Sentiment Classification of Arabic Documents: Experiments with multi-type features and ensemble algorithms

Amine Bayoudhi
ANLP Group, MIRACL
FSEGS, Sfax University
3018, Sfax, TUNISIA
bayoudhi.amine@gmail.com

Lamia Hadrich Belguith
ANLP Group, MIRACL
FSEGS, Sfax University
3018, Sfax, TUNISIA
l.belguith@rnu.fsegs.tn

Hatem Ghorbel
ISIC Lab, HE-Arc
Applied Science University
CH-2610, Switzerland
hatem.ghorbel@he-arc.ch

Abstract

Document 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. In the literature, most existing researches rely on n-grams as selected features, and on a simple basic classifier as learning model. In the context of our work, we try to improve document classification findings in Arabic sentiment analysis by combining different types of features such as opinion and discourse features; and by proposing an ensemble-based classifier to investigate its contribution in Arabic sentiment classification. Obtained results attained 85.06% in terms of macro-averaged F-measure, and showed that discourse features have moderately improved F-measure by approximately 3% or 4%.

1 Introduction

With the expanding growth of social networks services, user generated content web has emerged from being a simple web space for people to express their opinions and to share their knowledge, to a high value information source for business companies to discover consumer feedbacks about their products or even to decide future marketing actions. Therefore, opinion mining is becoming a potential research domain interesting more and more researchers who attempt to improve current results and to solve more advanced and complex issues in the domain. Typically, mining opinions is viewed as a classification problem called sentiment classification. 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, which is our focus in this research, the main goal is to mine the overall polarity of a document with the hypothesis that is expressed by a single author towards a single target.

Document 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. In the literature, most existing researches rely on n-grams as selected features, and on a simple basic classifier as learning model. The limit of these two choices is revealed when shifting from one domain to another. As a matter of fact, in one hand, each domain has generally his specific vocabulary. So, n-grams features produced from one domain fail to be discriminative in another. In the other hand, numerous studies showed that the performance of classification algorithms is domain dependent (Xia et al., 2011).

In the context of our work, we try to improve document classification findings in Arabic sentiment analysis by (*i)* combining different types of features such as opinion and discourse features; and by (*ii)* proposing an ensemble-based classifier consisting of a set of accurate basic classifiers to investigate its contribution in Arabic sentiment classification similarly to some other languages such as Chinese (Wang et al., 2014).

29th Pacific Asia Conference on Language, Information and Computation pages 196 - 205
Shanghai, China, October 30 - November 1, 2015

The rest of the paper is organized as follows. In section 2, we review a selection of related work to document sentiment classification for English and Arabic languages. In section 3, we detail our proposed approach and focus on the feature extraction and the classification model selection steps. In section 4, we describe the conducted experiments and discuss the obtained results. Finally, we summarize our conclusions and provide some perspectives.

2 Related Work

2.1 English Sentiment classification

In English sentiment classification, various strategies have been proposed, (Liu, 2012). The most effective ones are related to machine-learning paradigm, viewing the opinion and polarity detection as text classification tasks. These techniques vary from supervised to unsupervised learning, typically probabilistic methods such as Naïve Bayes (NB) and Maximum Entropy (MaxEnt), and linear discrimination methods such as Support Vector machine (SVM). As other possible classification schemes, we mention non-parametric classifiers such as k-Nearest Neighbor (KNN), as well as similarity scores methods (i.e. phrase pattern matching, distance vector, frequency counts and statistical weight measures).

Nevertheless, to get a good accurate classifier, we need to select the most effective set of textual predictors (Liu and Motoda, 2008). In sentiment classification, n-grams (Pang et al., 2002) are the most used features, however, there are some researches where other semantic features are tested such us opinion words and phrases, opinion operators such as negation (Mejova et al., 2011), parts of speech (Wang et al., 2014), and syntactic dependencies (Nakagawa et al., 2010). Some other researches attempt to integrate discourse features and report a significant added value of rhetorical roles in sentiment classification (Chardon et al., 2013). For instance, Somasundarun et al. (Somasundarun et al., 2009) proposed a supervised and unsupervised methods employing Discourse relations to improve sentiment classification. This is performed by adopting relational feature that exploit discourse and neighbor opinion information.

In general, most of adopted features tend to be domain specific (e.g., the term *television* has a negative polarity in a movie review, but may have a positive one in a book review). This problem can be solved by the second approach: the lexicon based approach.

Lexicon-based approach relies on a sentiment lexicon to calculate orientation for a document from the semantic orientation of words or phrases in the document (Taboada et al., 2011). Sentiment lexicon is a collection of classified opinion terms that can be compiled according to three approaches: dictionary-based approach, corpus-based approach, or combined approach.

