-Folds cross validation. Both [4] and [5] trained classifiers in order to classify the emotional significance of the tweet "Anger, fear, joy, sorrow and disgust".

When we talk about the tasks of finding sentiment according to the expected categories “Emotions, feeling”, I find a research entitled "Enhanced Sentiment Learning Using " A. Rappoprt et al. [6] presented an enhanced sentiment learning using twitter hashtags and smileys. They used KNN algorithm to classify emotions "happiness, sadness, anger” in English Tweets, and they used 50 Twitter tags and 15 smileys as sentiment labels which allow to build a classifier for dozens of sentiment types for short textual sentences.

On the other hand, J. Fing & L. Barbosa, [7] used SVM to classify sentiments[1] in tweets. They extracted words that determine sentiments from sample data and create a lexical glossary that contains these words with the feeling's degree for each word.

Another important work entitled “Mining Arabic Business Reviews" by Elfeky & Elhawary [8], where they used the official Arabic business website as source for the input data. Sentiment in these journals had been analysed but no evaluation has been done in this study.

4. DATASET

One of the most important challenges in this research is the lack of suitable Arabic datasets and Lexicons which serve the subject of the research. We found more than one Arabic dataset and lexicon, but it was in other dialects, such as Jordanian [11] and Saudi [12], and annotated with the sentiment, not the emotion of the user.

We built a web application which uses LINQ_To_TwitterCP library. This application pair with Twitter Search Application Programming Interface (API)[2] for corpus collection, which allows harvesting a stream of real-time tweets by querying their content in order to retrieve tweets. Then we filtered out the tweets to keep those written in Arabic- Syrian dialect with an emotional connotation in their content. We have a dataset of 1708 tweets, 1320 tweets for training and 388 tweets for testing.

Examples of Training “1320 instances”:

	Sad	Joy	surprise
Instances	220	220	220
	Disgust	Fear	Anger
Instances	220	220	220

Examples of Testing: 390 instances, ~32% of training set.

5. MEHODOLOGY

In this research, we will extract the emotions (Sad, Joy, Surprise, Disgust, Fear, Anger) from tweets written in Syrian dialect. The syntactical structure of a word will be considered by using N-grams (on the full-form words, and on the stemmed-form words).

After collecting the tweets for each emotions category as dataset, we processed them by:

- Eliminating @Users, #hashtags and URLs.
- Eliminating repeated letters, such as راااائع، حزيين
- Normalization (such as Hamza “أ” standardization)

Then we filtered out the dataset by filters which exclude the punctuation marks and stop words. All of this aims to decrease the dimension length of the word in the dataset.

The tweets were represented as feature vector taking the tweet’s words as features in this vector. We improved the feature vector to be represented by the six emotions categories to reduce the vector size.

Befor all of that, it was so important to collect several dictionaries. This dictionaries had been collected during our work. The words of these dictionaries had been taken from textual training data in addition to common Arabic words that social media users use.

5-1. Dictionaries

a. Negation dictionary

The method of negation in Arabic is divided into two parts:

- Open or explicit.
- Implicit or non-apparent denial.

Explicit is the negation in which one of negation words is used. The most famous of

[1] negative, positive, neutral.
[2] https://dev.twitter.com/

© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

these words are: "No – لا", "Not – ليس", "Non – لم" and "غير".
The implicit negation is a negation which is formulated without the words of negation, which carries the meaning of the explicit negation, but it comes with other words, such as conditional or questioning includes the meaning of negation, as: "لمّا", "لو" and "لولا".
We have built a dictionary that contains a list of Arabic negation words (about 70 words).

b. Modifiers dictionary

We mean by the modifiers' words, the set of words that may affect the intensity of the emotional significance of the subsequent or previous word in the tweet, such as: "much - كثيرا", "never – أبدا".
We have created a basic dictionary of these words (about 30 words) and we are contiously updating them.

c. Emoticons dictionary

We built an emoticons lexicon (about 110 emoticons-emotion), where each emoticon-emotion has impact on the emotions category by a determined value. (See figure 1)

	sad	joy	surprice	Disgust	fear	anger
:-(	1	0	0	0	0	0
:(	1	0	0	0	0	0
):	1	0	0	0	0	0
>:O	0	0	0.25	0	0	0.75
:O	0	0	1	0	0	0
>:(	0.5	0	0	0	0	0.5
:/	0	0	0	1	0	0
:'(	1	0	0	0	0	0
3:)	0	1	0	0	0	0
O:)	0	1	0	0	0	0
-_-	0.25	0.75	0	0	0	0
o.O	0	0	0.5	0	0.5	0
:-)	0	1	0	0	0	0
:)	0	1	0	0	0	0
(:	0	1	0	0	0	0
:V	0	1	0	0	0	0
^_^	0	1	0	0	0	0
8-)	0	1	0	0	0	0

