Microblogging Opinion Mining Approach for Kuwaiti Dialect

Jana Ben Salamah and Aymen Elkhlifi
LaLIC, Université Paris-Sorbonne, 28 rue Serpente, 75006, Paris, France
ENS Lyon, ICAR Laboratory, 15 parvis René Descartes - ENS de Lyon, 69342 Lyon, France
jana.bensalamah @paris4.sorbonne.fr
Aymen.khelifi@ens-lyon.f

ABSTRACT

In this paper, we present an approach to extract and classify opinion in microblogging. It is based mainly on linguistic resources produced for the Kuwaiti dialect used by an SVM classifier. The approach has been tested on a corpus of 340,000 tweets about "interrogation of ministers by the National Assembly of Kuwait"-"استجواب الوزراء" during the last two years. Tweets were collected automatically by a module developed in java. This corpus has been manually annotated by three Kuwaiti dialect native speakers. We obtained an average value of precision and recall respectively 76% and 61%.

KEYWORDS

Microblogging opinion mining, information extraction, opinion classification, tweet collector.

1 INTRODUCTION

With the growing popularity of Web 2.0, the opinion mining becomes more and more interesting. New kinds of opinion-rich sources appeared like blogs, social networks, wikis and online reviews sites. Due to the largest size of this available data, we need to automate their extraction and organization.

This requirement has motivated researches to detect opinions in text passages and assign them to subjective classes: positive or negative opinions.

Different techniques have been applied to this purpose such as machine learning classifiers generally based on lexical features [1] or syntactic features [2] associated with opinion. A wide range of statistical methods are also investigated in order to detect and classify subjective texts [3], [4]. But for dialect language we haven't the language resources or taggers to be able to use the methods mentioned above.

In this paper, we are interested in this issue, particularly on the Kuwaiti dialect from twitter. We examine an alternative extraction and classification strategy of opinions based on proposed linguistic resources.

Accordingly, we suggest solving this problem of opinion extracting in stages; starting with collect and preprocessing tweets. In fact, we take as unit opinion carrier a single word, and first classify each adjective, verb, and noun by its opinion. Then, we combine the elementary opinion information in order to find the overall opinion class.

The rest of the document is organized as follows: section (2) introduces the related works on opinion extraction in general then in Arabic language. In section (3), our approach to detect and the classify opinions is described in detail. The experimentation is described in section (4). In section (5) we evaluate the KDOEST (*Kuwaiti-Dialect Opinion Extraction System from Twitter*) in order to demonstrate its ability. In section (6), we conclude with a few notes on future works.

2 RELATED WORKS

Research in opinion mining includes distinguishing subjective from objective language [1], [2], [5], [6], [7] as well as distinguishing positive from negative language [2], [8], [9], [10].

Various approaches have been adopted to address the first question. For example, Yu and Hatzivassiloglou [8] use unsupervised statistical techniques for detecting opinions at the sentence level. While Bethard, Thornton, Hatzivassiloglou and Jurafsky [1] use a statistical approach divided on two methods for proposition opinion classification. The first method relies on differences in the relative frequency of a word in subjective documents, versus documents that contain mostly facts using the TREC 8, 9 and 11 text collections. The second method used co-occurrence information to extend a seed list of 1336 opinion adjectives. Wilson, Wiebe and Hwa [2] propose an automatic opinion classification approach to classify nested clauses in every sentence in the corpus. They use a wide range of features, including new syntactic features.

Although these methods achieve high precision, they are based on a large corpus, and need a large amount of manually tagged training data.

In our study we investigate the Kuwaiti dialect, we were asked to build these resources to apply similar methods.

For word sentiment classification, atzivassiloglou and McKeown [9] use a supervised learning algorithm to infer the semantic orientation of adjectives from conjunction constraints. While Turney [4] applies a specific unsupervised learning algorithm based on the mutual information between sentences and the words "excellent" and "poor", where the mutual information is calculated using statistical techniques.