In dictionary based approach, we attempt to find a set of opinion seed words and then enrich them by retrieving their synonyms and antonyms from dictionaries such WordNet and Thesaurus. For instance, Hu and Liu (Hu and Liu, 2004) and Esuli and Sebastiani (Esuli and Sebastiani, 2005) classify polarity using emotion words and semantic relations from WordNet, WordNet Gloss, WordNet-Affect and SentiWordNet respectively.

However, in corpus-based approach, we use patterns in particular syntactic ones to mine large domain specific corpora and extract opinion terms. Among well-known researches in lexicon-based approach, we mention those of Taboada et al. (Taboada et al., 2011) who developed a semantic orientation calculator called SO-CAL. They started by manually creating a sentiment lexicon by annotating a large corpus of reviews extracted from *Epinions* website. The lexicon was enhanced by positive and negative words from the General Inquirer dictionary. To calculate the semantic orientation of each review, the authors took in consideration intensification by multiplying intensifier words by a percentage, and they incorporated Negation by shifting the semantic orientation toward the opposite polarity by a fixed amount.

Note that some researches combined the machine learning and the lexicon based approaches by exploiting a sentiment lexicon in the framework of a supervised learning method (Mejova et al., 2011) (Maynard et al., 2011).

2.2 Arabic Sentiment Classification

Most of the work in sentiment analysis was devoted to the English language, an important number of resources and tools have been elaborated accordingly. When addressing the same issue to other target languages such as Arabic, several difficulties come out as potential challenges, including the lack of standard lexical and sentiment resources and of good accurate linguistic analyzers and parsers. That’s why, we consider that Arabic sentiment classification is still limited compared to English.

Nevertheless, there are many published research papers focusing on sentiment classification of Arabic documents. These researches have been the object of some surveys (Korayem et al., 2012) (Al-Twairesh et al., 2014). For example, we cite Abbasi et al. (Abbasi et al., 2008) who proposed a machine learning method based on entropy weighted genetic algorithms to classify movie reviews and forum comments in English and Arabic. Conducted experiments based mainly on stylistic features yielded an accuracy of 93.62% but with a high computational cost.

Rushdi-Saleh et al. (Rushdi-Saleh et al., 2011) have introduced in their research a new collected corpus of movie reviews called OCA (Opinion Corpus for Arabic). They reported as well as some experiments based on n-grams words and carried out with SVM and NB classifiers. The best F-measure attained 90.73% with SVM classifier.

Mountassir et al. 2013 (Mountassir et al., 2013) investigated three classification settings in an n-grams framework based on three classifiers namely NB, SVM and KNN. The tested settings are stemming type, term frequency thresholding and term weighting. Experiments are performed on two data collections: OCA and ACOM (collected by the authors). Best results in terms of F-measure attained 93% on OCA with KNN classifier and 87.5% and 76.4% respectively on ACOM DS1 and ACOM DS2 with NB classifier.

El-Halees (El-Halees, 2011) followed an hybrid sequential approach by applying lexicon-based method with a seed word list enriched from online dictionaries. Classified documents were then used to train a MaxEnt based classifier. Classified documents of the two previous steps were finally used to train a KNN based classifier. Experiments were conducted on a multi-domain corpus consisting of 1143 documents. Achieved accuracy was around of 80%.

3 Proposed Approach

In this section, we present our approach proposed for the sentiment classification of Arabic documents. This approach, based on multi-type features, is using a set of publicly available linguistic resources and tools. It takes as input an Arabic review about a given target and predicts its polarity which can be Positive or Negative. The approach consists chiefly of three sequential phases which are composed of one or more steps. The three phases are: document pre-processing, feature extraction, and sentiment classification.

3.1 Data Description

In Arabic language, sentiment resources are in general rare. However, in the task of document sentiment classification, there are many used data collections since they are easy to collect and to annotate. In fact, we remark that each researcher has collected his own datasets and used in the evaluation of his classification approach, which does not allow comparing properly the obtained results. Therefore, we have decided to use in our experiments existent datasets that have been widely used by the NLP research community.

According to the literature, there are few publicly available sentiment corpora for document sentiment classification. They are derived from different domain such as social networks (Abdullah et al., 2013), product reviews (Abbasi et al., 2008) (Rushdi-Saleh et al., 2011) (Mountassir et al., 2013) and news (Ahmad et al., 2006) (Almas et al., 2007). Among these corpora, the most used one is OCA (Rushdi-Saleh et al., 2011) and the largest one is ACOM (Mountassir et al., 2013). That’s why, we have chosen these two corpus to evaluate our approach and to compare our results.