Figure 1: An example of the Emoticons dictionary

d. Cursing Words dictionary

In Arabic, the words of cursing can be used to negatively affect the emotional significance of the sentence in order to indicate anger or disgust.
We have built a basic dictionary that contains a list of Arabic cursing words (about 100 words) as "Trifle - تافه", "Dirty - قذر", "Selfish – أناني" and "Varmint - حقير".

e. Emotion words dictionary "COR-Emotional lexicons"

We need to take into consideration the emotion that each word implies. Therefore, for each word in the training data "tweets", we calculated the weight of that word within each emotion category. Then, we built a set of dictionaries that indicate the weight of the words within the emotions category.

In this research, we compared among three methods to measure this weight.

1- Tf-IDF Term Frequency – inverse Document Frequency [4] :
"Q1" number of the term [x] in tweets which refer to emotion [e]. "Q2" number of the terms within tweets which refer to emotion[e].

$$fij = {}^{Q1}/_{Q2}$$

"N" number of all tweets within training data. "dfi" number of all tweets which contain term [x] within training data.

$$Freq\ of\ term_x\ in\ emotion_e = \left(\frac{fij}{max_{fij}}\right) * \log\left(\frac{N}{dfi}\right)/\log(2)$$

2- Weighed-TwF [5] :

"D" number of all tweets within training data. "ne" number of the tweets which refer to emotion [e], and contain term [x]. "Q" number of all tweets which refer to emotion [e].

$$\text{Normalized\_Tweet\_Frequency} = {}^{ne}/_{Q}$$

$$\begin{cases} (ne < D)\ \ then\ weight = {}^{1}/_{ne} \\ \quad else\ \ weiht = 0 \end{cases}$$

$$Freq\ of\ term_x\ in\ emotion_e = NTF * weight$$

3- We modified TF-IDF to be:

"Q" number of tweets which refer to the emotion [e].

"Z" number of the term [x] within tweets which refer to emotion [e]. "V" number of the terms within all tweets.

$$Freq\ of\ term_x\ in\ emotion_e = \frac{Z}{Q}/V$$

Journal of Theoretical and Applied Information Technology
15th October 2017. Vol.95. No 19
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

We also used "Saif Dictionary" [9], NRC-Emotion-Lexicon-v0.92-InManyLanguages, to compare its effect on the tweets emotion identification.

5-2. Classification Algorithms

We traind three types of classifiers using 10 Folds Cross-Validation:

- SVM(SMO) "Support Vector Machine".
- NB "Naive Bayes".
- CRF "Conditional random fields".

The result of the classification algorithm models was evaluated according to the "Precision, Recall and F-measure":

Precision = #correct guesses / #total guesses
Recall = #correct guesses / #total
F-measure = 2PR / (P+R)

where #correct guesses is the number of statements marked correctly as expressing an emotion X by the classifier, #total guesses is the total number of statements that are marked by the classifier as expressing the emotion X (including correct and wrong guesses), and #total is the number of statements expressing the emotion X in the dataset.

5.3. Vector Space Modelling

The tweets are represented as a feature vector, by considering each word in the tweets as an attribute in the feature vector. Its value may take multiple forms:

- The simplest form represents each word by a number that reveals the term frequency in the source text "tweet". The process starts by taking all the unigrams and bigrams from tweets then keeping those which exceed a certain frequency to be represented as features in the final vector.
- The second form is more advanced, as when assigning weights to the words within the training data, we take into consideration both of syntactic features (n-grams frequency, total number of the words and characters) & stylistic features (stems frequency and punctuation marks).

Actually, the problem with all of these methods is the high dimensionality of the feature vector. (See figure 2)

N:	وثق	الله	غفر	رحم	راود	قلق	غريب	حيال	أزمة	سورية	..	..	..	Emotion
1	0.45	0.78	0.89	0.9	0	0	0	0	0	0	0	0	0	#ثقة
2	0	0	0	0	0.72	0.56	0.98	0.55	0.345	0.24	0	0	0	#قلق
3	..	..	..	..	..	..	..	..	..	..	..	..	..	..
4	..	..	..	..	..	..	..	..	..	..	..	..	..	..

Figure 2: An example of textual feature vector

The solution was to reduce the feature vector containing all training data words to a feature vector that consists of six features only "a feature for each emotion category". The value of each feature will be an accumulative number that stands for the total weighted sum of sentences' words according to the emotion category that this value refers to. For example, the following feature vector extremely indicates "surprise".