In earlier work [4] only singletons ("excellent" and "poor") were used as seed words. In our work, we use many seed words having more than one subjective category: harm, approval, joy, criticism, pleasure, etc., in order to test whether multiple seed words have a positive effect in extraction performance.

However, symbolic rules are developed manually, based on annotated data so well. Moreover, such approaches are language-dependents.

In contrast to statistical approaches, symbolic approaches such as proposed by Maurel and Dini [10] and Vernier, Monceaux and Daille [7] provide a better text analysis to represent the grammatical and semantic structure of analyzed text.

Concerning the Arabic language, most works based on lexical based classification using the machine algorithm [11], [12], [13], [14], but they have resources about Standard Arabic. For example in [12], [13] they use a tagger to prepare dataset. In the same way, Yet in [11] used Token, Lemma, Word forms, POS Tagging standard features (Unique Gender, User ID) with a polarity lexicon for subjectivity and sentiment analysis for Arabic social media. On the other hand, El-halees [15] works on document level using the combined feature by three methods. The first method, establish on lexicon based method to classify documents. And then using as training set for maximum entropy method which subsequently classifies some other documents. Finally, k-nearest method used the classified documents from these two last methods to classify the rest of the documents.

As part of our studies, we have no a tagger on the Kuwaiti dialect. We also don't have dictionaries for this dialect. That is why we decided to create our own linguistic resources to organize a lexicon around opinions.

3 THE PROPOSED APPROACH OF OPINIONS EXTRACTION AND EMOTION CLASSIFICATION

Microblogging today has become very popular communication tool. Millions of messages are posted in websites providing microblogging features such as Twitter, Facebook, Pinnit, etc. We are particularly interested in twitter.

We propose to extract the opinions in four steps. (See Figure 1):

1. Tweets Collector: It involves collecting tweet using keywords (with or without #). The user specifies a list of keyword with the number of tweets to collect and a Time period. This period circle the start and end publication date of tweets.
2. Preprocessing: it consists on the one hand, in the tokenization and the segmentation of tweets, on the other hand, in the extraction of htag, user mention, url form tweet.
3. Opinions-oriented words extraction: is to extract the opinion-oriented words through language resources that we have developed for the Kuwaiti dialect.
4. Opinion classification: it consists in classifying the polarity of opinion using an SVM classifier. It takes as an attribute; inter alia, term frequencies extracted in the previous step.

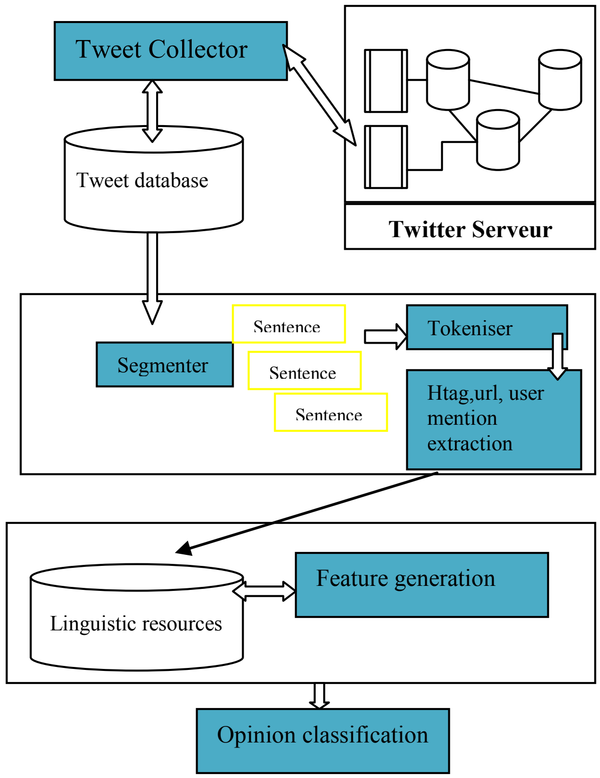

Figure 1. Proposed approach of opinions extraction and classification

3.1 Tweets Collector:

It is used to collect a large number of tweets in a reasonable time. An average of 11,000 tweets per hour for 3 computers connected at the same time. The restriction concerning the number of tweets automatically downloaded is imposed by twitter. For example, twitter cut the connection from server every quarter of an hour as a security measure. For this, we have created a user account table, in which we created and set 30 accounts. The module 1 alternates between these accounts in the process of collection.