OCA (Opinion Corpus for Arabic) consists of 500 documents divided equally into positive and negative (Table 1). The corpus was collected by extracting reviews about movies from Arabic web pages and blogs. 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. The classification of documents into positive and negative were automatically performed by exploiting the review rating score given by the user. This annotation strategy avoids wasting time in manual annotation, but, it does not always succeed to assign the right class to the annotated review. In fact, reviewers can mention much more negative feedbacks than positive ones, but give a weak positive rating score to the movie.

Property	**Neg.**	**Pos.**
Total documents	250	250
Total tokens	94,556	121,392
Avg. tokens in each file	378	485
Total sentences	4,881	3,137
Avg. sentences in each file	20	13

Table 1: Statistics on OCA

ACOM (Arabic Corpus for Opinion Mining) is a multi-genre corpus collected from Aljazeera polls and forums. It consists of three datasets of different domains. The first dataset DS1 consists of 594 documents and falls within movie review domain. The second dataset DS2 is sport-specific dataset and consists of 1492 comments about 18 sport topics. The third dataset DSP2 is a collection of 1082 comments about a political issue titled “Arab support for the Palestinian affair”. ACOM were manually annotated according to four classes: positive, negative, neutral and dialectal. Then, neutral and dialectal categories were eliminated since the authors were interested in classification by polarity of documents written only in Modern Standard Arabic (Table 2).

Dataset	Positive	Negative	Total
DS1	184	284	468
DS2	486	517	1003
DS3	149	462	611
Total	819	1263	2082

Table 2: Statistics on the collected ACOM

In addition, the authors proceeded to eliminate a number of negative comments from each dataset in a way to equalize the number of documents for each category (Mountassir et al., 2013). The final number of documents used in experiments is 1368 documents: 698 negative and 670 positive (Table 3).

Property	Negative	Positive
Total documents	698	670
Total tokens	45697	38819
Avg. tokens in each file	65.46	57.93

Table 3: Statistics on the datasets of ACOM used in experiments

3.2 Document preprocessing

Before going on with the classification task, some preprocessing steps are necessary to prepare the raw documents to the feature extraction step. This step requires to search and to identify a set of lexical cue words and markers. To this end, three main steps are required: segmentation, stemming and stop-word removal.

Segmentation: This step, which we carried out using Stanford word segmenter (Monroe et al. , 2014), includes text normalization and word segmentation. Normalization aims to normalize the spelling of some Arabic characters which can be written in different ways. Arabic text can be vowelized, non-vowelized, or even partially vowelized. To ensure the detection and extraction of all orthographic word forms, we decided to eliminate discretization from the reviews. Normalization is also applied to some characters such as alef by transforming all his forms (Alef Hamza above "أ" and Alef Hamza below "إ") into bare Alef "ا". This process is applied because many reviewers omit or confuse these similar letters and use them interchangeably.

Stemming: MADAMIRA (Pasha et al., 2014) is used to apply a light stemming on the reviews. Light stemming aims, to transform nouns in singular and to conjugate verbs with the third personal pronoun. 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.

Stop-word removal: To accelerate the detection process of the lexical cues, we have profited from the stop-word list of Khoja stemmer tool (Khoja and Garside, 1999) and revised it. In fact, this Stop-word list was established to serve information retrieval applications. However, in sentiment classification task, a more reduced list is required, because many non-informative bearing words (such as negation operators and discourse markers) can be helpful cues in sentiment classification.

3.3 Extraction of classification features

In English language, several features ranging from lexical to deep analysis features were tested in the sentiment classification task. However, in Arabic, research works were focused on lexical or statistical features in particular n-grams. This is due to many reasons basically the lack of sentiment resources (i.e. lexicons, standard annotated corpora) and high accurate linguistic tools (i.e. syntactic parser, segmenter). That’s why, we propose to adopt a set multi-type features. Our selected features are: opinion features, discourse markers, stylistic features, domain dependent features and morpho-lexical features. In feature extraction step, a set of linguistic resources and tools are required.

Opinion features: include opinion bearing words and opinion operators. Opinion bearing words were detected using a sentiment lexicon called LAP (Bayoudhi et al., 2014). It is an Arabic lexicon that contain over than 8,000 entries, semi-automatically constructed from the MPQA Arabic translated lexicon (Elarnaoty et al., 2012). It is also fed by mapping synonyms from Arabic Wordnet (Boudabous et al., 2013), by manual annotation of sentiment corpora and by entries from multilingual sentiment lexicons. Statistics on this lexicon are illustrated in Table 4.