"*انا متفاجئ انو الوقت رح يمر بشكل جيد*" which mean

"*I am surprised that the time will pass well*". (See figure 3.)

	Sad	Joy	Surprise	Disgust	Fear	Anger
فاجئ - surpris	0	0	0.8662509	0	0.2993424	0
وقت - time	0	0.3415608	0	0.2894182	0.2993424	0
يمر - pass	0	0	0.2887503	0	0	0
شكل - state	0.6516933	0	0.5775006	0.5788364	0.2993424	0.8213903
جيد - well	0	0	0	0	0	0
F. vector	0.6516933	0.3415608	1.732502	0.8682545	0.8980272	0.8213903

Figure 3: An example of the modified feature vector

5.4. Factors affecting our feature vector

5.4.1. Modifiers

When a modifier word appears in a sentence, it may take many forms and can be placed at different locations in the sentence. For example, we can write:

- "*أنا واثق من النجاح كتير*"
- "انا كتير واثق من النجاح"
- "*أنا واثق كتير من النجاح*"

which all mean "*I am very confident of success*".

The algorithm detects the emotion category "E_{MAX}" which refer to "the attribute that has the maximum value within the feature vector ". Then it doubles the value which refer to emotion category "E_{MAX}" for both previous and following word of the Intensifier (here the word كتير).

Figure 4 presents the feature vector of the sentence:

" اناكتير *متفاجئ انو الوقت رح يمر بشكل جيد* "

wich mean "*I am very surprised that the time will pass well*", (compare with Figure 3).

© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	Sad	Joy	Surprise	Disgust	Fear	Anger
F. vector	0.6516933	0.3415608	4.620004	0.8682545	0.8980272	1.642781

Figure 4: Feature vector for the sentence containing Intensifier

5.4.2. Emoticons

After extraction and analysing the final attributes vector of the tweet, the analytical algorithm re-checks the tweet to insure the existence of any emoticons within it. If emoticon [S] appears, the algorithm will increment the value [V] of the cumulative counter [C] of each emotional category [i] by the result of the multiplication of [V] value of this counter with the impact strength of this emoticon [SM] on this emotional category [i].

$$V_{C_i} = V_{C_i} + (V_{C_i} * SM_i)$$

The following feature vector extremely indicates surprise. :

" انا متفاجئ انو الوقت رح يمر بشكل جيد "

which mean *"I am surprised that the time will pass well* ". (See Figure 5, and compare with Figure 3)

	Sad	Joy	Surprise	Disgust	Fear	Anger
F. vector	0.6516933	0.3415608	2.526565	0.8682545	0.8980272	1.642781

Figure 5: Feature vector for statement containing a smiley face

6. EXPERIMENTS

We have conducted multiple tests in order to compare between a number of Arabic tweets emotion classifying methods.

6.1. "Tweet words" vs. "6 emotion" feature vector

Initially, we represented the dataset as feature vector taking the tweet's words as attributes of these vectors, and we applied the classifiers (SVM "SMO", Naïve Bayes, Conditional random fields) using this feature vector. Figure 6 presents the results.

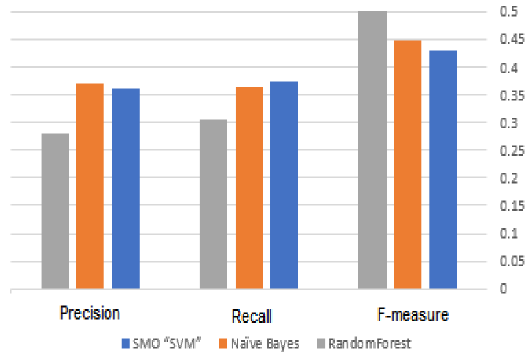

Figure 6: Comparison of different classification results with "Tweet words" feature vector experiment

6.2. (S1, S2, S3, S4) Experiments

In order to explore the different feature vector implementation, we have implemented different models, using S1, S2, S3, S4 sets (see figure 7).

- Weighted words Model (S1):Using TF-IDF weighted model, weighted TWF model, or Modified TF-IDF weighted model (see Section 5-1-e).
- Negation model (S2): Negation_Feature or Negation_Swap.

  For that, when a negation tool is detected, we divide the sentence into two parts, a previous part until the Negation tool and the next part. We have conducted a comparison between these two methods:

  1- Negation_Feature: Add special attribute inside the feature vector that represents the if the sentence includes a Negation tool or not.