Another constraint imposed by twitter: only 100 tweets per automatic collection. We used a specific source code that overcomes this restriction: starting by a thread connection in a loop and a point to an exception for container collection.

Our component of Tweet Collector consists of a java-client module distributed on several computers and a centralized relational database. The database model is the structure of twitter network.

3.2 Preprocessing

In our study, preprocessing comprise the segmentation of tweet into sentences then the tokenization of its words.

The segmentation: is the determination of the sentences borders. It is a hardly-realizable task. Given that punctuation is not enough to detect the end or the beginning of a segment it is necessary to take into account all typographical markers and specific linguistic marker like "و". Moreover, other linguistic bases are engaged like the syntactic structure of a sentence and the significance of each typographical marker in a well-defined context.

We developed our own segmenter while basing ourselves on punctuation marks. Due to the great number of the linguistic rules to program, we have to integrate in our knowledge base all the rules developed in the system Segatex [16].

The sentences obtained, we apply the Stanford Arabic tokenzier. It is a class form Stanford POS parser.

For tweets that contain the search keywords, we store them in the database, username transmitter, date, tweet text, retweet number, favorite number etc.

Given the fallowing tweet: @hesham @salwa_z. ماكو منها استدعاء #فاتن_حمزة. مو صج الاستجواب . عجيب القانون فقد هيبته

We extract the flowing information:

- TweetID : 440480688630800000
- User emitter : ***bahraini_soul***, ID User : 44726
- Date : ***Mon Mar 03 14:35:54*** CET 2014
- Htag : ***#فاتن_حمزة***
- Number of retweet : ***2***
- Langue : ***ar***
- Url : 0;
- Number of favorites : ***0***

For our experiments we collected 340,000 tweets over the keywords "استجواب الوزير", "الكويت + الاستجواب", "الاستجواب + مجلس الامه الكويتي".

3.3 Opinions-Oriented Words Extraction

This module is used to extract the opinion-oriented words through language resources that we have developed for the Kuwaiti dialect. We describe them in the flowing section.

Linguistic Resources for the Kuwaiti Dialect:

We have implemented a hierarchical set of resources for the Kuwaiti dialect. They cover, inter alia, opinions and emotions indicators (Linguistic unity that marks). We divided the linguistic indicators opinions into four sets: adjectival, nominal, verbal and adverbial indicators. For each set, we defined a list of subclasses of indicator. In this article, we describe only opinion resources;

We identified and classified 4213 adjectives written in Kuwaiti dialect. Each item has been manually assigned to one of 22 classes: 12 positive and 10 negative as follows:

Positive classes (adjectives):

1. attraction - الجاذبيه
2. love - الحب
3. joy and happiness - الفرح و السعادة
4. Pleasure - المتعة
5. goodness - الصلاح
6. good judgment - الحكم الجيد
7. satisfaction - الرضا
8. generosity - الكرم
9. emotions - العواطف
10. elegance - الأناقة
11. complement - المدح
12. others - آخرون

Negative classes (adjectives):

1. criticism - النقد
2. badness - الوضع السيء
3. annoyance - الإزعاج
4. harm & injury - الأذى و الضرر
5. terror - الرعب
6. sadness - الحزن
7. moral - الضرر المعنوي
8. anger - الغضب
9. unpleasantness - البغض
10. others - آخرون

For example, concerning the first adjectival class "***attraction"***, we identified 13 subclasses with a list of adjectives for each subclass as shown in the following table:

