Amine Bayoudhi, Hatem Ghorbel, Houssem Koubaa, Lamia Hadrich Belguith

Sentiment Classification At Discourse Segment Level: Experiments on multi-domain Arabic corpus

Abstract

Sentiment classification aims to determine whether the semantic orientation of a text is positive, negative or neutral. It can be tackled at several levels of granularity: expression or phrase level, sentence level, and document level. In the scope of this research, we are interested in the sentence and sub-sentential level classification which can provide very useful trends for information retrieval and extraction applications, Question Answering systems and summarization tasks. In the context of our work, we address the problem of Arabic sentiment classification at sub-sentential level by *(i)* building a high coverage sentiment lexicon with semi-automatic approach; *(ii)* creating a large multi-domain annotated sentiment corpus segmented into discourse segments in order to evaluate our sentiment approach; and *(iii)* applying a lexicon-based approach with an aggregation model taking into account advanced linguistic phenomena such as negation and intensification. The results that we obtained are considered good and close to state of the art results in English language.

1 Introduction

Sentiment analysis refers to the computational study and processing of opinions, sentiments and emotions of people found in text (Al-Radaideh et al., 2014). Recently, this domain has significantly evolved and attracted widespread attention especially with the expanding growth of social networks services and user-generated web content. This situation provided a great opportunity to easily access and mine public opinions and sentiments about any subject. Business companies, for example, exploited this information source to discover consumer feedbacks about their products or even to decide future marketing actions.

Sentiment analysis includes several tasks. According to Liu (Liu, 2012), it can be divided into six main tasks: 1) extract and categorize all entity expressions from documents, 2) extract all aspect expressions and categorize them into clusters, 3) extract and categorize opinion holders, 4) extract the times when opinions are given and standardize the time formats for all opinions, 5) determine whether an opinion is positive, negative or neutral, 6) produce all opinion quintuples expressed in a document. Among these tasks, the fifth task, namely sentiment classification, is the one having received the most researcher attention.

Specifically, sentiment classification aims to determine whether the semantic orientation of a text is positive, negative or neutral. It can be tackled at many levels of granularity: expression or phrase level, sentence level, and document level. Expression sentiment classification aims to determine the prior sentiment class or valence of an expression. As for sentence level, the objective is to calculate the contextual polarity of a sentence. Concerning document level, the main goal is to mine the overall polarity of a document with the hypoth-

esis that is expressed by a single author towards a single target. In the scope of this research, we are interested in the sentence and sub-sentential level classification. This level of granularity can provide very useful trends for information retrieval and extraction applications, Question Answering systems and summarization tasks.

Sentence sentiment classification is often processed by applying machine learning techniques, in particular supervised learning which consists basically of two major steps: feature extraction and training the learning model. Though this approach has proved to be successful in producing high accuracy, it suffers from certain shortcomings. It requires building a huge corpus (dataset), which needs to be labeled manually by human experts (Abdulla et al., 2014). The process of manual annotation can be very difficult even for native speakers due to sarcasm and cultural references. It can also be expensive and time-consuming (He and Zhou, 2011). Moreover, the model built could be a domain-biased. That is, it could give low accuracy when applied in a domain, different than the domain from which it was learned (Read and Carroll, 2009). Due to these reasons, many researchers were oriented towards a second approach, namely the lexicon-based one.

In the context of our work, we address the problem of Arabic sentiment classification at sub-sentential level by *(i)* building a high coverage sentiment lexicon with semi-automatic approach; *(ii)* creating a large multi-domain annotated sentiment corpus segmented into discourse segments in order to evaluate our sentiment approach; and *(iii)* applying a lexicon-based approach with an aggregation model taking into account advanced linguistic phenomena such as negation and intensification. In fact, most of the recent works in Arabic language have not yet released their resources and some of them have common weak points such as not handling negation in the statement. In addition, the redundancy in the training data causes an ambiguity in sentiments (Ibrahim et al., 2015).

Compared to related Arabic sentiment classification work, the main contributions of this research are: *(i)* adopting a lexicon-based approach handling negation and intensification by applying state of the art strategies and establishing an extensive list of word and phrase operators, *(ii)* addressing Arabic sentiment classification at discourse segment level (first work according to our knowledge), *(iii)* experimenting Arabic sentiment classification on discussions and debates other than short comments and reviews, which is more difficult.

The rest of the paper is organized as follows. In section 2, we review a selection of key papers related to the sentiment classification for English and Arabic languages. In section 3, we describe our semi-automatic approach to build a sentiment lexicon of opinion words. In section 4, we outline our efforts to annotate two sentiment corpuses at discourse segment level. In section 5, we detail our proposed approach for sub-sentential sentiment classification, and we present and discuss the experiment results. In section 6, we sum up and provide some perspectives for future work.

2 Related work

Sentiment analysis is an emerging domain in natural language processing. Due to the explosion of user-generated web content, the interest in this field is continually increasing. Currently, dozens of research papers are published in this field in each year and many work-

shops are organized about this topic, such as WASSA (Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis) and SENTIRE (Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction). Due to the large number of publications and the variety of sentiment analysis tasks, we are interested in this section in only sentence and sub-sentential level and with an emphasis especially in works related to English and Arabic languages.

In the literature, we discern two main approaches: supervised machine learning approach and lexicon-based approach. In machine learning approach, sentiment classification is viewed as a special case of text categorization task. The sentiment classifier is built by extracting discriminative features from manually annotated sentiment data and applying a learning algorithm such as Support Vector Machines, Naïve Bayes and Maximum Entropy. Generally, the best performance is achieved by using n-grams feature, but also Part-of-speech and syntactic information can be important effective features (Pang et al., 2002); (Pak and Paroubek, 2010); (Ghorbel and Jacot, 2011).

At the sentence level, recently researches have started to study more advanced linguistic traits. For instance, Tang et al. (Tang et al., 2014) proposed a joint segmentation and classification framework for sentiment analysis in order to handle the inconsistent sentiment polarity between a phrase and the words it contains. Specifically, they used a log-linear model to score segmentation candidates, and utilize the phrasal information as features to build the sentiment classifier. The effectiveness of the joint model has been verified by applying it on the benchmark dataset of Twitter sentiment classification in SemEval 2013.

Yang et al. (Yang and Cardie, 2014) proposed an approach that allows structured modeling of sentiment while taking into account both local and global contextual information. Specifically, they incorporated intuitive lexical and discourse knowledge as expressive constraints while training the conditional random field model via posterior regularization. According to the experiments, these constraints allow achieving better accuracy than existing supervised models for the sentence-level sentiment classification.

Liu et al. 2014 (Liu et al., 2014) have focused in sentiment analysis of sentences with modality. They presented a linguistic analysis of modality and detailed its types. Then, they proposed some general linguistic features, and specific modality features to train a support vector machine classifier predicting the sentiment orientation in sentences with modalities. The reported experimental results outperformed traditional lexicon-based, unigram-based SVM and Naive Bayes classifiers.

With regard to lexicon-based approach, it typically uses a lexicon of opinion words where a sentiment polarity or intensity is associated to each entry. In addition to detected words, linguistic phenomena such as intensification and negation are often taken into account to aggregate the sentiment polarity of sentences. One of the pioneer researches in the lexicon-based approach is the work of Turney (Turney, 2002) who used a part-of-speech tagger to identify phrases that contain adjectives or adverbs, and then estimated the semantic orientation of each extracted phrase by using Pointwise-Mutual Information (PMI). The sentiment class is finally assigned to the review based on the average semantic orientation of the extracted phrases. Kim and hovy (Kim and hovy, 2004) extended Turney work by using a seed

list enriched by wordnet synonyms. They also proposed two other models to combine sentiment words, which are the product model and the geometric mean model.

Ding et al. (Ding et al., 2008) proposed a holistic lexicon-based approach to solve the problem of opinion words whose semantic orientations are context dependent in reviews by exploiting the review context. They also proposed an effective function based of sum method for aggregating multiple conflicting opinion words in a sentence.