  2- Negation_Swap: Deal with the Negation tool impact on the emotional significance of the next word in the sentence.

  In the second method, when a sentence "tweet" includes one of Negation tools, the emotion of the next word will be inverted
- lexicons usage choices (S3) :Using COR Emotional lexicons , Saif lexicons [9], in addition to both COR+NRC Emotion lexicons.
- N-gram choices (S4) :Using 1-grams or 2-grams.
- Stemming choices: Dealing with Full Form Words or Stemmed Form (using"ISRI stemmer").

Journal of Theoretical and Applied Information Technology
15th October 2017. Vol.95. No 19
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

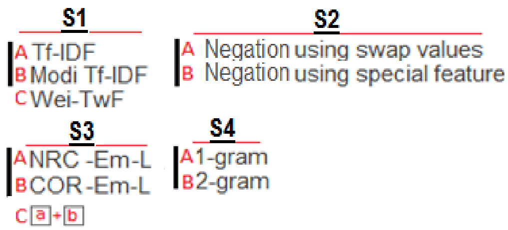

Figure 2: Methods sets

After that, we compared these models with the different feature vector implementations, using CRF classifier. Figure 8 presents the results and shows that the models ABCA, ABCB, ABBA, ABBB had the best results with an F-Measure of 0.98%.

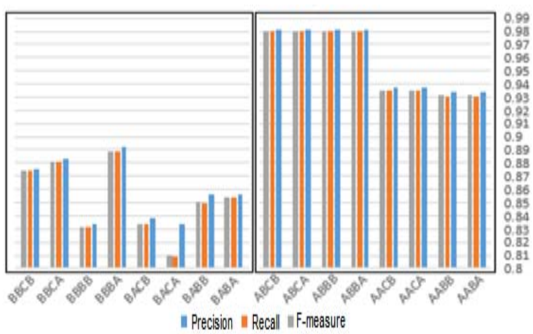

Figure 3: The comparison chart of the models using CRF

S1	S2	S3	S4	Average P	R	F
A	A	C	A	0.934	0.931	0.932
A	A	C	B	0.934	0.931	0.932
A	A	B	A	0.937	0.935	0.935
A	A	B	B	0.937	0.935	0.935
A	B	C	A	0.981	0.98	0.98
A	B	C	B	0.981	0.98	0.98
A	B	B	A	0.981	0.98	0.98
A	B	B	B	0.981	0.98	0.98
B	A	C	A	0.856	0.854	0.854
B	A	C	B	0.856	0.85	0.851
B	A	B	A	0.834	0.809	0.81
B	A	B	B	0.838	0.834	0.834
B	B	C	A	0.892	0.889	0.889
B	B	C	B	0.834	0.832	0.832
B	B	B	A	0.883	0.881	0.881
B	B	B	B	0.875	0.874	0.874
			Average	0.908	0.904	0.904

Table 1: Comparison table of the models using CRF

We can notice, when representing the dataset using words as feature vector attributes the classifier results were less by 60% to 70% than when we improved the feature vector to be represented by the six emotions categories.

6.2.1. S1 Model

In this experiment, we study the effect of the weighting models used *Tf-IDF*, *Modified TF-IDF*, *Weighted-TwF*.

Table 2 shows that using "*Tf-IDF*" enhanced the results of classifiers by 10% to 11% more than when using "*Modiefied TF-IDF*", and by 2% more than when using "*Weighted-TwF*". Figure 9 presents the results.

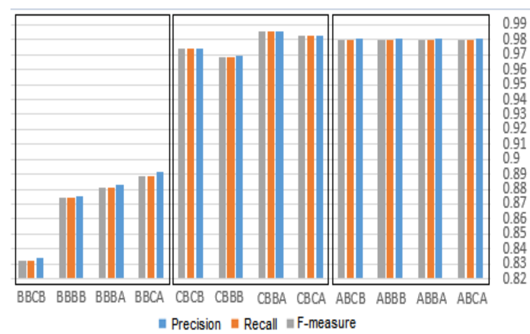

Figure 9: Comparison of the weighting model using CRF

S1	S2	S3	S4	Average P	R	F
A	B	C	A	0.981	0.98	0.98
A	B	B	A	0.981	0.98	0.98
A	B	B	B	0.981	0.98	0.98
A	B	C	B	0.981	0.98	0.98
			Average	0.981	0.98	0.98
C	B	C	A	0.983	0.983	0.983
C	B	B	A	0.986	0.986	0.986
C	B	B	B	0.969	0.968	0.968
C	B	C	B	0.974	0.974	0.974
			Average	0.978	0.977	0.977
B	B	C	A	0.892	0.889	0.889
B	B	B	A	0.883	0.881	0.881
B	B	B	B	0.875	0.874	0.874
B	B	C	B	0.834	0.832	0.832
			Average	0.871	0.869	0.869

Table 2: The Comparison table of the weighting model using CRF

Figure 10 shows the result of using the three classifiers (NB, CRF, SVM(SMO)) with the best tow models in this experiment (CBBA, CBCA).