Table 1. Adjectival subclasses of "attraction" in Kuwaiti dialect

<table>
<tr><th>subclass of attraction</th><th>Kuwaiti dialect adjectives</th></tr>
<tr><td>1. spellbinding</td><td>ملك جمال | صوره | آيه | ياخذ العقل | ياخذ القلب | روعه | قطعه</td></tr>
<tr><td>2. seductive</td><td>إشد الواحد | مغري | مغوي</td></tr>
<tr><td>3. fascinating</td><td>زوقه | زوغه | اينن | اعور القلب | روعه</td></tr>
<tr><td>4. beautiful</td><td>حلو | مملوح | اشحلاته | يحليله | سنع | شاتر | ناطع</td></tr>
<tr><td>5. attractive</td><td>يقتل | عذاب | نالر</td></tr>
<tr><td>6. gorgeous</td><td>كامل و الكامل الله | اهبل</td></tr>
<tr><td>7. Bonny</td><td>حياوي | نشيط | صحته ممتازه</td></tr>
<tr><td>8. splendid</td><td>ويه نور | ويهه امنور | امنور | منجز | امبدع | فنان | عايش ابنعمه</td></tr>
<tr><td>9. Nice</td><td>حبوب | طيوب | كتكوت | كيوت | امرتب | ناعم</td></tr>
<tr><td>10. shapely</td><td>شكله حلو | منسق | جسمه خيال | جسسم | رشيق</td></tr>
<tr><td>11. Pretty</td><td>ماشي حاله | خفيف طينه | دمه خفيف | زين</td></tr>
<tr><td>12. tempting</td><td>اشوق | سكسي | دميره</td></tr>
</table>

We added to the list of adjectives above, a second of adjective in standard Arabic. They are also used in Kuwaiti dialect. As shown in the following table:

Table 2. Adjectival subclasses of "attraction" in Arabic standard language

<table>
<tr><th>subclass of attraction</th><th>Standard Arabic adjectives</th></tr>
<tr><td>spellbinding</td><td>فتان | فاتن | ملفت للنظر | مفتون بجماله | تصويري</td></tr>
<tr><td>seductive</td><td>مغر | مغو | جذاب</td></tr>
<tr><td>fascinating</td><td>ساحر | آسر</td></tr>
<tr><td>beautiful</td><td>جميل | ذات جمال | ملك جمال | ذات حسن | مليح | حسن | حسناء | وسيم</td></tr>
<tr><td>attractive</td><td>جذاب | خلاب</td></tr>
<tr><td>gorgeous</td><td>كامل الاوصاف | فائق الجمال | بهي | رائع</td></tr>
<tr><td>Bonny</td><td>ممتاز| ممتلئ صحة | حيوي | حي</td></tr>
<tr><td>splendid</td><td>باهر | بديع | مبدع | بارع | مشرق | فاخر | جليل | لامع | فخم | مترف | ساطع</td></tr>
<tr><td>Nice</td><td>لطيف | ظريف | أنيق | حلو | مليح | عذب | رقيق</td></tr>
<tr><td>Shapely</td><td>متناسق | حسن الشكل | حسن | متألق | جميل القوام | رشيق القوام</td></tr>
<tr><td>Pretty</td><td>حسن | ظريف | جيد</td></tr>
<tr><td>tempting</td><td>مغر | مغري | مغو</td></tr>
</table>

Regarding the verbal indicators, we identified 23 classes, 11 positive and 12 negative. The following is an excerpt lists defined for each positive class.

Table 3. Verb subclasses indicators (positive)