Taboada et al. (Taboada et al., 2011) developed a Semantic Orientation CALculator (SO-CAL) based on some dictionaries where words are annotated with polarity and strength scores. SO-CAL introduced state of the art methods to deal with negation and intensification. The authors used Amazon's Mechanical Turk service to collect validation data to their dictionaries and performed their experiments on four different corpora with equal numbers of positive and negative reviews.

Chardon et al. (Chardon et al., 2013) proposed to compute the opinion orientation at the sub-sentential level using a parabolic model that accounts for the effects of negation and modality on opinion expressions based on several linguistic experiments. The parabolic model represents an opinion expression as a point on a parabola, negation as functions over this parabola and modality as a family of parabolas of different slopes. The reported evaluation of the model showed that it has good agreement with the way in which humans handle negation and modality in opinionated sentences.

Research done on Arabic sentiment analysis is considered very limited compared to other languages like English whether at document-level or sentence-level (Shoukry and Rafea, 2012). Indeed, Ibrahim and Salim demonstrated in their literature review (Ibrahim and Salim, 2013) that there is a lack of studies focusing on multilingual twitter sentiment analysis and especially on Arabic tweet opinion and Arabic tweet subjectivity. They pointed out also that the most features used for twitter SA for Arabic tweets are n-grams features, and the most methods used in twitter SA for Arabic tweets is Naive Bayes (NB) and Support Vector Machines (SVM). For instance, Shoukry and Rafea (Shoukry and Rafea, 2012) compared two machine learning techniques which are SVM and NB classifiers on 1000 collected tweets. The task is considered a sentence-level sentiment classification since tweets length was restricted to 140 characters. The authors used unigrams and bigrams as features and concluded that there is no difference in the results between them. Final classification results showed that SVM outperformed NB in sentiment analysis with an accuracy of 72.6%.

Abdul-Mageed et al. (Abdul-Mageed et al., 2014) developed the SAMAR system for subjectivity and sentiment analysis of Arabic social media using some Arabic morphological features. They used the SVM^{light} as classification algorithm and a multi-genre dataset collected from four different genres of social media websites.

Arafat et al. (Arafat et al., 2014) implemented the Aara' system for polarity classification over informal colloquial Arabic comments. The classification scheme consisted of four categories: strongly positive, positive, negative and strongly negative. Experiments were carried out on 815 comments collected from online newspapers and achieved 82% in terms of accuracy.

Recently, ElSahar and El-Beltagy (ElSahar and El-Beltagy, 2015) conducted an extensive set of experiments for the sake of benchmarking their collected datasets and testing their

viability for both two and three class sentiment classification problems. Yielded results showed that the best performing classifier was SVM and that the best effective feature representations were the combination of the lexicon based features with the other features.

Regarding lexicon-based approach, very few researches relative to Arabic were conducted. Al-Subaihin et al. (Al-Subaihin et al., 2011) introduced a sentiment analysis tool for Arabic social media. The tool is relies in merging human computation with natural language processing. Human computing aims to harness knowledge of humans in a novel way (such as a computer game), and use the gained results to solve certain steps in an otherwise fully automated system expressions. This technique was used by the authors to build sentiment lexicons. Given these lexicons and the set of negative, positive and neutral sentence patterns, user reviews are classified according to their sentiments.

Oraby et al. 2013 (Oraby et al., 2013) proposed a scalable opinion-rating system following a rule-based approach tailored to the Arabic language. The approach takes into account language-specific traits that allows for closer analysis of opinion-bearing queues such as polar words, basic negation words, intensifiers, and conjunction modifiers. The overall document rating were calculated by taking the positive polarity score over the total document polarity score to give an estimate of the document's polarity score as a ratio of polar units.

Abdulla et al. (Abdulla et al., 2014) presented detailed steps of building the main two components of the lexicon-based SA approach: the lexicon and the sentiment analysis tool. In particular, the sentiment tool was designed to take into account negation and intensification. To aggregate the review polarity score, the authors used the sum method.

3 Our approach: building the sentiment lexicon

While there has been a recent progress in the area of Arabic Sentiment Analysis, most of the resources in this area are either of limited size, domain specific or not publicly available (ElSahar and El-Beltagy, 2015). Therefore, we decided to build our own lexicon. However, since building a sentiment lexicon "from scratch" is a relatively expensive task, we have chosen to benefit from the available lexicons, enhance and enrich them in order to build our sentiment lexicon. Thus, we propose a semi-automatic approach exploiting a set of Arabic linguistic tools and resources (i.e. translator, tagger, dictionaries) and exploiting other English or multilingual sentiment lexicons (i.e. MPQA lexicon, SentiStrength lexicon). The approach includes also a manual annotation process of multi-domain collected corpus. The starting lexicon that we used to build our lexicon is MPQA Arabic translated lexicon (Elarnaoty et al., 2012).

Our approach consists of three phases including manual and automatic steps. These phases are: phase of study and cleaning, phase of enrichment, and phase of reforming and revision (Figure 1).

3.1 Phase of study and cleaning

This phase consists in manually reviewing the MPQA translated lexicon in order to detect possible defects. The process allowed us to identify the following anomalies:

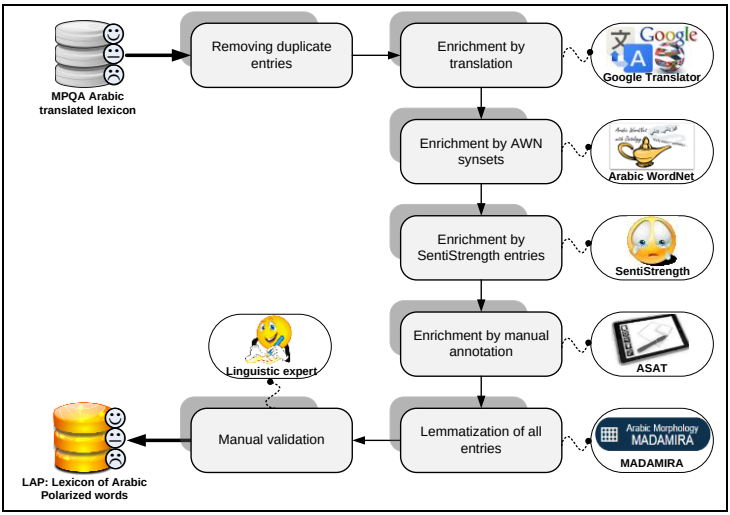

Figure 1: Creation step of the lexicon LAP

- Existence of many duplicate entries in the same class. For example: the words احتقار (contempt), أحمق (stupid), أزعج (discomfort), are found in many entries of the Negative Strong class.
- Existence of objective words (do not have **a priori** polarity). For example: معنى (sense), قرار (decision), يعامل (treat).
- Existence of misclassified words (wrong assigned polarity). For example: الخوف (fear) is classified as Strong Positive word when it should be classified as Strong Negative, ممتاز (excellent) is classified as Weak Positive instead of Strong positive, ألم (pain) is classified as Strong Positive instead of Strong Negative.
- Lack of Part of Speech (POS) tags. Indeed, MPQA translated lexicon is stored in four TXT files representing each sentimental class. POS tags are missing despite the fact that they are very useful in resolving morphological disambiguation (i.e. ذهب can be a name and means "gold", and can be a verb and means "go").

In order to remedy some of these anomalies, we performed a manual cleaning operation by eliminating duplicated words. These words are of two types: (1) duplicated words in the same class (they are unnecessary words that can be deleted without any problem), (2) duplicate words in classes that must be removed to avoid ambiguity. In fact, at this stage, the lexicon has no disambiguation technique.

3.2 Phase of enrichment

This phase consists of four steps, namely, enrichment by translation, enrichment by importing synsets of ArabicWordNet, enrichment by importing entries from the SentiStrength

lexicon and enrichment by manual annotation. In each step, only new words that do not exist in the lexicon are added to it.