Regarding Opinion operators, they are linguistic elements which do not intrinsically bear opinions, but they are altering the characteristics of opinion words located in their scope (Chardon, 2013).

Class	Number of entries
Negative Strong	2,281
Negative Weak	2,689
Positive Strong	1,726
Positive Weak	1,437
Total	8,133

Table 4: Statistics on the lexicon LAP

In the course of our research, we propose to classify opinion operators in three categories: intensifiers, negation operators, and epistemic modality operators. A list of each opinion operator is prepared by a linguistic expert.

- *Intensifiers*: they are operators altering the intensity of the opinion expression. We distinguish two types of intensifiers: *(i)* amplifiers (i.e. very, much, extremely) strengthen the intensity of the opinion expression, *(ii)* attenuators (i.e. little, less) weaken the intensity of the opinion expression.

- *Negation operators*: affect the polarity of the opinion expression (i.e. not, never, neither). This effect is handled at the sentence level by following different strategies such as switch polarity (Sauri, 2008) and linear shift polarity (Taboada et al., 2011) and angular shift polarity (Chardon, 2013).

- *Epistemic modality operators*: Epistemic modality serves to reveal how confident writers are about the truth of the ideational material they convey (Palmer, 1986). There are two types of epistemic modality operators: hedges and boosters. Hedges (i.e. perhaps, I guess) are words employed by the speaker to reduce the degree of his liability or responsibility towards the expression. Boosters (i.e. definitely, I assure that and of course) are elements used by the speaker to emphasize the expression. Both hedges and boosters modify polarity of the opinion expression, either strengthen or weaken it (Abdul-Mageed et al., 2012).

Discourse features: In document sentiment classification, many research studies have investigated the integration of deep analysis techniques through syntactic parsing and dependency relations, or through discourse analysis and role relation detection. Accordingly, we propose in our research to follow the same approach by adopting discourse features. In fact, compared to dependencies relations, discourse relations contain, in addition to the structural aspect, a semantic aspect which can be exploited in the sentiment classification. However, unfortunately, discourse processing researches in Arabic are very limited. It focuses on either annotating corpus with discourse information (Al-saif and Market, 2010) or proposing taxonomies of discourse relations (Khalifa et al., 2012). Therefore, it is not possible to profit, in Arabic language, from an automatic generated discourse structure or an automatic recognition of discourse relations to improve sentiment classification. Hence, discourse analysis can be exploited only through discourse markers called also discourse connectives (DC) (Asher, 1993). To use these discourse markers, we have adopted the list of Arabic Text Segmenter (Keskes, 2015), an Arabic tool that segments text into elementary Discourse Units. This list is structured in a discourse relation hierarchy containing 24 relations categorized in four main classes: thematic, temporal, causal and structural. In the context of this work, we started by exploiting only the structural class. This class contains 7 relations illustrated in Table 5.

Relation	Sample of DCs
Contrast	في المقابل، إلا أن، بينما
Antithetic	في حين أن، ليس،
Concession	غير أن، لكن، بيد أن
Correction	لا بل، كلا، إنما
Alternation	سواء، أم، أو
Parallel	كذلك، كما، مع
Conditional	لو، شرط أن، إذا

Table 5: Discourse relation hierarchy that we used in sentiment classification

To exploit these DCs in our classification model, we have grouped them according to their effect in opinion expressions into three feature categories: polarity propagation, polarity switch, conditional polarity (Figure 1).

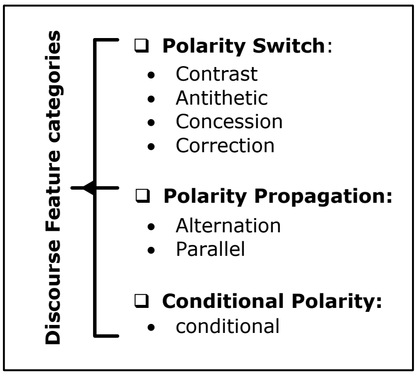

Figure 1: Proposed discourse features for the sentiment classification

Stylistic features: consist mainly of:

- Punctuation marks: three punctuation marks are considered in our research: period (full stop), question mark and exclamation mark. Comma is not taken into account since it is not often used in Arabic writing.

- Number of words per document.

- Polarities of the first and last expressed opinion words: based on the assumption that the first and last sentences of a document are the most informative sentences, we added to our stylistic features the polarity of the first expressed opinion word and the polarity of the last expressed opinion word.