Verbal classes	Verb indicator (positive)
1. attraction – الجاذبية	charm - سحر, mesmerize - فتن, etc.
2. love - الحب	admire - عجب, adore - عشق, love - حبب, favor أصلح ,prefer - فضل
3. desire - الرغبة	Hanker – تاق, desire - رغب, like - ود, want - أراد, wish - تمنى, hope - أمل
4. approval & praise - الموافقة و الثناء	Applaud - أشاد ,praise - مدح , congratulate – هنأ, compliment – إطراء, recommend - أوصي, exalt - كثف, laud - مجد
5. agreement - القبول	Consent - وافق , admit إعترف
6. pleasure & happiness - المتعة و السعادة	Revel - عربد, gratify - أشبع ,delight - بهج
7. joy - الفرح	Exult - تهلل ,triumph - انتصر , joy - فرح, jubilate - تهلل, rejoice - ابتهج
9. sympathy- التعاطف	sympathize - تعاطف, assort – تجانس, resemble – تماثل, console - واسى
10. goodness - الصلاح	ameliorate - تحسن, beautify - جمل consummate – بارع
11. other positives - أخرى	astonish – ذهل, endear - مات, dote - حب الجنون ,enthuse - تحمس, brave - شجع, exonerate - أبرئ ,forgive - غفر , persuade - حث , edify - هذب

The following is an excerpt lists defined for each negative class.

Table 4. Verb subclasses indicators (negative)

Nominal classes	Verb indicator
1. criticism – النقد	Belabor - هاجم وكرر, censure - لام , chastise – عاقب, criticize – نقد, scold - أنب, fulminate - شجب , rebuke - وبخ
2. disapproval - الاستنكار	Belittle - استخف, disparage - انتقص etc.
3. opposition – المعارضة	deny - نفى ,contravene - خالف , disapprove - رفض , discourage – تثبيط, oppose - عارض ,disbelieve – كفر, refuse – رفض
4. dislike – الكراهية	Dislike - كره, detest - بغض
5. annoyance – الإزعاج	Annoy - ضايق , beleaguer , حاصر ,vex - غيظ bother - أزعج
6. displeasure – الاستياء	complain – إشتكى, grouch - تذمر, grouse - احتج, grumble - تذمر
7. pain - الألم	Anguish – كرب, terrify – جنن, embarrass – أربك, abash – خجلpain – ألم, hurt – جرح, ache – وجع, inflict - أصاب
8. anxiety – القلق	pain – ألم, hurt – جرح, ache – وجع, inflict - أصاب
9. sadness – الحزن	Fear - أخاف, worry - أقلق
10. misbehavior - سوء السلوك	Bemoan - تحسر على, bewail - نوح, mourn - ندب, lament - مرثاة, rue - أسف, regret - ندم
11.harm - الضرر	Slander - افترى , malign - شتم, vilify - ذم,falsify - دحض, discredit - شوه السمعة, blandish - تملق,corrupt – أفسد, debase – أهان, deceive - خدع, delude - ضلل , misinform - تضليل
12.anger – الغضب	Endanger - وقع في الخطر, jeopardize - عرض للخطر, harm - ضر, prejudice - تحيز, aggravate - تفاقم , exacerbate - إستفحل
13.other negatives – أخرى	Anger – غضب, enrage - حنق

With regard to the names, we sum both class has satisfied: positive and negative. The following is an extract positive noun:

Table 5. Nouns indicators (positive)

Nominal classes	Noun indicator
Nouns	الهتاف , الاحتضان ,الفطنة, الإعجاب , التأييد , الذهول, الود, المودة, العاطفة, التأكيد, الجاذبية الحب, التسلية, الرسول, التأليه, التقدير, الاعتقال, الحماسة, الإثارة, الصاعد, الطموح, التوكيد, الدهشة, الجاذبية, الرعب, الصفقة, الجمال, المستفيد, الخير, النعيم , الجرأة, الشجاعة, الصراحة, السبي, الاحتفال, السحر, الهتاف, الكياسة, القرب, التماسك, الشفقة, التكملة, الثقة, التهنئة, الإجماع, الثبات, الاطمئنان, الشجاعة, المجاملة, الحنين, التصديق, المحبوب, المدافع, الإذعان, البهجة, الرغبة, التقرير, الإخلاص, الكرامة, التوق, الجدية, الحماسة , الوقاحة, الراقة, الفيض, العجب , الكهربة, الأناقة, البلاغة, الربوة, التشجيع, المتعة, الحماسة, الاحترام, خلود, النشوة , التمجيد, الامتياز, الإثارة, النموذج, الابتهاج, الوفرة, الشهرة , الإخلاص, الإيمان, التهلل etc.