- Enrichment by automatic translation: the Arabic version of MPQA does not contain the translation of all the words in the original English version. Indeed, the number of words in the English version of MPQA is 8222 and the number of words in the Arabic version is 7434. Therefore, we used the Google translator to translate the rest of the words. Among the words that have been added in this step, there are the words نزيه (probe), ثأر (revenge), بارز (marking). Note that during the translation process, there are words that can be added to the lexicon because of the existence of several translation possibilities. Similarly, there are words that can be eliminated because their corresponding Arabic words already exist in the lexicon.
- Enrichment by synsets of WordNet: This step is based on the assumption that the words having the same synonyms have the same polarity. Therefore, we have enriched our lexicon by ArabicWordNet synsets (Boudabous et al., 2013) corresponding to each word of Arabic MPQA lexicon. This is performed by exploiting the semantic relation "Near_synonym" of ArabicWordNet. Examples of the words added to the lexicon in this step are شديد (severe) that is the near synonymous of عنيف (violent), منفصل (separate) which is the closest synonym to مفكك (disassembled) and واقعي (realistic) which is the near synonymous حقيقي of (real).
- Enrichment by SentiStrength entries: To enrich our lexicon, we have exploited some other multilingual sentiment lexicons which are freely available for scientific research, such as SentiStrength (Thelwall et al., 2010). The Arabic version of this lexicon has a small size and includes only 1,434 entries. Nevertheless, this step was useful since it has fed our lexicon by new opinion words, for example, أناني (selfish) الأمثل (ideal) and اطمئنان (contentment).
- Enrichment by manual annotation: This step allows enriching the lexicon by annotating sentiment corpora. The annotation allows to mark subjective words of the text and to add them to the lexicon LAP according to their sentimental class. The annotation process is performed using an annotation tool developed for this task and called ASAT tool (Arabic Sentiment Annotation Tool). ASAT offers the possibility to the annotator to mark words with different colors according to their sentimental classes. The annotation is carried out according to the annotation scheme of MPQA translated lexicon. This scheme is composed of four classes: Negative Strong, Weak Negative, Positive Weak and strong Positive. The annotated corpus used for the enrichment of the lexicon is COPARD2 (Corpus of Arabic Opinion Debates), a collected corpus from the political domain. It consists of a set of TV political debates type of programs broadcast by Aljazeera (see section 4.1.2).

3.3 Phase of reforming and revision

To ensure good coverage of our lexicon and a size compression, we saved the opinion words as lemmas using MADAMIRA tool (Pasha et al., 2014). This allows, on the one hand, detecting all morphological variants of the opinion words, and on the other hand, saving memory space and execution time. Finally, to improve the quality of the lexicon, a validation step was conducted by a linguistic expert. The aim is to check that the polarity assigned to each word corresponds to the most frequent context in which it is used. Table 1 illustrates the evolution of the size of the lexicon after accomplishing each phase of our approach.

Table 1: Evolution of the size of the lexicon

Class	Arabic MPQA	Ph1	Ph2	Ph3
Negative Strong	2,860	1,752	2,991	1,544
Negative Weak	2,057	741	3,056	1,719
Positive Strong	1,386	922	1,990	1,278
Positive Weak	1,131	571	1,608	761
Total	7,434	3,986	9,645	5,302

4 Annotation of sentiment corpuses

In this section, we describe our efforts in annotating two sentiment corpuses at discourse segment level. In our context, we will use the annotated corpuses to evaluate the approach that we will propose afterwards (section 5) to classify Arabic discourse segments according to their polarity. However, these corpuses can be as well exploited for machine learning purposes. First, we present our data collection and provide statistics on each corpus. Second, we argue the adoption of discourse segmentation and present the used tool. Third, we explain the annotation process and outline the annotation model and guidelines. Finally, we discuss the obtained results.

4.1 Data collection

4.1.1 OCA (Opinion Corpus for Arabic)

It consists of 500 documents divided equally into positive and negative (Table 2). The corpus was collected by extracting reviews about movies from Arabic web pages and blogs (Rushdi-saleh et al., 2011). After that, many processing steps on each review were carried out in order to obtain a formatted document. The main steps were removing HTML tags and special characters, correcting spelling mistakes, filtering out nonsense and nonrelated comments, fixing Romanized comments and comments in different languages. OCA was annotated at document level by classifying its documents into positive and negative. This classification was automatically performed by exploiting the review rating score given by the user.

4.1.2 COPARD2 (Corpus of OPinion Arabic Debates 2)

It consists of 20 episodes of political debates broadcast on Aljazeera satellite channel. Most of the discussed issues were related to the political situations in the countries of Arab Spring especially in Tunisia and Egypt. All selected debates were multiparty conversations involving speakers of different profiles: politicians and political activists, economists, sociologists, security experts etc. These debates were transcribed and made publicly available in PDF format in Aljazeera website. After downloading the debates, some preliminary preprocessing steps were performed such correcting spelling and conversion mistakes, normalization and standardization of some Arabic letters, and removing thematic header of the debates.

Table 2: specificities of OCA and COPARD2

Property	OCA	COPARD2
Domain	Movie reviews	Politics
Number of documents	500	20
Number of words	209,733	97,172
Avg. of words per document	419	4,858
Number of segments	18,377	8,234
Avg. of segments per document	36	411
Avg. of words per segment	11.41	11.8

4.2 Segmentation into discourse segments

The objective of this work is to investigate sentiment classification at local level which is coarser than expression level and finer than document level. This requires splitting text into several segments of tokens connected by a structural or semantic relation. Most researchers have gone for considering sentences as these portions of text, and a lot of studies in sentiment classification were focused on sentence level. Nevertheless, in our current research, we do not share the opinion that sentences are the appropriate segmentation unit to study sentiment classification at local level, and this is due to many reasons.

First, theoretical definition of a sentence as "a part of a speech or a written discourse that has a complete and independent meaning" (Khalifa et al., 2011) do not offer, in Arabic case, enough cues to estimate sentence boundaries. As a matter of fact, unlike Indo-European languages, in Arabic there is no capitalization, which makes recognizing sentence boundaries a harder task. In addition, there are no strict rules of punctuation. This is leads to a rare use of punctuation marks, and even if they are used, they are not decisive cues to guide the segmentation process (Belguith et al., 2008). Moreover, Arabic discourse tends to use long and complex sentences. The average number of words per sentence is larger than the average in English sentence. For example, we can easily find too long paragraph with only one punctuation mark at the end (Keskes, 2015); (Aliwy, 2012).

Second, given the additional length of Arabic sentence compared to other languages, this sentence contains often more than one opinion expression and opinion target. Or, dealing with more than one opinion target per segmentation unit is a very complicated task in sentiment classification. That's why, most researches adopt an ignorance strategy in this issue and start from the hypothesis that the detected opinion expressions are expressed towards a single target, even if the problem is tackled at a document level (they are expressed throughout a whole document). Hence, it will be more beneficial to adopt a segmentation unit which is shorter than sentence in order to maximize chances to have only one target per segmentation unit.

Third, adopting a minimal segmentation unit shorter than sentence will help to resolve another problem in sentiment classification which is defining the scope of opinion operators. Actually, local polarity is highly affected by opinion operators in particular negation operators. Estimating opinion words affected by these operators is a very challenging task.

Among efficient solutions proposed to this problem is to adopt a minimal segmentation unit and to propagate the negation effect to all opinion terms of this segmentation unit.

Given the reasons explained above, we decided to split our texts into discourse segments and to study sentiment classification at this local level of granularity. Indeed, According to Chardon (Chardon, 2013), after splitting his corpus into discourse segments or Elementary discourse Units (UDE), 90% of these units contain only one opinion expression. An EDU is mainly a sentence or clause in a complex sentence that typically correspond to a verbal clause, as in [*I loved this movie*]$_a$ [*because the actors were great*]$_b$ where the relative clause introduced by the discourse connective because, indicates a cutting point. An EDU can also correspond to other syntactic units describing eventualities, such as prepositional and noun phrases, as in [*After several minutes,*]$_a$ [*we found the keys on the table*]$_b$ (Keskes et al., 2014).