© 2005 – ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

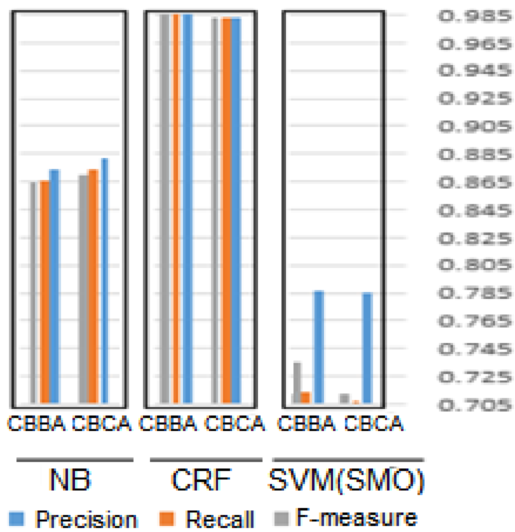

Figure 10: Comparison chart of multiple classifiers with the best tow results

S1	S2	S3	S4		Average P	R	F
C	B	C	A	SVM (SMO)	0.785	0.707	0.712
C	B	B	A	SVM (SMO)	0.787	0.714	0.735
				Average	0.786	0.710	0.723
C	B	C	A	CRF	0.983	0.983	0.983
C	B	B	A	CRF	0.986	0.986	0.986
				Average	0.984	0. 984	0. 984
C	B	C	A	NB	0.882	0.874	0.87
C	B	B	A	NB	0.874	0.866	0.864
				Average	0.878	0.87	0.867

Table 3: Results table of multiple classifiers with the best tow results

6.2.2. S2 Model

In this experiment, we study the effect of using two models of considering "Negation" in the system. Table 4 shows that "Negation_Feature" method enhanced the accuracy by 4% more than when using "Negation_Swap" method. Figure 11 presents the compared results.

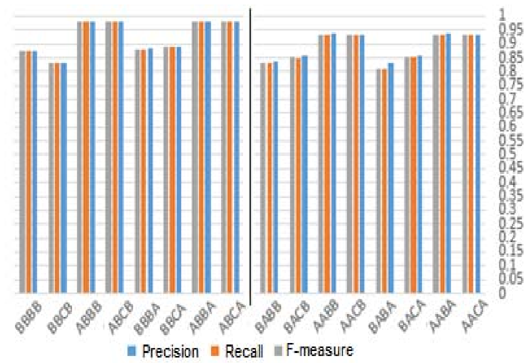

Figure 11: Comparison chart of Negation model using CRF

S1	S2	S3	S4		Average P	R	F
A	**A**	C	A	CRF	0.934	0.931	0.932
A	**A**	B	A	CRF	0.937	0.935	0.935
B	**A**	C	A	CRF	0.856	0.854	0.854
B	**A**	B	A	CRF	0.834	0.809	0.81
A	**A**	C	B	CRF	0.934	0.931	0.932
A	**A**	B	B	CRF	0.937	0.935	0.935
B	**A**	C	B	CRF	0.856	0.85	0.851
B	**A**	B	B	CRF	0.838	0.834	0.834
				Average	0.890	0.884	0.885
A	**B**	C	A	CRF	0.981	0.98	0.98
A	**B**	B	A	CRF	0.981	0.98	0.98
B	**B**	C	A	CRF	0.892	0.889	0.889
B	**B**	B	A	CRF	0.883	0.881	0.881
A	**B**	C	B	CRF	0.981	0.98	0.98
A	**B**	B	B	CRF	0.981	0.98	0.98
B	**B**	C	B	CRF	0.834	0.832	0.832
B	**B**	B	B	CRF	0.875	0.874	0.874
				Average	0.926	0.924	0.924

Table 4: Comparison table of Negation model using CRF

6.2.3. S3 Model

In this experiment, we study the effect of using NRC Emotion Lexicon in addition the lexicon built form the dataset (COR). Table 5 shows that extracting feature vector using simultaniously COR + NRC lexicons enhanced the results by 0.2% more than when using COR Lexicon. (see Figure 12) .