ISBN: 978-0-9891305-5-4 ©2014 SDIWC

Following a negative extracted noun:

Table 6. Nouns indicators (negative)

Nominal classes	Noun indicator
Nouns	الانحراف, الاشمئزاز, المقت, المقت شديد, السخافة, السخف, الهاوية, الاتهام, الحدة , اللذع, العتاب, الغش, الإهانة, التفاقم , العدوان, العدوانية, المعتد, الإثارة, المهيج, سكرة الموت, الادعاء, الغموض, التناقض, الفوضوية, الفوضوي , الفوضى, الغضب, الكرب, الإبادة, العداء, الإزعاج, التضاد, المضادة لل, الكراهية, القلق, الحرص, اللا مبالاة, سفر الرؤيا, المدافع عن دين أو قضية, الغطرسة التشهير, القاتل, الفظاعة, الجرأة , الوقاحة, المستبد, الجشع, النفور, الفظاعة, الحرج, التخلف, الحيرة, الهلاك , النفي, الهمجي, الهمجية, القاحل, ساحة المعركة, خط المعركة, البهيمة, الهرج والمرج, المخبول, العداء, الثكل, الخيانة, الضلالي, الحيرة, الانحياز, المرارة, التجديف, الكآبة, اللوم, الأبله, سفك الدماء, التخبط, القصف, العبودية, الملل, الإزعاج, الجرأة , الشقاوه , التبجح, الوقاحه , الكبريت, الوحشية, الوحشي, العامل غير بارع , العبء, الفضولي etc.

We have defined a list of adverb and quantifier classified according to their polarities. Following an intensifier extracted adverb:

Table 7. Quantifier subclasses indicators (postive)

quantifier classes	adverb indicator
quantifiers	حييل, حاد, مرعب, فظيع, مهول, مخيف, وايد , خيالي, متشدد, مو معقول, حييل بالقوه, ابقساوه, بحدة , منحط, بطريقه مو طبيعيه, خطير, مو متأثر, ميؤوس منه, حده مؤثر, ما تقدر تتجنبه, اعلي شي, قمه, نشيط , ابدقة, دايما, ابراعة, خصوصا, علي طول, معقد, كله / بكل معنى, خاصة, مهم / ضروري, كلش, استثنائي, باستثناء, سريع / افشقه, إلى الأبد, فل / كامل, اساسا / في الاساس, بشراسة, يستغنى عنه, بدون تفكير, بشكل مكثف, بشكل معقد, بدون حدود etc.

Following a diminished extracted adverb:

Table 8. Quantifier subclasses indicators (negative)

Quantifier classes	adverb indicator
quantifier	ما عنده هدف, مو متحمل, بانتباه / ابحذر, متقلب / متغير, كافي, اشويه, معتدل, لحظات, كلش, مو محتمل, بس, مو مستقر, نادر, مو ثقه, مناسب / معقول, علبونه, تقريبا, مستبد / استبدادي, بالياله, حوالي, من وقت لي وقت, جزئيا, يمكن, تسبيا, اقريب,اباي طريقه تكون, ظاهريا, اونس, ضعيف, مرات etc.

This work of identification and classification of linguistic units was released by 3 for Kuwaiti dialect native speakers. We obtained a final database of 16784 linguistic units classified. They are stored in a relational database. We made available community resources.

3.4 Opinion Classification:

We are preparing a training set for applying some machine learning algorithms. Features calculate that includes the following variables:

- Number of terms belonging to the list of ***positive*** nouns indicator.
- Number of terms belonging to the list of ***negative*** nouns indicator.
- 22 variables measuring the frequencies of terms in relation to the 22 classes of adjective.
- 23 variables measuring the frequency of terms in relation to the 23 classes of verbs.
- N-grams variables for each term defined lists (noun, verb, adjective), we combine several Boolean variables incorporating n-grams for (n = 2, 3, 4).
- The number of positive emotions terms like ";)", ":)", ":D", "-.-", "-_-"
- The number of positive emotions terms like ":(", ";'(", ":o" etc.
- etc.