To ensure the best segmentation quality, segmentation process was semi-automatically performed and conducted by an Arabic native speaker annotator using two segmentation tools. The main tool was ATS (Arabic Text Segmenter) developed by Keskes (Keskes, 2015). ATS was designed to divide documents into EDU following the Segmented Discourse Representation Theory (SDRT) principles. It adopts a pattern based approach relying on punctuation marks and discourse connectors as cues. ATS, evaluated in a collected corpus of elementary school books, has achieved good results around 85.5% in terms of F-measure. Nevertheless, in our case, our data collection consists of spontaneous user generated content and transcribed dialogues which are much more difficult to segment than school books written with regular Arabic discourse style. Hence, in order to ensure that the segmented output do not contain too long sentences neither broken sentences or clauses, another segmentation proposition is offered to the annotator by using STAr tool. STAr was developed by Belguith (Belguith et al., 2005) to segment non-vowelled Arabic texts into paragraphs and sentences. The tool adopts an approach based on a contextual analysis of the punctuation marks and a list of particles, such as the coordination conjunctions. Evaluation results on a corpus of elementary school books yields a precision of 80.65%. Given the two segmentation propositions, the annotator intervenes to choose the appropriate segmentation output consisting of minimal segments or clauses holding a complete and independent meaning. Final segmentation results are illustrated in Table 2.

4.3 Annotation process and guidelines

In the literature, many annotation models were proposed in sentiment classification either at expression level or at document level. For instance, in MPQA Project (Wilson et al., 2006), the proposed model to annotate news articles, included three attributes: polarity, intensity with 4 values, and explicit or implicit character. A more detailed model is proposed by Daille et al. (Daille et al., 2011) which contains five semantic categories: opinion, appreciation, agreement/disagreement, judgment and acceptation/refuse.

In Arabic language, the most known annotation work was realized by Abdul-Mageed et al. (Abdul-Mageed et al., 2012) who introduced AWATIF, a multi-genre Arabic corpus labeled at sentence level. The corpus was labeled using both regular and crowdsourcing methods with two types of annotation guidelines: simple and linguistically-motivated. For the simple annotation, two examples of positive, negative and neutral sentences were provided to the

annotators as guidelines, then, they were asked to label the data according to these three categories. As for the linguistically-motivated annotation, before starting the annotation task, annotators were exposed to a linguistics background, and the nuances of the genre to which each dataset belongs were explained to them. The annotated datasets consisted of collections of newswire stories from various domains (e.g., political, economic, sports), 30 Wikipedia Talk Pages, and web forum extracts comprising 2532 threaded conversations from 7 forums. Depending on these datasets, the annotation model was updated by adding "MIXED" or "OBJECTIVE" classes. Annotator Agreement reached good rates that vary from 0.79 to 0.82 depending on the dataset.

In the current research, we have assigned the task of annotating our data collection for sentiment analysis to two students familiar with natural language processing domain. The two students were Arabic native speakers and postgraduate. This profile was selected based on Abdul-Mageed et al. funding's who assert that linguistics background can be very useful for sentiment labeling since the concepts of subjectivity and sentiment are fuzzy. Indeed, the authors reported an achieved improvement of linguistically-motivated annotation when compared to simple annotation guidelines. Our Annotation model and guidelines are described in the two next sections.

4.3.1 Annotation model

Our annotation model (Figure 2) is based on two levels of granularity: expression level and discourse segment level. In each level, we tried to rely only on basic sentiment attributes that are essential for our task. The objective is first to reduce the expansive cost of the annotation task in terms of time consumption and money; and second, to maximize the annotator agreement which can be degraded when the annotation model contains too many categories.

Briefly, our annotation model consists:

- At the expression level of three attributes:
  - Polarity: polarity of the expression which can be positive or negative. Neutral expressions are not labeled.
  - Intensity: intensity of the expression which can be strong or weak.
  - Introducer: term used to introduce an opinion expression. It can be evaluative (polarized) or non-evaluative (non-polarized).

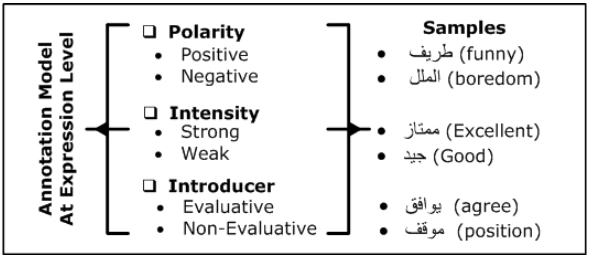

Figure 2: Annotation model at Expression level

- At discourse segment level it only consists of the attribute "Polarity" which allows classifying discourse segments depending on their semantic orientation into positive, negative and neutral. Neutral class includes objective segments that do not express any position or sentiment and also segments that express neutral position or sentiment.

Table 3. Sample of each type of discourse segments

Polarity	Sample
Positive	كما لعب العمل الجماعي دورا هاما في نجاح الفيلم
	Teamwork has also played an important role in the success of the movie
Negative	ولكن ليس هناك اتفاق حول المخرج من هذه الأزمة
	But there is no agreement on the way out of this crisis
Neutral	استفساري الثاني يتعلق بالمغربية سناء موزيان
	The second question is about the Moroccan Sana Mouziane

4.3.2 Annotation guidelines

Annotation guidelines are a set of instructions communicated to the annotators to help them to further understand the task and resolve the difficult encountered cases. Therefore, in our guidelines, we provided several examples to the annotators illustrating each sentiment attribute in the annotation model. In addition, to simplify the task, we made the assumption that each segmentation unit contains only one opinion target, but of course it may contain more than one opinion expression. This will help the annotators to focus more on opinion expressions than extracting the target or the holder, which is beyond the scope of this paper. Moreover, the annotators were asked to label all morphological forms that can bear an opinion or a sentiment: adjectives, adverbs, nouns and verbs. In fact, adjectives are significant indicators of opinion expressions. However, that does not mean that other Part-Of-Speech tags do not contribute to opinion expressions. Indeed, a lot of researchers point out that adjectives and adverbs are better than adjectives alone and certain verbs and nouns are also strong indicators of sentiment (Xia et al., 2011).

Nevertheless, to ensure a good progress of the annotation process, the two annotators were trained separately by annotating, under our supervision, the first 10 documents of OCA corpus and the first political debate of COPARD2 corpus.

4.4 Annotation Results

In order to evaluate the annotation task at discourse segment level, we used the confusion matrix to visualize the numbers of segments in which annotators agree and disagree according to each annotation category. Each column of the matrix represents the instances labeled by the first annotator, while each row represents the instances labeled by the second annotator. Figure 3 and Figure 4 illustrate respectively the confusion matrix related to OCA and the confusion matrix related to COPARD2.

		Annotator 2			
		Positive	**Negative**	**Neutral**	**Total**
Annotator 1	**Positive**	3,579	435	1,203	5,217
	Negative	260	2,966	1,013	4,239
	Neutral	978	1,946	6,485	9,409
	Total	4,817	5,347	8,701	18,865

Figure 3: confusion matrix relative to OCA

Then, to measure the annotator agreement, Cohen's kappa coefficient is computed. The results we obtined were 0.51 and 0.35 respectively for OCA and COPARD2. This agreement rate is considered very poor and subsequently, the annotated data cannot be considered enough homogenous to compare the machine results to it.

		Annotator 2			
		Positive	**Negative**	**Neutral**	**Total**
Annotator 1	**Positive**	408	27	93	528
	Negative	21	393	53	467
	Neutral	809	1,100	5,464	7,373
	Total	1,238	1,520	5,610	8,368

Figure 4: confusion matrix relative to COPARD2

By observing the two confusion matrices, we can easily find the cause of the rather poor agreement rates. Indeed, we can clearly see that the high number of instances where annotators disagree concerns always the "Neutral" category. Therefore, to resolve the problem, we decided to abandon neutral segments and to transform the problem to a binary classification task. Hence, we have removed agreed neutral segments (segments agreed by the two annotators to be neutral) as well as disagreed neutral segments (segments labeled by of one the annotators as neutral). The updated confusions matrices are illustrated in Figure 5 and Figure 6.