Domain dependent features: n-grams are widely used as features in text classification and sentiment classification for their capacity to encode word order information and substantially the context of the document (Pang et al., 2002). However, these features are domain dependent; they cause a big decrease in the performance of the classifier when testing it with other data collections. Therefore, we have decided to minimize the effect of domain dependent features by excluding unigrams and relying only on bigrams and trigrams. Choosing bigrams and trigrams is explained also by the fact that the lexicon LAP does not contain compound words. Hence, to feed the classifier with compound words, we selected a set of bigrams and trigrams based on their frequency.

Morpho-lexical features: Since adjectives and adverbs are the most morphological forms expressing opinions, and since the lexicon LAP do not include Part-of-speech information, we propose to consider, as additional features, the number of positive adjectives and adverbs and also the number of negative adjective and adverbs in each document.

3.4 Feature transformation

Feature transformation step determines the numerical representation used in the classification process. It’s performed by applying a weighting scheme on the extracted textual data of the corpus. We distinguish three weighting schemes: binary, term frequency and TF-IDF representation. Binary schema takes into account presence or absence of a term in a document. Term frequency considers the number of times a term occurs in a document (Li et al., 2009). TF-IDF (Term Frequency - Inverse Document Frequency) considers not only term frequencies in a document, but also the relevance of a term in the entire collection of documents (Manning et al., 2008).

Many researches confirm that the most suitable representation for sentiment classification is binary since overall sentiment may not usually be highlighted through repeated use of the same terms. In fact, Pang et al. (Pang et al., 2002) showed in their experiments that better performance is obtained using presence rather than frequency, that is, binary-valued feature vectors in which the entries merely indicate whether a term occurs or not formed a more effective basis for review polarity classification. Whereas, Mountassir et al. (Mountassir et al., 2013) point out that TF-IDF is also a suitable weighting for SVM and KNN.

3.5 Attribute selection

Attribute selection aims to evaluate the effectiveness of features by identifying relevant features ones to be considered in the learning process. This is allows performing an intense dimensionality reduction without losing on the classifier accuracy.

There are many algorithms for attribute selection such as information gain (Abbasi et al., 2008), mutual information, and chi-square (Li et al., 2009). None of them has been widely accepted as the best feature selection method for sentiment classification, despite the fact that information gain has often been competitive: it ranks terms by considering their presence and absence in each class (Moraes et al., 2013).

3.6 Learning Algorithm

Apart from classification features, Sentiment classification task depends highly on the used learning algorithm. According to the literature, the most popular algorithms are NB, SVM, MaxEnt, Artificial Neural Networks (ANN). Many studies were interested in evaluating and comparing these learning techniques and experimental findings confirm that a given learning algorithm can outperform all others only for a specific problem or an exact subset of the input data, it is abnormal to find a single algorithm achieving the best results on the overall problem domain (Kuncheva, 2004). For instance, a lot of authors reported that they achieved the best performance with SVM in their experiments (liu et al., 2011) (Rushdi Saleh, 2011). Moraes et al. (Moraes et al., 2013) affirm that ANN produce superior or at least comparable results to SVM. Other researchers claim that they yield the best performance by applying KNN and NB (Mountassir et al., 2013).

The proposed solution for this problem is adopting the ensemble technique. This technique consists in combining, in an efficient way, the outputs of several classification models to form an integrated output. We distinguish in the literature many combination types such as sum, voting, weighted combination and meta-classifier (Xia et al., 2011). In the context of our research, we are focusing on four well-known ensemble algorithms namely Bagging (Breiman, 1996), Boosting (Schapire et al., 1998), Voting (Kuncheva, 2004) and Stacking (Syarif et al., 2012).

4 Experiments and discussion

In this section, we carried out two types of experiments. The first type of experiments focus on evaluating a set of base learning algorithms versus a set of ensemble based classifiers. The objective is to find the combination configuration that ensures the best and stable performance across different domains. The second type of experiments concentrates on the feature set used in the classification process. The objective is to evaluate, in particular, the effectiveness of discourse features in Arabic sentiment classification.

All results reported in this section are obtained by applying an evaluation method based on 10-folds cross validation. Attribute transformation, attribute selection and learning algorithms are applied using the Weka data mining software (Hall et al., 2009). Binary representation is used for opinion features, and the TF-IDF representation for bigrams and trigrams.

4.1 Base Classifier evaluation

In this first type of experiments, we conducted a comparative evaluation of three well-known classification algorithms namely SVM, MaxEnt and ANN. The objective is to determine the best accurate base algorithm in each dataset. Many attempts with different parameters are made to achieve the best performance. These experiments were performed on the following data collections: OCA, ACOM DS1, ACOM DS2, ACOM DS3, ACOMB (ACOM Balanced data), ACOMA (ACOM All data).