Initially, we applied to this training set several learning algorithms based on Decision Tree to understand the rules built. We noticed that the training set is multidimensional, so we applied the SVM technique because it is adequate for this kind of corpus.

4 Experimentation

A system, called *KDOEST (Kuwaiti-Dialect Opinion Extraction System from Twitter)* based on the proposed techniques has been implemented in Java under Eclipse platform.

Kuwaiti-Tweet includes the four following modules:

- Module 1: Tweets Collector,
- Module 2: Preprocessing,
- Module 3: Opinion Oriented terms Extraction,

- Module 4: Opinion classification.

Using module of Tweets Collector (see section 3) we collected a corpus containing 340,000 tweets about "interrogation of ministers by the National Assembly of Kuwait" - "استجواب الوزراء" during the last two years.
All tweets were annotated by three Kuwaiti dialect native speakers. The first task was to indicate whether a tweet is subjective (containing opinions). If this is the case, the second step is to be classified as positive, negative.

The average length of tweet is 12.23 words, with an average of 1.45 opinions per tweet, for a total of approximately 4 158 200 words;

For module 3 we use the *Weka* libraries edition 3.6 [17] for decision tree and SVM algorithms (*LibSVM*).
An *ARFF* file (input format of *Weka*) is generated automatically for each tweet. It will be useful like a data source for the algorithms of classification. We adopted *J48*, *ADTREE*, *Random Tree* and *SVM* with the cases of the events.

5 Results

The evaluation is done at several levels. We start with the classification evaluation by using the PCC, then, the clustering by measuring the Precision and the Recall. We exploit the following algorithms:
To evaluate our approach, we employ the precision and recall measurements. We assign each sentence to one of the three following categories:

- a: sentence correctly assigned to a given class value,
- b: sentence not detected (silent),
- c: sentence misclassified.

The Precision, the Recall and F1 prove that to be calculated as:

$$P = \frac{a}{a+c}, \; R = \frac{a}{a+b} \text{ and } F1 = \frac{2 \times P \times R}{(P+R)}$$

J48: implementation of C4.5 algorithm [18] which selects for each level the tree node as the attribute which differentiate better the data. Then, it divides the training set into sub-groups in order to reflect the values of the attribute of the selected node. We repeat the same treatment for under group until we obtain under homogeneous groups (all the instances or the majority have the same attribute of decision).

ADTree: construction of the decision trees extended to the cases of multiclass and multi-labels.

Random Tree: begin with tree random and chosen by the majority best vote.
We obtained the following results:

J48

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.525	0.158	0.469	0.325	0.383	positive
0.442	0.375	0.427	0.502	0.461	negative

ADTREE

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.411	0.333	0.512	0.476	0.493	positive
0.420	0.242	0.421	0.453	0.463	negative

RandomTree

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.5	0.211	0.601	0.5	0.548	positive
0.489	0.5	0.452	0.489	0.469	negative

SVM

Precision	Recall	F-Measure	Class
0.767	0.634	0.694	positive
0.658	0.789	0.717	negative

6 Conclusion and Future Word

In this article, we have proposed four stages to annotate opinion on tweets starting, in a first stage, by the collecting them using keyword in a specific model. In a second stage, the preprocessing consists on segment and extract some information form tweet. In the third stage, we generate features for machine learning process based on developed linguistic resources to Kuwaiti dialect. Finally, we use SVM and decision Tree to classify tweet in positive or negative classes.

Our approach was evaluated on 340,000 tweets about "interrogation of ministers by the National Assembly of Kuwait" during the last two years. Tweets were collected automatically by a module developed in java. This corpus has been manually annotated by three Kuwaiti dialect native speakers. We obtained an average value of precision and recall respectively 76% and 61%.