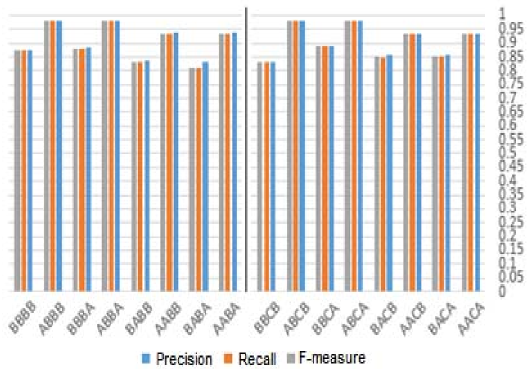

Figure 12: Comparison chart of Lexicon cases model using CRF

S1	S2	S3	S4		Average P	R	F
A	A	**C**	A	CRF	0.934	0.931	0.932
B	A	**C**	A	CRF	0.856	0.854	0.854
A	A	**C**	B	CRF	0.934	0.931	0.932
B	A	**C**	B	CRF	0.856	0.85	0.851
A	B	**C**	A	CRF	0.981	0.98	0.98
B	B	**C**	A	CRF	0.892	0.889	0.889
A	B	**C**	B	CRF	0.981	0.98	0.98
B	B	**C**	B	CRF	0.834	0.832	0.832
				Average	0.908	0.905	0.906
A	A	**B**	A	CRF	0.937	0.935	0.935
B	A	**B**	A	CRF	0.834	0.809	0.81
A	A	**B**	B	CRF	0.937	0.935	0.935
B	A	**B**	B	CRF	0.838	0.834	0.834
A	B	**B**	A	CRF	0.981	0.98	0.98
B	B	**B**	A	CRF	0.883	0.881	0.881
A	B	**B**	B	CRF	0.981	0.98	0.98
B	B	**B**	B	CRF	0.875	0.874	0.874
				Average	0.908	0.903	0.903

Table 5: Comparison table of Lexicon cases model using CRF

6.2.4. S4 Model

In this experiment, we study the effect of using more word context by using 2-grams instead of 1-gram. Table 6 shows that extracting the feature vector using "2-gram" enhanced the results by 0.5% more than when using "1-gram". Figure 13 presents the comparsion results.

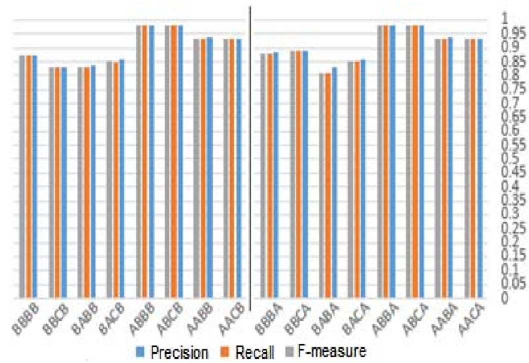

Figure 13: Comparison chart of N-gram model using CRF

S1	S2	S3	S4		Average P	R	F
A	A	C	**A**	CRF	0.934	0.931	0.932
A	A	B	**A**	CRF	0.937	0.935	0.935
A	B	C	**A**	CRF	0.981	0.98	0.98
A	B	B	**A**	CRF	0.981	0.98	0.98
B	A	C	**A**	CRF	0.856	0.854	0.854
B	A	B	**A**	CRF	0.834	0.809	0.81
B	B	C	**A**	CRF	0.892	0.889	0.889
B	B	B	**A**	CRF	0.883	0.881	0.881
				Average	0.912	0.907	0.907
A	A	C	**B**	CRF	0.934	0.931	0.932
A	A	B	**B**	CRF	0.937	0.935	0.935
A	B	C	**B**	CRF	0.981	0.98	0.98
A	B	B	**B**	CRF	0.981	0.98	0.98
B	A	C	**B**	CRF	0.856	0.85	0.851
B	A	B	**B**	CRF	0.838	0.834	0.834
B	B	C	**B**	CRF	0.834	0.832	0.832
B	B	B	**B**	CRF	0.875	0.874	0.874
				Average	0.904	0.902	0.902

*Table 6: **Comparison table of N-gram model using CRF***

6.2.5. "Punctuation and cursing words" Model

In this experiment, we study the effect of using the punctuations (exclamation, and question marks) in addition to "cursing" words in the system. Extracting feature vector taking into consideration these atributes enhanced the accuracy by 0.3% (see Figure 14 vs Figure 10).

© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

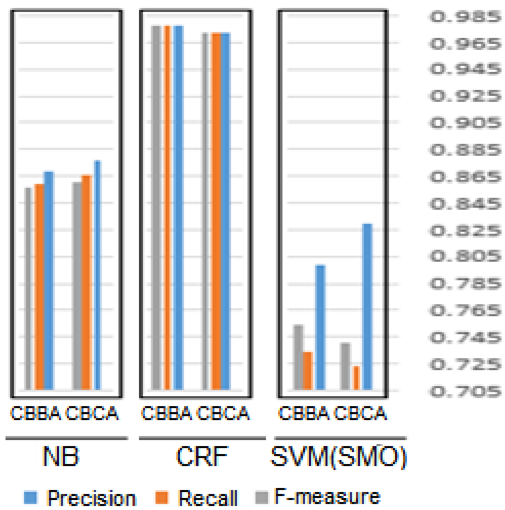

Figure 4: Comparison of multiple classifiers considering Punctuations and Cursing words

6.2.6. Word Form

In this experiment, we study the effect of the word form used in the system (Full-form vs. stemmed words). Table 7 shows that extracting feature vector using "Full-Form" words enhanced the classification results by 0.8% to 1.5% more than ISRI Stemmed words. Figure 15 presents the comparsion results.

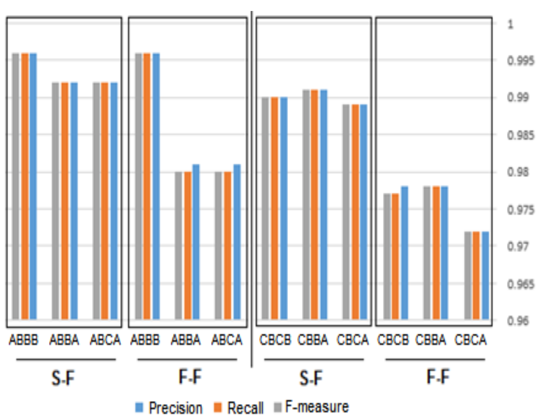

Figure 5: Comparison chart of Word Form model using CRF

					Average		
S1	S2	S3	S4		P	R	F
C	B	C	A	SF	0.972	0.972	0.972
C	B	B	A	SF	0.978	0.978	0.978
C	B	C	B	SF	0.978	0.977	0.977
Average					0.976	0.975	0.975
C	B	C	A	FF	0.989	0.989	0.989
C	B	B	A	FF	0.991	0.991	0.991
C	B	C	B	FF	0.99	0.99	0.99
Average					0.99	0.99	0.99
A	B	C	A	SF	0.981	0.98	0.98
A	B	B	A	SF	0.981	0.98	0.98
A	B	B	B	SF	0.996	0.996	0.996
Average					0.986	0.985	0.985
A	B	C	A	FF	0.992	0.992	0.992
A	B	B	A	FF	0.992	0.992	0.992
A	B	B	B	FF	0.996	0.996	0.996
Average					0.993	0.993	0.993

Table 7: Comparison table of Word Form model using CRF

6.2.7. Stop Words

After taking into consideration each of "exclamation, question marks and cursing words" as attribute in the feature vector, we compared the results of keeping versus removing stop words when extracting feature vector. Figure 16 presents the results and shows that keeping the stop words has a positive effect on the emotion detection. The raison could be that many stop words can affect the emotion detection and eliminating them could have an impact on the Recall measure. For example, the stop word "ما" serves as an indicator of a surprise, such in "ما أصغر هذا العالم!" which mean "what a small world!".

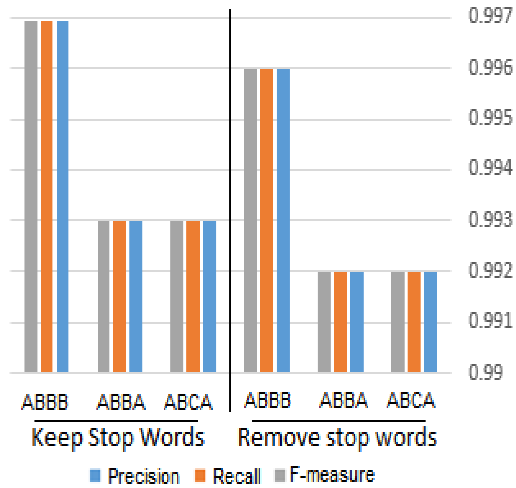

Figure 6 : Comparison chart of Stop Words model using CRF

S1	S2	S3	S4		Average P	Average R	Average F
A	B	C	A	Remove	0.992	0.992	0.992
A	B	B	A	Remove	0.992	0.992	0.992
A	B	B	B	Remove	0.996	0.996	0.996
				Average	0.993	0.993	0.993
A	B	C	A	Keep	0.993	0.993	0.993
A	B	B	A	Keep	0.993	0.993	0.993
A	B	B	B	Keep	0.997	0.997	0.997
				Average	0.994	0.994	0.994