Obtained results are expressed in terms of F-measure (Table 6). But, since we are applying our model to several sets of data, we need an averaging evaluation metric to get an idea about the best performance in the overall experiments. There are two types of averaging methods: macro-averaged and micro-averaged. Macro-averaging gives equal weight to each dataset, whereas micro-averaging gives equal weight to each document. Because the F-measure ignores true negatives and is mostly determined by the number of true positives, large datasets dominate small datasets in micro-averaging. Micro-averaged results are therefore really a measure of effectiveness on the large datasets in a test collection (Van Asch, 2013). Hence, we have used in Table 6 macro-averaged F-measure to compare the performance of the three algorithms on the different datasets.

Compared to earlier work, our results overstep state of the art existing performances. Indeed, MaxEnt classifier tested in OCA has achieved 95% of F-measure, which exceed 93% reported by Mountassir et al. with KNN classifier (Mountassir et al., 2013) and 90.73% reported by Rushdi-Saleh et al. with SVM classifier (Rushdi-Saleh et al., 2011). Similarly, our obtained results in ACOM DS1 and ACOM DS2 which are respectively 89.3% and 80.1% exceed Mountassir et al. reported results (87.5% and 76.4%). Concerning ACOM DS3, ACOMA and ACOM, these data collections are not yet evaluated by any earlier work.

Dataset	SVM	MaxEnt	ANN
OCA	91.8	95	90.6
ACOM DS1	86	87.5	89.3
ACOM DS2	80.1	80	76.2
ACOM DS3	89.5	86.2	86.8
ACOMB	80.5	80.1	77.8
ACOMA	79.6	75.7	76.1
Macro-avg	84.58	84.08	82.8

Table 6: Results of the base classifiers

According to our experiments, SVM seems to be the most stable classifier among the three classifiers. In fact, it achieved the best results on ACOM DS2, ACOM DS3, ACOMB and on ACOMA.

Regarding OCA and ACOM DS1, the best performance was yielded respectively by MaxEnt and ANN, although these two datasets are derived from the same domain (movie review) and have a relatively close size. The difference in terms of F-measure between the two classification results is considered significant since it exceeds 5.5%. As for DS2, results were less good than the other datasets. In fact, although the documents are in the same domain, they talk about 18 different sports, which make the dataset relatively heterogeneous. On the other hand, DS3 is derived from political specific domain which is a very large domain, but all documents discuss only one political issue, which explain why the classification results were good. Concerning ACOMA and ACOMB, as expected, results were better in ACOMB in which the number of documents is much less than ACOMA.

4.2 Ensemble classifier evaluation

In addition to the evaluation of base classifiers, we conducted another set of experiments to evaluate ensemble classifiers with the same datasets and evaluation metrics. The combination of the classifiers is performed according to the four methods: boosting, bagging, voting and stacking. Several experiments are performed to choose the base classifiers and the combination method that reach the best performance. At the end, we have maintained these four experiments: *(i)* bagging MaxEnt, *(ii)* boosting MaxEnt, *(iii)* majority voting with SVM, MaxEnt and ANN as base classifiers, *(iiii)* stacking SVM and MaxEnt with Linear regression as meta-classifier. The results achieved in each experiment are illustrated in terms of F-measure and macro-averaged F-measure in Table 7.

Dataset	Bag.	Boost.	Vot.	Stack.
OCA	95	94	93.2	94.8
ACOM DS1	92.9	89.4	90.4	87.8
ACOM DS2	80.3	79.6	80.6	79.7
ACOM DS3	88.2	84	90.3	88.2
ACOM B	79.4	79.9	81.4	80
ACOM A	75.9	74	79.7	79
M. Avg	84.7	83.38	85.06	84.26

Table 7: Results of ensemble based classifiers

Compared to Table 6, Table 7 indicates that most of the selected ensemble classifiers have exceeded the results yielded by base classifiers in terms of macro-averaged F-measure. In particular, majority voting of MaxEnt, SVM and ANN has achieved the best results with a macro-averaged F-measure of 85.06%. In five among six datasets, this ensemble classifier has performed better results than the best base classifiers.

4.3 Discourse feature evaluation

In order to evaluate the effectiveness of the discourse features, we have reapplied the best accurate base algorithms on our datasets with removing discourse features. This was performed with respecting all pre-mentioned constraints of attribute transformation and attribute selection steps. Table 8 presents the new achieved F-measure in each dataset with the best accurate algorithm obtained according to the experiments described in section 4.1.