In short term, one of the first future works which we propose is to adopt this word on other dialect like Saudi dialect.

In long term, we look forward to fuse the opinion. In effect, we have the idea of adopting, to the case of the events, the MCT model proposed by Smets [19] for the fusion of information in general.

7 References

[1] S.Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou and Jurafsky D, “Automatic extraction of opinion propositions and their holders”. Proc. The Association for the Advancement of Artificial Intelligence. 2004. (AAAI-04).

[2] Wilson T., Wiebe J. and Hwa R.; "Just how mad are you? Finding strong and weak opinion clauses". Proc. AAAI’04. 2004.

[3] Kim S.-M. and Hovy E. (2004). "Determining the sentiment of opinions". Proc. COLING-04, pp. 1267–1373.

[4] Turney P. D. (2002). "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews". Proc. ACL’02.

[5] Riloff E. and Wiebe J. (2003). "Learning extraction patterns for subjective expressions". Proc. The Conference on Empirical Methods in Natural Language Processing (EMNLP-2003), pp. 105–112.

[6] Riloff E., Wiebe J. and Wilson T. (2003). "Learning subjective nouns using extraction pattern bootstrapping". Proc. The Seventh Conference on Natural Language Learning (CoNLL-03).

[7] [Vernier M., Monceaux L. and Daille B.1. (2009). "Catégorisation des évaluations dans un corpus de blogs multi-domaine". Revue des nouvelles technologies de l'information 2009, pp. 25.

[8] Yu H. and Hatzivassiloglou V. (2003). "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences". Proc. EMNLP’03, pp. 129–136.

[9] [Hatzivassiloglou V. and McKeown K. R. (1997). "Predicting the semantic orientation of adjectives". Proc. ACL’97, pp.174–181.

[10] Maurel S. and Dini L. (2009). "Exploration de corpus pour l’analyse de sentiments". Proc. DEfi Fouille de Textes (DEFT-09), p. 11-23.

[11] Rushdi-Saleh M., Teresa Martín-Valdivia M., Alfonso Ureña-López L., and José M. Perea-Ortega (2011) “OCA: Opinion Corpus for Arabic”, Proc. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY—October 2011

[12] Abdul-mageed M. , Sandra K. , Kbler S. and M. Diab (2012) ”SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media to Subjectivity and Sentiment” , no. July 2012, pp. 1928.

[13] Shoukry A. and Rafea A. (2012) ”Sentence-level Arabic sentiment analysis” Proc. 2012 International Conference on Collaboration Technologies and Sys-tems (CTS), May 2012, pp. 546550.

[14] L. Albraheem and H. S. Al-Khalifa, ”Exploring the problems of senti- ment analysis in informal Arabic,” Proceedings of the 14th International Conference on Information Integration and Web-based Applications and Services - IIWAS 12, 2012,p. 415.

[15] A. El-halees, ”Arabic opinion mining using combined classification approach,” Proceedings of the International Arab Conference on Infor- mation Technology, ACIT. 2011

[16] Gh. Mourad, “Analyse informatique des signes typographiques pour la segmentation de textes et l’extraction automatique de citations. Réalisation des Applications informatiques: SegATex et CitaRE“. 2002. Thèse de doctorat Université Paris-Sorbonne.

[17] E. Frank, I. Witten, Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). Morgan Kaufmann, 2005. 525 pages.

[18] Quinlan, J. R. 1993. Programs for Machine Learning. Morgan Kaufmann Publishers.

[19] Ph. Smets, The Transferable Belief Model and other Interpretations of Dempster-Shafer's Model. Uncertainty in Artificial Intelligence 6, P.P. Bonissone, M. Henrion, L.N. Kanal, J.F. Lemmer (Editors), Elsevier Science Publishers,1991, pp. 375-383.

[20] A. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[21] J. John Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[22] I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[23] K. Elissa, “Title of paper if known,” unpublished.

[24] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[25] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

ISBN: 978-0-9891305-5-4 ©2014 SDIWC