UNIVERSIDAD DE JAÉN

ESCUELA POLITÉCNICA SUPERIOR
DE JAÉN
DEPARTAMENTO DE
INFORMÁTICA

TESIS DOCTORAL

SENTIMENT ANALYSIS IN ARABIC: OPINION POLARITY DETECTION

PRESENTADA POR:
MOHAMMED RUSHDI SALEH

DIRIGIDA POR:
DR. D. L. ALFONSO UREÑA LÓPEZ
DRA. DÑA. MARÍA TERESA MARTÍN VALDIVIA

JAÉN, 7 DE OCTUBRE DE 2013

ISBN 978-84-8439-855-4

UNIVERSIDAD DE JAÉN

Sentiment analysis in Arabic: opinion polarity detection

Memoria que presenta

Mohammed Rushdi Saleh

Para optar al Grado de Doctor en Informática

con mención de Doctorado Internacional

Fdo. *Mohammed Rushdi Saleh*

MINISTERIO
DE ASUNTOS EXTERIORES
Y DE COOPERACIÓN

AGENCIA
ESPAÑOLA DE
COOPERACIÓN
INTERNACIONAL

جامعة القدس
AL-Quds University

Tesis Doctoral subvencionada por La Agencia Española de Cooperación Internacional para el Desarrollo (AECID), y parcialmente por Al-Quds University.

D. Luis Alfonso Ureña López y **Dña. María Teresa Martín Valdivia**, Profesores Titulares de Universidad del Departamento de Informática de la Universidad de Jaén

CERTIFICAN

Que la memoria titulada: *"Sentiment analysis in Arabic: opinion polarity detection "* ha sido realizada por **D. Mohammed Rushdi Saleh** bajo su dirección en el Departamento de Informática de la Universidad de Jaén.

Jaén, a 15 de Abril de 2013

Fdo. Luis Alfonso Ureña López	María Teresa Martín Valdivia
Directora de la Tesis	Director de la Tesis

Dedication

To the soul of my mother
To my father, brother, sisters and wife,
For all their endless love, encouragement, prayers and support

Acknowledgment

First and foremost, I must thank the almighty Allah who has again helped me and provided me with the strength to surmount another challenge in my life.

I can't fully express my gratitude to my honorific supervisor Professor Luis Alfonso Ureña López, for his encouragement, guidance and invaluable support and advice. His constructive comments and insightful views have helped me to progress in this research. I've been fortunate to learn from his vast experience. Without his help I could not have finished my dissertation successfully. I am genuinely grateful to Dr. María Teresa Martín Valdiviav for her full support and her careful and wise guidance made this work possible. Our conversations enlightened my way of thinking. You have been a great co-promoter.

During my research, Drs. Arturo Montejo Ráez, Drs. José Manuel Perea Ortega and Drs. Cristóbal José Carmona del Jesus offered me a lot of friendly help; they transferred to me different technical experiences which were necessary in pursuing Ph.D. research. I would therefore like to give my sincere thanks for their generous help.

My mother sadly passed away in March 1993 it is a great pity that she could not see this day. I'm writing these words on the anniversary of her death. I'm very grateful to my mother. Her understanding and her love encouraged me to work hard and to continue pursuing a Ph.D. project. Her firm and kind-hearted personality has helped me to steadfast and never to bend to difficulty. She always let me know that she was proud of me, which was motivated me to work harder and do my best. My thanks and gratitude are also due to my brother Rajeh, he is the best and kind-heart I have ever met. They are also owed to my wife Marwa for her encouragement and patience, without which this work would not have been possible. There is one person I need to mention especially. I'm highly indebted to my uncle, Dr. Khalid Kanan, without whom I would not have reached this peak. He is an inspiration and my role model for how to be an academic. My cousins Waleed and Wesam are also a wonderful companions and friends.

I'd like to convey my heartfelt thanks to Professor Sari Nusseibeh, president of Al-Quds University, without whose support it would not have been possible for me to pursue and to complete this Ph.D. Many thanks go to The AECID as well, for granting me support to finish my thesis.

Last but not least, to all of the staff in my epartment and my collogues at the Universidad de Jaén in both offices N^{o} 154 and N^{o} 103, and my friends Husain and Samer, those happy memories will not be forgotten.