		Annotator 2		
		Positive	**Negative**	**Total**
Annotator 1	**Positive**	3,579	435	4,014
	Negative	260	2,968	3,228
	Total	3,839	3,403	7,242

Figure 5: confusion matrix relative to OCA without Neutral category

With these new confusion matrices values, Kappa rates for OCA corpus has reached 0.8 and for COPARD2 corpus 0,89. These new results are considered good enough for the purpose of using the annotated data for the sentiment classification task.

		Annotator 2		
		Positive	**Negative**	**Total**
Annotator 1	**Positive**	408	27	435
	Negative	21	393	414
	Total	429	420	849

Figure 6: confusion matrix relative to COPARD2 without Neutral category

4.5 Creation of Gold Standard

Although removing neutral category has resolved the problem of the poor annotator agreement, it has severely decreased the number of annotated segments in our collected data from 27,233 segments to 8,089 segments. In order to increase the size of our annotated data, we have revised our decision of abandoning neutral category, and we have decided to reject only agreed neutral segments. For the disagreed neutral segments, an adjudication operation is applied by a senior annotator (me) to add them to our final Gold standard version. The adjudication process included also disagreed polarized segments which are 695 segments in OCA corpus and 48 segments in COPARD2 corpus. The final properties of the Gold Standard versions of OCA and COPARD2 are presented in Table 4.

Table 4. Statistics on the Gold Standard versions of OCA and COPARD2

	OCA	COPARD2
Positive segments	7,455	1,794
Negative segments	4,931	1,110
Total	12,386	2,904

4.6 Discussion

Improving researches in sentiment analysis relies basically on availability of linguistic resources, in particular sentiment corpus. Such resource is required to conduct the linguistic study of the problem, to carry out a machine learning technique, or to evaluate the implemented proposed solution. In spite of that, sentiment corpuses annotated at local level are very rare. This is due to the high and expansive cost necessary to build them. Actually, according to our knowledge, Abdul-Mageed et al. (Abdul-Mageed et al., 2012) work is the only consistent attempt to create a sentiment corpus annotated at sentence level for Arabic language, and the annotated data are not yet released. In comparison to their work, we tried in this research to perform sentiment classification at a finer level, which is discourse segment. The strength of using this level of granularity was explained in section 4.3. In addition, similarly to Abdul-Mageed et al. work, our data collection is multi-domain including movie

review and political discussions. The final number of annotated segments in the standard Gold version is over than 15,000 segments.

5 Our approach for Sentiment classification

As seen in the literature survey, sentiment classification can be tackled by adopting a machine learning approach or by setting up a lexicon-based model. Our proposed approach (Figure 7) to classify discourse segments according to their polarity is based on a rule model and a sentiment lexicon to detect opinion expressions. It consists of three phases: preprocessing steps, detection of opinion expression, and computation of polarity score of the discourse segment. The first and second phases exploit Arabic linguistic resources, while, the third phase is language-free.

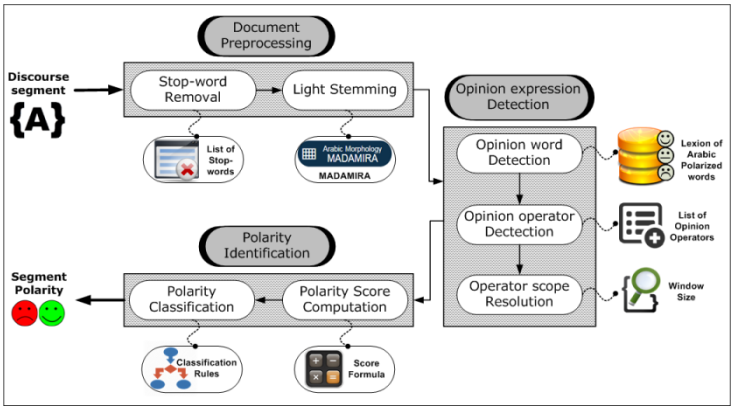

Figure 7: Steps of the proposed approach

5.1 Preprocessing steps

Preprocessing steps are required to accelerate and optimize the detection of opinions. They consist of two steps: stop-words removal and word stemming.

5.1.1 Stop-word removal

To accelerate the detection process of opinion words, we have profited from the stop-word list of Khoja stemmer tool (Khoja and Garside, 1999). In fact, this Stop-word list was widely used in Arabic processing community (Al-kabi, 2013) (Ababneh et al., 2012) (Sawalha and Atwell, 2008), but it was established to serve information retrieval applications. In sentiment classification task, a more reduced list is required, because many non-informative bearing words (such as negation operators and discourse markers) can serve as helpful cues in sentiment classification. Therefore, the stop-word list was revised to be tailored to sentiment classification constraints.

5.1.2 Stemming

Unlike Indo-European languages such as English and French, stemming in Arabic language is more difficult, mainly due to the fact that Arabic is an agglutinative and derivational language. Indeed, in Arabic, there are many more affixes than English and this leads to a large number of word forms. Besides, as a property of words in Semitic language, Arabic stem has an additional internal structure consisting of a two parts namely "root" and "pattern". The root is usually a sequence of three consonants and has some abstract meaning. The pattern is a template that defines the placement of each root letter among vowels and possibly other consonants. For example, in Figure 8, the word "AlkitabAn", meaning the two books, has a root composed of the letters k, t, and b; and his pattern is "Root1+i+Root2+a+Root3" (Heintz, 2010).

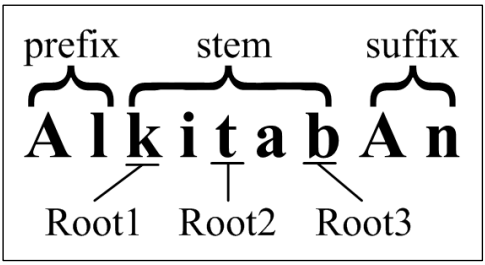

Figure 8: Example of Arabic stemming (Heintz, 2010)

Many Arabic tools, in particular morphological taggers, allow extracting roots from words. But, very few of them provide stems. To our knowledge, MADAMIRA (Pasha et al., 2014) is the only available light stemmer which performs morphological analysis and disambiguation of Arabic. Therefore, we used MADAMIRA to apply a light stemming on each document. Light stemming aims to reduce words to their lemma forms: for verbs, this is the 3rd person masculine singular perfective form and for nouns, this corresponds to the singular default form (Abdul-Mageed et al., 2014). In fact, stemming, which reduces words to their roots, is not convenient in Arabic language, because it may affect the word sense. Light stemming will be helpful to detect all morphological variations of the word.

5.2 Opinion expression detection

Opinion can be defined as a quadruplet Op=(w, t_w,h_w,$opers_w$), where:

- w is the opinion expression,
- t_w is the opinion target or topic,
- h_w is the opinion holder,
- $opers_w$ is the operator list affecting the opinion expression (Chardon, 2013).

In the same way, opinion expression is defined as "the minimal portion of text bearing an opinion". Hence, to identify the sentiment class of an opinion, it is necessary to identify the polarity of the words forming its opinion expression, and to analyze the effects of its operators.

5.2.1 Opinion bearing word detection

Once a sentiment lexicon is available, detecting opinion bearing words becomes a relatively simple task. In fact, preprocessing steps allow reducing the search scope by removing stop-words, and they allow also optimizing the detection process by mining the stems instead of the words themselves. Subsequently, the detection task becomes a naive comparison of two strings.

However, a more advanced treatment of this task may invoke the semantic disambiguation problem. Some word sense ambiguities are addressed by taking part of speech (POS) into account. For instance, *plot* is only negative when it is a verb, but should not be so in a noun dictionary; *novel* is a positive adjective, but a neutral noun (Taboada et al., 2011). Nevertheless, in Arabic language, this problem is much more challenging since most Arabic texts are non-vowelized. This leads to a high number of possible candidate solutions. For instance, "كرم" with POS=Noun can be vowelized as "كَرَمٌ" (generosity) which is positive, or as "كَرْمٌ" (vineyard) which is neutral.

In the current research, given the structure of the used sentiment lexicon LAP, opinion word detection was reduced to simple task especially that LAP is still under construction and his entries do not include POS information yet.