Dataset	Best classifier	F-meas. (%)	Diff (%)
OCA	MaxEnt	92.6	-2.4
ACOM DS1	ANN	85.3	-4
ACOM DS2	SVM	79.7	-0.4
ACOM DS3	SVM	89.2	-0.3
ACOMB	SVM	80.5	0
ACOMA	SVM	79.82	0

Table 8: Discourse feature evaluation

Obtained results show that discourse features are more efficient with OCA and ACOM DS1 derived from movie review domain. In fact, removing discourse features with OCA and ACOM DS1 has respectively decreased F-measure by 2.4% and 4%. This can be explained by the fact that in movie review domain in particular, discourse markers are frequently employed.. Nevertheless, regarding ACOM DS2 and ACOM DS3, the results were not very altered by removing discourse features since F-measure has decreased only by 0.3% and 0.4%. So, this type of features is not very efficient for sport or political domain. Concerning the two last experiments, removing discourse features while evaluating ACOMA and ACOMB has not revealed any impact on the classification results.

5 Conclusion

In this paper, we have proposed a supervised classification approach of Arabic documents. The proposed approach is based on multi-type feature set including opinion features, discourse markers, stylistic features, domain dependent features and morpho-lexical features. In addition, we have carried out a comparative study between some well-known base classifiers and some ensemble-based classifier with different combination methods. Obtained results showed that MaxEnt, SVM and ANN combined with majority voting rules have achieved the best results with a macro-averaged F1-mesaure of 85.06%. Furthermore, experiments showed that discourse features have improved F-measure by approximately 3% or 4%.

As perspectives, we intend to integrate discourse structure and relations as features. This is can be performed by exploiting cross lingual discourse parsing of parallel sentiment corpora, since there is no Arabic discourse parser. In addition, following the same approach, we intend to adopt also syntactic information and dependency relations as classification features.

References

Abbasi A., Chen H., Salem A. 2008. Sentiment analysis in multiple languages: Feature selection or opinion classification in Web forums. ACM Transactions on Information Systems 26(3):12.

Abdul-Mageed M., Diab M. 2012. AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In Proc LREC 2012, pp. 3907-3914.

Ahmad K., Cheng D., Almas Y. 2006. Multi-lingual sentiment analysis of financial news streams. In Proc. of the 1st International Conference on Grid in Finance.

Almas Y., Ahmad K. 2007. A note on extracting sentiments in financial news in English, Arabic & Urdu. In Proc of Workshop on Computational Approaches to Arabic Script-based Languages.

Al-Saif A., Markert K. 2010. The Leeds Arabic discourse treebank: Annotating discourse connectives for Arabic. In Proc of LREC, 17-23 May, Malta.

Al-Twairesh N., Al-Khalifa H., Al-Salman A. 2014. Subjectivity and Sentiment Analysis of Arabic: Trends and Challenges. In Proc. Of the 11th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA' 2014), Doha, Qatar, 10-13 November.

Asher N. 1993. Reference to Abstract Objects in Discourse. Kluwer Dordrecht.

Bayoudhi A., Koubaa H., Ghorbel H., Hadrich Belguith L. 2014. Vers un lexique arabe pour l’analyse des opinions et des sentiments, In Proc of the 5th International Conference on Arabic Language Processing CITALA’14, Oujda, Morocco, 26 – 27 November.

Boudabous M.M., Chaâben Kammoun N., Khedher N., Hadrich Belguith L., Sadat F. 2013. Arabic WordNet semantic relations enrichment through morpholexical patterns, in Proc. of the First International Conference on Communications, Signal Processing, and their Applications (ICCSPA’13), Sharjah, UAE, February 12-14.

Breiman L. 1994. Bagging predictors. Technical Report 421, Department of Statistics, University of California, Berkeley.

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

Elarnaoty M., AbdelRahman S., Fahmy A. 2012. A machine learning approach for opinion holder extraction in Arabic, International Journal of Artificial Intelligence & Applications; March 2012, Vol. 3 Issue 2, p45.

El-Halees A. 2011. Arabic Opinion Mining Using Combined Classification Approach. In Proc of the international Arab Conference on Information Technology, 11-14 December, Riyadh, Saudi Arabia.

Esuli A., Sebastiani F. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In Proc of the fifth international conference on Language Resources and Evaluation, 22-28 May, Genoa, Italy.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H. 2009. The WEKA data mining software: an update, ACM SIGKDD Explorations Newsletter, Volume 11 Issue 1, pp 10-18, June 2009.