Table 8: Comparison Table of Stop Words model using CRF

6.2.8. Testing on a new Dataset

To experiment the effect of the data set on the results, we have applied the best model we had on a new dataset espacially built to test the system. The tweets in the new dataset is quite different from the tweets used in the previous experiments. The new dataset contained 390 tweets , with 65 tweets for each emotion. Table 9 presents the results on the new dataset, with about 66.9% F-measure using CRF classifier. We remark a decrease in the results, which is quite logic due to the changement of the words in the new dataset.

Precision	Recall	F-measure	Emotions Category
0.661	0.804	0.726	*Sadness*
0.679	0.809	0.738	*Joy*
0.828	0.462	0.593	*Surprise*
0.695	0.837	0.759	*Disgust*
0.484	0.596	0.534	*Fear*
0.862	0.521	0.649	*Anger*
0.7	0.669	0.664	←Avg

Table 9: CRF's results on the New Dataset.

7. CONCLUSION

In this research, we proposed different models to recognize the basic six emotions - Sad, Joy, Surprise, Disgust, Fear, Anger - of Arabic tweets (Syrian dialectal tweets). In order to test our models, we have built a balanced dataset of Syrian Tweets, and manually annotated it. We compared the results of several machine learning algorithms such as SVM, Naive Bayes, CRF. We also compared the results of our proposed models.

In the future, we intend to expand the size of our labeled dataset, by acquiring more emotional Arabic Syrian tweets. We will validate this dataset by annotating it by 3 annotators, which will enable us to calculate the interagreement of the annotators. We also intend to expand our special lexicons (cursing words used in dialectal case, emotional dialectal words and idioms….) using automatic methods.

REFERENCES

[1] Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Transactions on Information Systems (TOIS), 26(3), 12.

[2] Al-Shalabi, R., & Obeidat, R. (2008, March). Improving KNN Arabic text classification with n-grams based document indexing. In Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt (pp. 108-112).

[3] Alhajjar, A., Hajjar, M., & Zreik, K. (2009, April). Classification of arabic information extraction methods. In 2nd International conference on Arabic language resources and tools.

[4] Danisman, T., & Alpkocak, A. (2008, April). Feeler: Emotion classification of text using vector space model. In AISB 2008 Convention Communication, Interaction and Social Intelligence (Vol. 1, p. 53).

[5] Do, H. J., & Choi, H. J. (2015, October). Korean Twitter Emotion Classification Using Automatically Built Emotion Lexicons and Fine-Grained Features. In PACLIC.

[6] Davidov, D., Tsur, O., & Rappoport, A. (2010, August). Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd international conference on computational linguistics:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

posters (pp. 241-249). Association for Computational Linguistics.

[7] Barbosa, L., & Feng, J. (2010, August). Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 36-44). Association for Computational Linguistics.

[8] Elhawary, M., & Elfeky, M. (2010, December). Mining Arabic business reviews. In Data Mining Workshops (ICDMW), 2010 IEEE International Conference on (pp. 1108-1113). IEEE.

[9] . NRC Hashtag Emotion Lexicon - Version 0.2 - 2013. Saif Mohammad (saif.mohammad@nrc-cnrc.gc.ca). www.*saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm*

[10] Rabie, O., & Sturm, C. (2014). Feel the heat: Emotion detection in Arabic social media content. In the International Conference on Data Mining, Internet Computing, and Big Data (BigData2014) (pp. 37-49). The Society of Digital Information and Wireless Com

[11] Twitter Data set for Arabic Sentiment Analysis. April, 2014. (a) Creator: N. A. Abdulla and N. Mahyoub (b) Donor: N. A. Abdulla (naabdulla11@cit.just.edu.jo) "collection of 2000 labelled tweets (positive tweets and negative ones) These tweets written in both Modern Standard Arabic (MSA) and the Jordanian dialect".

[12] AraSenti Lexicon. May, 2016 (a) Nour AlTwairesh (Twairesh@ksu.edu.sa) (b) Hend Alkhaifa (hendk@ksu.edu.sa) (c) abdulMalik Alsalman (salman@ksu.edu.sa)

"All resources created and used in Arabic Sentiment Analysis of Arabic Tweets. Includes Sentiment lexicon generated from Arabic tweets and a corpus of Arabic tweets in the Saudi dialect annotated with four labels: positive, negative, neutral, mixed"