5.2.2 Opinion operator detection

Opinion operators or modifiers are linguistic elements which do not intrinsically bear opinions, but they are altering the characteristics of opinion words located in their scope (Chardon, 2013). In the course of our research, we consider only the two main opinion operators: intensifiers and negation operators. A limited list of each opinion operator category is prepared by a linguistic expert. Other operators such modality operators (Liu et al., 2014) and conditional operators (Narayanan et al., 2009) are left for future work.

- *Intensifiers*: they are operators altering the polarity or the intensity of the opinion expression. We distinguish two types of intensifiers: *(i)* amplifiers (i.e. very, much, extremely) which strengthen the intensity of the opinion expression, *(ii)* attenuators (i.e. little, less) which weaken the intensity of the opinion expression. It is notable that most intensifiers are adverbs and that many of them are term-compound such as "إلى حدّ كبير" (pretty much).

- *Negation operators*: Negation is a very common linguistic construction that affects polarity and, therefore, needs to be taken into consideration in sentiment analysis (Wiegand et al., 2010). Similarly to other language such as English (Taboada et al., 2011) and French (Benamara et al., 2012), negation can be introduced by different ways, through: *(i)* negators such as "not" and "without", *(ii)* quantifiers such as "never" and "nobody", *(iii)* lexical negation such as "absence" and "lack of".

In practice, according to their relative position to the opinion expression, we have classified negation operators into two categories: right operators and left operators. Right operators are the main negation words, while left operators can coexist in the same segment to play the role of quantifiers.

The detection of these operators follows the same technique described above concerning opinion bearing words. However, they are stored in specific separate lists since they do not

bear opinions or sentiment and subsequently have not prior polarity scores. Table 5 illustrates samples of Arabic opinion operators.

Table 5. Samples of opinion operators

Opinion operators	Samples
Amplifiers	جدا، كثيرا، (very, much)
Attenuators	قليلا، بعض الشيء، (little, slightly)
Right negation operators	لا، ما، لن، دون (not, less)
Left negation operators	أبدا، بتاتا، (never, at all)

5.2.3 Resolution of the opinion operator scope

While identifying intensifier scope is a simple task and can be performed by locating the closest opinion word to the intensifier, identifying the negation scope is among the challenging tasks in sentiment classification. In fact, negation scope and its effects have been a subject of interest for many researchers, not only in sentiment analysis domain, but also in many other fields such as philosophy, logic, and psycholinguistics (Morante and Sporleder, 2012). Basically, negation scope can be defined as the parts of a sentence whose meaning is inverted by a negation word. To resolve this problem, two major approaches were proposed in the literature: rule-based approach and machine learning approach. The rule-based approach relies upon linguistic rules which seek for negation words in the sentence and invert the meaning of certain surrounding parts based on different predefined window sizes (Prollochs et al., 2015). For instance, Hogenboom et al. (Hogenboom et al., 2011) have achieved a significant increase in overall sentiment classification accuracy when applying a two word window in a set of English movie review sentences. Concerning the machine learning approach, many techniques were applied to predict negation scope such conditional random fields (Councill et al., 2010) and Hidden Markov Models (Prollochs et al., 2015)

However, since machine learning approach, in particular supervised methods, requires an annotated training data which is unavailable and difficult to create, we have chosen to follow a window-based method to resolve negation scope. Hence, a set of experiments were carried out to determine the most effective window size. Obtained results are presented in the section 5.4.1.

5.3 Identification of the segment polarity

Identifying the polarity of the discourse segment depends, as we mentioned earlier, on the detected opinion bearing words and on the opinion operators affecting them. In this section, we explain the mapping process from the expression level to the discursive segment level; in other words, how can we exploit opinion words and operators to identify the polarity of a segment called the contextual polarity?

5.3.1 Prior Polarity

After detecting opinion bearing words, a polarity score P_w is assigned to each word according to its polarity and its intensity. This score, representing the prior polarity of the word or

the out of context polarity, is calculated according to the formula: $P_w = Pol_w * int_w$ where Pol_w is the polarity of the word and int_w is its intensity. Pol_w and int_w are respectively determined according to the lexicon sentiment class ant the intensity class into which the word belongs. Pol_w for the "Positive" class is 1 and for the "Negative" is -1. Int_w for the "Weak" class is 1 and for the "Strong" is 2. So, for example, the word "احتفل" (celebrate) which belongs to the "Positive" polarity class and the "Strong" intensity classe, P_w(احتفل)= 1*2=2.

5.3.2 Operator effect

Concerning negation operators, their effect on opinion expression is addressed at the local level by following one of three main strategies:

- Polarity reversal: called also switch negation. It is the classic approach for dealing with negation in sentiment analysis. It consists of changing the polarity sign of the opinion expression (Sauri 2008). For example, if P_w("good")=3 in a scale of [-5..5], then P_w("not good") will be -3.

- Polarity linear shift: first introduced by Taboada et al. (Taboada et al., 2011) who pointed out that polarity reversal works well in certain cases but fails drastically in others. For example, if P_w("Excellent")=5, we cannot say that P_w("not excellent")=-5. Therefore, they proposed treat negation by shifting the intensity towards the opposite polarity by a fixed amount. The amount used in implementation was 4. So, P_w("not excellent") will be 1.

- Polarity angular shift: first introduced by Chardon et al. (Chardon et al., 2013) who represent opinion expression by a point E of a parabola of focus F and summit O. The angle OḞE allows measuring the polarity score of the opinion expression. The negation effect is computed by adding/subtracting π to/from the angle OḞE. For example, if P_w("Excellent")=5π/6 in a scale of]-π..π[(5 in a scale of [-5..5]) , then P_w("not excellent") will be –π/6 (-1 in a scale of [-5..5]).

Concerning intensifiers, there are also three strategies in the literature to handle their effect on opinion expression:

- Addition and subtraction: It is the simplest way to deal with intensifiers. It consists in adding or subtracting a fixed value to/from the intensity of the opinion expression depending on the intensifier type. For example, if P_w("tired")=-3, then P_w("very tired") will be -4 and P_w("bit tired") will be -2 (Kennedy and Inkpen, 2006).

- Multiplicative factor: first introduced by Taboada et al. (Taboada et al., 2011) who considered that intensification should depend on the item being intensified. So, to each intensifying word, they have associated a percentage from -50 to 100 and created a separate dictionary for adjectival intensifiers. The polarity of an expression containing an intensifier operator is computed as $P_{exp} = P_w*(100\%+Perct_{int})$ where P_w is the polarity of the opinion word and $Perct_{int}$ is the percentage of the intensifier.

- Angle adjustment: introduced by Chardon et al. (Chardon et al., 2013) who treated intensification by increasing or decreasing the angle OḞE in their parabolic model.

Although Chardon et al. and Taboada et al. approaches have adopted different shift forms and values, they share the same strategy considering that negation is not always polarity reverser; It is basically polarity shifter and it affects also the intensity of the opinion expression. In the course of our research, we adopt the same strategy of Taboada et al. concerning

negation and intensification. As a matter of fact, we have applied polarity shift and multiplicative factor with different shift values from the one proposed by Taboada et al. since our intensity scale is different.

5.3.3 Computation of segment polarity score

After taking into account the different components affecting the sentiment classification of the segment, we have to put them together in order to compute the contextual polarity. In the literature, different heuristics are applied to perform the mapping from expression level to segment, sentence or document level. For instance, Yuan et al. (Yuan et al., 2013) proposed a simple sentiment word-count method to classify domain specific datasets. The method identifies the polarity of the text on the basis of the number of detected positive and negative opinion word. In other words, if the number of positive words bigger than the number of negative words, the text is positive; otherwise, it is negative.

Another research that addressed this issue is the work of Kim and Hovy who built and compared three models to assign a sentiment category to a given sentence (Kim and hovy, 2004). The first model computes the product of the signs of the sentiment polarities in the region (parts of the sentence in which sentiments would be considered). The second is the harmonic mean (average) of the sentiment strengths in the region, and the third is the geometric mean. The authors pointed out that the first model achieved the best overall performance.