Hu M., Liu B. 2004. Mining Opinions Features in Customer Reviews. Proc of 19th international conference on Artificial Intelligence AAAI, 25-29 July, San Jose, California, pp. 755-760.

Keskes I. 2015. Discourse Analysis of Arabic Documents and Application to Automatic Summarization. PhD dissertion, UPS France.

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

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

Korayem M., Crandall D., Abdul-Mageed M. 2012. Subjectivity and Sentiment Analysis of Arabic: A Survey. in Advanced Machine Learning Technologies and Applications AMLTA, 322.

Kuncheva L. 2004. Combining pattern classifiers: methods and algorithms. John Wiley & Sons.

Li S., Xia R., Zong C., Huang C.-R. 2009. A framework of feature selection methods for text categorization. In Proc of the 47th Annual Meeting of the Association for Computational Linguistics, 2-7 August, suntec, Singapore.

Liu B. 2009. Sentiment Analysis and Opinion Mining. Boca Raton: Morgan & Claypool Publishers.

Liu B. 2011. Web data mining: Exploring hyperlinks, New York: Springer, Lovins, J. B. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistic.

Liu H., Motoda H. 2008. Computational Methods of Feature Selection, Boca Raton: Chapman & Hall.

Manning C.D., Raghavan P., Schtze H. 2008. Introduction to information retrieval. Cambridge University Press.

Maynard D., Funk A. 2011. Automatic detection of political opinions in tweets. In Proc of the 10th International semantic web conference, 23-27 October, Bonn, Germany.

Medhat W., Hassan A., Korashy A. 20140. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal 5, pp. 1093–1113.

Mejova Y., Srinivasan P. 2011. Exploring Feature Definition and Selection for Sentiment Classifiers. In Proc of the Fifth International AAAI Conference on Weblogs and Social Media ICWSM, 17-21 July, Barcelona, Spain.

Monroe W., Green S., Manning C.D. 2014. Word Segmentation of Informal Arabic with Domain Adaptation. In Proc of the 52nd Annual Meeting of the Association for Computational Linguistics, 22-27 June, Baltimore, USA.

Moraes R., Francisco Valiati J., Gavião Neto W.P. 2013. Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications 40, pp. 621-633.

Mountassir A., Benbrahim H., Berrada I. 2013. Sentiment classification on Arabic corpora: A preliminary cross-study. Document Numérique 16(1): 73-96.

Nakagawa T., Inui K., Kurohashi S. 2010. Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables, Human Language Technologies, The 2010 Annual Conference of the North American Chapter of the ACL NAACL HLT, pp 786–794, 1-6 June, Los Angeles, California.

Palmer F. 1986. Mood and Modality. Cambridge University Press.

Pang B., Lee L., Vaithyanathan S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques, In Proc of the Conference on Empirical Methods in Natural Language Processing EMNLP, 6-7 July, Philadelphia, PA, USA.

Pasha A., Al-Badrashiny M., Diab M.T., El-Kholy A., Eskander R., Habash N., Pooleery M., Rambow O., Roth R. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In Proc of the The 9th edition of the Language Resources and Evaluation Conference LREC, 26-31 May, Reykjavik, Iceland.

Rushdi-Saleh M., Martın-Valdivia M., Urena-Lopez L., Perea-Ortega J. 2011. OCA: Opinion corpus for Arabic. Journal of the American Society for Information Science and Technology 62(10).

Saurı R. 2008. A Factuality Profiler for Eventualities in Text. PhD dissertation.

Schapire R. E., Freund Y., Bartlett P., Lee W.S. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics.

Somasundaran S., Namata G., Wiebe J., Getoor L. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In Proc of the conference on Empirical Methods in Natural Language Processing EMNLP, 6-7 August, Suntec, Singapore.

Syarif I, Zaluska E., Prugel-Bennett A., Wills G. 2012. Application of bagging, boosting and stacking to intrusion detection. In Proc of the 8th International Conference on Machine Learning and Data Mining MLDM, 13-20 Jul, Berlin, DE.

Taboada M., Brooke J, Tofiloski M., Voll K., Stede M. 2011. Lexicon-based methods for sentiment analysis. Computational Linguistics 37, pp 267 -307.

Van Asch V. 2013. Macro- and micro-averaged evaluation measures [[basic draft]].

Wang G., Sun J., Ma J., Xue K., Gud J. 2014. Sentiment classification: The contribution of ensemble learning. Decision Support Systems 57, pp. 77–93.

Xia R., Zong C., Li S. 2011. Ensemble of feature sets and classification algorithms for sentiment classification, Information Sciences 181, pp 1138–1152.