However, the most used model to compute sentence polarity score is the sum model (Ding et al., 2008); (Oraby et al., 2013); (Tromp and Pechenizkiy, 2013); (Ghosh and Kar, 2013). It consists of summing up all opinion expression scores, upon which the result can be normalized depending on the used scale.

In the context of this research, we have used the sum model and normalized it by dividing it with the sum of the number of opinion words in the segment. Hence, $Pol_{seg} \in [-1..1]$ and it is computed according to the formula:

$Pol_{seg} = \frac{\sum_{i=1}^{n} Pol_{w_i}}{n}$ where Pol_{wi} is the polarity of the word w_i an n is the number of opinion words in the opinion expression.

5.3.4 Polarity identification

Given the polarity score of the discourse segment, a set of three rules are applied in to order to identify its polarity:

if $Pol_{seg} > 0$, the polarity of the segment is positive.

if $Pol_{seg} < 0$, the polarity of the segment is negative.

if $Pol_{seg} = 0$, the polarity of the segment is undetermined. Since, our classification scheme is binary and we do not take into account neutral segments, these segments are considered as misclassified segments when evaluating the approach.

5.4 Evaluation and discussion

In this section, we describe our performed experiments to evaluate negation scope resolution, negation effect, and the sentiment classification approach.

5.4.1 Negation scope resolution

To determine the most effective window size to use for negation scope, we conducted a set of experiments with different window sizes on OCA corpus. These experiments (Table 6) are performed with considering only negation operators (e.g. without intensifiers) and with applying polarity reversal as effect.

Table 6. Obtained results with different window sizes

Window size	Accuracy	Precision	Recall	F-score
1	70.63	67.38	89.69	76.95
2	71.39	68.33	88.85	77.25
3	71.96	69.17	87.85	77.40
4	71.83	69.30	87.04	77.16
5	71.50	69.23	86.14	76.77

Results show that, although the three-sized window has achieved the best F-score value, there is no significant difference in the classification between the different applied window sizes. This is similar to Dadvar et al. results (Dadvar et al., 2011) who have evaluated the effect of different window sizes in negation detection on 2000 movie reviews. Obtained accuracies were very close along the 5 windows. This may be explained by the fact that in movie reviews, many of sentiments are expressed implicitly. In addition, a more detailed approach about double negations and combined negation intensification has to be studied.

5.4.2 Negation effect

Two major strategies are proposed to deal with negation effect in sentiment analysis: polarity reversal and polarity shift. In these experiments (Table 7), we have evaluated these two strategies on OCA corpus by ignoring intensification and by adopting a three-sized window for negation scope as this was the size achieving the best performance in the previous experiments.

Table 7. Evaluation of the negation effects

Negation effect	Accuracy	Precision	Recall	F-score
Polarity reversal	71.96	69.17	87.85	77.40
Polarity shift	67.74	64.63	90.53	75.42

Although polarity shift strategy seems to be a more relevant strategy, it has achieved slightly worse F-score than the polarity reversal strategy. A possible explanation to this result may be the choice of the shift value. In reality, there is no rule which determines the fixed amount to shift from the intensity of the opinion expression when the negation word is encountered. This is can be a subject of an empirical study that investigates the shift value on the basis of the intensity scale.

5.4.3 Sentiment classification

In this section, we present our final classification results on OCA and COPARD2 with three-sized window as negation scope, polarity shift as negation effect, and by taking intensification into account.

Table 8. Obtained results with the proposed method

Corpus	Accuracy	Precision	Recall	F-score
OCA	70.48	67.91	87.18	76.35
COPARD2	71.41	67.79	83.58	74.86

Despite the fact that Arabic is a morphologically rich language that faces many challenging issues in sentiment analysis (Ibrahim et al., 2015), the proposed approach achieved relatively close results to the state of the art of English sentiment classification especially for OCA corpus. This proves that our build lexicon has a good coverage quality and that the implemented rules can constitute a good start for a high accurate classifier. Nevertheless, much more work is required to take into account more linguistic forms such as modal auxiliaries and conditional.

6 Conclusion and future work

In this paper, we have presented a lexicon-based approach for Arabic sentiment classification at sub-sentential level. First, we started by building a sentiment lexicon by following a semi-automatic approach. The lexicon entries were used to detect opinion words and assign to each one a sentiment class. Second, we proceeded to the annotation of new sentiment corpuses at discourse segment level. These corpuses are then used for the evaluation of the lexicon-based approach. This approach relies on an aggregation model taking into account advanced linguistic phenomena such negation and intensification. Evaluation results were considered good and not too far from state of the art results in English language.

As perspectives, we intend to enhance the lexicon quality by implementing a semantic disambiguation component based on POS information. This will improve the detected sentiment classes of the opinion words. In addition, we intend to improve existing strategies of treating negation and intensification by conducting more experiments especially on the effect of negation. Other opinion influencer cues like modalities and conditional sentences can be studied at this stage. Moreover, we intend to exploit the annotated corpuses to train a machine learning classifier for the sentence sentiment classification.

References

Abdulla et al. (2014). Nawaf A. Abdulla, Nizar A. Ahmedn Mohammed A. Shehab, Mahmoud Al-Ayyoub, Mohammed N. Al-Kabi,Saleh Al-rifai, Towards Improving the Lexicon-Based Approach for Arabic Sentiment Analysis, International Journal of Information Technology and Web Engineering, 9(3), 55-71, July-September 2014

Abdul-Mageed et al. (2012). Abdul-Mageed, M., & Diab, M. T. (2012). AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis (pp. 3907–3914). LREC.

Abdul-Mageed et al. (2014). M. Abdul-Mageed, M. Diab, and S. Kübler. Samar: Subjectivity and sentiment analysis for Arabic social media. Computer Speech & Language, 28(1):20–37, 2014.

Ababneh et al. (2012). Ababneh M., Al-Shalabi R., Kanaan G., Al-Nobani A.; Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness, International Arab Journal of Information Technology (IAJIT) . Jul2012, Vol. 9 Issue 4, p368-372.

Aliwy. (2012). Ahmed H. Aliwy. Tokenization as Preprocessing for Arabic Tagging System. International Journal of Information and Education Technology, Vol. 2, No. 4, August 2012.

Al-Kabi (2013). Al-Kabi M.N.; Towards improving Khoja rule-based Arabic stemmer, Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference,p1-6, 3-5 Dec. 2013, Amman, Jordan.

Al-Radaideh et al. (2014). Qasem A. Al-Radaideh, Laila M. Twaiq, Rough Set Theory for Arabic Sentiment Classification, 2014 International Conference on Future Internet of Things and Cloud.

Al-Subaihin et al. (2011). Al-Subaihin, A., Al-Khalifa, H., & Al-Salman, A. (2011). A proposed sentiment analysis tool for modern Arabic using human-based computing. the 13th International Conference on Information Integration and Web-based Applications and Services ACM.

Arafat et al. (2014). Arafat, H., Elawady, R., Baraka,S. and Elrashidy, N. Different Feature Selection for Sentiment Classification, International Journal of Information Science and Intelligent System.

Belguith et al. (2005).Lamia Hadrich Belguith, Leila Baccour et Ghassan Mourad, Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules, TALN 2005, Dourdan, 6-10 juin 2005.

Belguith et al. (2008). Lamia Hadrich Belguith, Chafik Aloulou et Abdelmajid Ben Hamadou, MASPAR : De la segmentation à l'analyse syntaxique de textes arabes. Revue Information Interaction Intelligence I3, vol 7, N2, mai 2008.

Benamara et al. (2012). Farah Benamara, Baptiste Chardon, Yannick Mathieu, Vladimir Popescu, Nicholas Asher, How do Negation and Modality Impact on Opinions?, Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM-2012).

Boudabous et al. (2013). M.M. Boudabous, N. Chaâben Kammoun, N. Khedher, L. Hadrich Belguith, F. Sadat, "Arabic WordNet semantic relations enrichment through morpho-lexical patterns", ICCSPA'13, , February 12-14, Sharjah, UAE, 2013.

Chardon et al. (2013).Baptiste Chardon, Farah Benamara, Yannick Mathieu, Vladimir Popescu, Nicholas Asher.Sentiment Composition Using a Parabolic Model. In International Workshop on Computational Semantics (IWCS 2013), Potsdam Germany.

Chardon. (2013). B. Chardon, "Chaîne de traitement pour une approche discursive de l'analyse d'opinion", Phd dissertion, UPS, France, 2013

Councill et al. (2010).I. G. Councill, R. McDonald, and L. Velikovich, "What's Great and What's Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis," in Proceedings of the Workshop on Negation and Speculation in Natural Language Processing.

Dadvar et al. (2011). Dadvar, Maral and Hauff, Claudia and Jong de, Franciska (2011) Scope of negation detection in sentiment analysis. In: Dutch-Belgian Information Retrieval Workshop.

Daille et al. (2011). Daille, Béatrice, Estelle Dubreil, Laura Monceaux, and Matthieu Vernier. Annotating Opinion evaluation of Blogs: The Blogoscopy Corpus. Language Resources and Evaluation 45 (4) (June 29): 409–437.

Ding et al. (2008). X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 International Conference on Web Search and Data Mining.

Elarnaoty et al. (2012). Elarnaoty, M., AbdelRahman, S., & Fahmy, A. (2012). A machine learning approach for opinion holder extraction in Arabic language.

ElSahar and El-Beltagy. (2015). H. ElSahar and S. R. El-Beltagy. Building large Arabic multidomain resources for sentiment analysis. In Computational Linguistics and Intelligent Text Processing.

Ghorbel and Jacot (2011). Ghorbel H., Jacot D. Sentiment Analysis of French Movie Reviews. In Studies in Computational Intelligence - Advances in Distributed Agent-Based Retrieval Tools. Ed. Springer, Vol 361. pp 97–108.

Ghosh and Kar (2013). Ghosh M., Kar A., Unsupervised Linguistic Approach for Sentiment Classification from Online Reviews Using SentiwordNet 3.0, International Journal of Engineering Research & Technology (IJERT), vol.2 Issue 9, September - 2013.

He and Zhou. (2011). He, Y., & Zhou, D. (2011). Self-training from labelled features for sentiment analysis. Information Processing & Management, 47(4), 606–616.

Heintz (2010). Heintz I., Arabic Language Modeling with stem-derived morphemes for automatic speech recognition, Phd dissertation, The Ohio State University, USA.

Hogenboom et al. (2011). A. Hogenboom, P. van Iterson, B. Heerschop, F. Frasincar, and U. Kaymak, Determining Negation Scope and Strength in Sentiment Analysis, in Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, 2011, pp. 2589–2594.

Ibrahim and Salim. (2013). Ibrahim, M. and Salim, N. opinion analysis for twitter and Arabic tweets: a systematic literature review. Journal of Theoretical and Applied Information Technology.

Ibrahim et al. (2015). Hossam S. Ibrahim , Sherif M. Abdou and Mervat Gheith, Sentiment analysis for modern stanadard Arabic and colloquial, International Journal on Natural Language Computing.

Kennedy and Inkpen. (2006). A. Kennedy and D. Inkpen. Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, 22(2):110–125, 2006.

Keskes et al. (2014). Iskandar Keskes, Farah Benamara and Lamia Hadrich Belguith. Splitting Arabic Texts into Elementary Discourse Units. Journal ACM Transactions on Asian Language Information Processing. Volume 13, Issue 2, June 2014.

Keskes. (2015). Iskandar Keskes, Discourse Analysis of Arabic Documents and Application to Automatic Summarization, Phd dissertation, UPS, France, 2015

Khalifa et al. (2011). I. Khalifa, Z. Feki, A. Farawila, Arabic discourse segmentation based on rhetorical methods, In Electric Computer Sciences. 11, 1, 2011

Khoja and Garside (1999) Khoja S. and Garside R., Stemming Arabic Text, UK: Computing Department, Lancaster University.1999

Kim and hovy. (2004). Soo-Min Kim, Eduard H. Hovy: Determining the Sentiment of Opinions. COLING 2004.

Liu et al. (2014). Yang Liu, Xiaohui Yu, Bing Liu, and Zhongshuai Chen, Sentence-Level Sentiment Analysis in the Presence of Modalities, CICLing 2014, Part II, LNCS 8404, pp. 1–16, 2014.

Liu. (2012). Liu, B. Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers.

Morante and Sporleder. (2012). Roser Morante, Caroline Sporleder: Modality and Negation: An Introduction to the Special Issue. Computational Linguistics 38(2): 223-260 (2012)

Narayanan et al. (2009). Narayanan, R., Liu, B., Choudhary, A.: Sentiment analysis of conditional sentences. In: EMNLP, pp. 180–189. Association for Computational Linguistics (2009)

Oraby et al. (2013). Shereen Oraby, Yasser El-Sonbaty, Mohamad Abou El-Nasr:Finding Opinion Strength Using Rule-Based Parsing for Arabic Sentiment Analysis. MICAI (2).

Pak and Paroubek. (2010). Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. in Proceedings of LREC'10, Valletta, Malta, 2010.

Pang et al. (2002). B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proceedings of EMNLP 2002.

Pasha et al. (2014). A. Pasha, M. Al-Badrashiny, M.T. Diab, A. El-Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, R. Roth, "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic". LREC 2014.

Prollochs et al. (2015). Nicolas Prollochs,Stefan Feuerriegel, Dirk Neumann, Enhancing Sentiment Analysis of Financial News by Detecting Negation Scopes, 48th Hawaii International Conference on System Sciences, 2015 .

Read and Carroll. (2009). Read, J., & Carroll, J. (2009). Weakly supervised techniques for domain-independent sentiment classification. the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (pp. 45-52). ACM.

Rushdi-saleh et al., (2011) Rushdi-Saleh M., Martín-Valdivia M. T., Ureña-Ló L. A., Perea-Ortega J. M.. OCA: Opinion corpus for Arabic. Journal of the American Society for Information Science and Technology, 62(10), 2045–2054.

Shoukry and Rafea. (2012). Shoukry, A., & Rafea, A. (2012). Sentence-level Arabic sentiment analysis. Collaboration Technologies and Systems (CTS) (pp. 546–550). IEEE

Sawalha et al. (2008). Sawalha M. and Atwell E.S. Comparative evaluation of Arabic language morphological analysers and stemmers. In Proceedings of 22nd International Conference on Comptational Linguistics COLING 2008, 18-22 August, Manchester.

Taboada et al. (2011).Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267–307.

Tang et al. (2014). Duyu Tang, Furu Wei, Bing Qin, Li Dong, Ting Liu, Ming Zhou, A Joint Segmentation and Classification Framework for Sentiment Analysis, Proceedings of (EMNLP), pages 477–487, October 25-29, 2014, Doha, Qatar.

Thelwall et al. (2010). Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.

Tromp and Pechenizkiy. (2013). Erik Tromp, Mykola Pechenizkiy: RBEM: a rule based approach to polarity detection. WISDOM 2013: 8.

Turney. (2002). Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proceedings of ACL 2002 (pp. 417-424).

Wiegand et al. (2010). Michael Wiegand, Alexandra Balahur, Benjamin Roth and Dietrich Klakow, Andrés Montoyo, A Survey on the Role of Negation in Sentiment Analysis, Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, Uppsala, July 2010

Wilson et al. (2006). Wilson, Theresa, Janyce Wiebe, and Rebecca Hwa. 2006. “Recognizing Strong and Weak Opinion Clauses.” Computational Intelligence 22 (2): 73–99

Xia et al. (2011). R. Xia, C. Zong, S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181.

Yang and Cardie (2014). "Bishan Yang, Claire Cardie,Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization, In Proceedings of the ACL 2014.

Yuan et al. (2013). Yuan, Ying Liu and Hui Li, Sentiment Classification in Chinese Microblogs: Lexicon-based and Learning-based Approaches, International Proceedings of Economics Development and Research, V68. 1., 2013.