Contents

Chapter 1

Memory

1.1 Introduction

The Internet is growing at a staggering rate as new users all over the world log on, while daily usage of the Internet increases according to the growing demand from as developing nations. Referring to statistics available on the Internet we can find that billions of Internet users everyday upload different types of information onto the Web, such as multimedia (images, music, video, etc.) and text (emails, posts, reviews, articles, etc.). The Internet has become a huge virtual space on which information can be generated and updated in a moment.

One effect of the information overload on the Web is the difficulty of making a decision when searching for specific data. The general types of information that can be retrieved from the Internet are objective and subjective. The objective information comes in a variety of forms as something observable or factual. This form of the data is as close to the truth as we can get. Subjective information refers to opinions, judgments, assumptions, beliefs, suspicions, and rumors. The subjective information varies from person to person and from day to day and this type of information may cause confusion when a person has a choice to make. Many researchers in the field of text mining and information retrieval have witnessed a swell of interest in the automatic identification and extraction of attitudes, sentiments and opinions. The motivation of this interest comes from the desire to provide an automatic system that can analyze information and track the attitudes in governmental, political and commercial domains and on different on-line forums.

The proliferation in the use of the World Wide Web and the rapid growth of e-commerce has increased the number of reviews and opinions on the Web. Processing a huge number of reviews has become a challenging task for the researcher in the field of text mining and information retrieval. This type of information processing belongs to the field of Opinion Mining (OM) or Sentiment Analysis (SA), which is itself a sub-field of Text Mining. Opinion Mining and Sentiment Analysis refer to the same field of study, which is focused on the computational treatment of opinions, sentiments and subjectivity in the text and on classifying them according to their polarity, either positive or negative [PL08].

OM is a discipline that involves several interesting tasks. For example, opinion extraction can be considered to be a specialization of the information extraction task. Its aim is to detect expressions denoting the key components of an opinion within a sentence or document. Another popular OM task focuses on detecting the subjectivity in a document, i.e., whether the document or part of the document is subjective or objective (informative). One of the most widely studied tasks is that of determining the polarity of a document, sentence, or feature (positive or negative) and measuring the degree of the polarity expressed in it.

A common task in opinion mining is to classify an opinionated document a positive or negative opinion. This type of classification is known as document-level sentiment classification, because the whole document is considered as the basic information unit. Sentiment classification at the document-level can be defined as follows: Given a set of opinionated documents D, it determines whether each document $d \in D$ expresses a positive or negative opinion of an object. For example, given a set of document reviews, the system classifies them into two classes of positive or negative reviews.

Typically, opinion mining is performed by statistical approaches at the document level. Such approaches are implemented by many researchers using techniques based on supervised learning, specifically Support Vector Machines (SVM), Naïve Bayes (NB), and Maximum Entropy (ME). For example [Tur02] used Naïve Bayes to classify reviews, using the Pointwise Mutual Information PMI-IR measure to detect the semantic orientation of subjective phrases. Here, a review is considered to be a recommendation if the accumulation of subjective phrases is positive, otherwise it can be considered to have a negative orientation. [PLV02] applied machine learning methods (NB, ME and SVM) to a data set downloaded from the Internet Movie Database (IMDb)[1]. They used 700 negative reviews and 700 positive reviews, to which they applied machine learning algorithms by using unigrams and bigrams of the part of speech features used to classify documents.

This thesis is organized as follows. In Part I, we introduce opinion mining applications and challenges that can be found in this field. We also explain the different objectives and goals of this study, introducing some techniques and related researches to our investigation. Following this, we summarize and discuss the results we obtained from the different experiments we have carried out. In the last section of this part, we have concluded all of our works and results with new motivations for continuing investigating in this line of research. In Part II, we present six publications distributed in three sections illustrating the investigation of our objectives.

- Polarity classification.

[1] http://www.imdb.com

- Arabic polarity detection.
- Resource generation.

1.1.1 Applications and task challenges

1.1.1.1 Opinion mining applications

Before the rise of the different weblogs and web pages concerned with expressing people's opinions, one needed a recommendation or advice in order to buy a car or other products, or even to vote in a local election, which was normally provided by friends or relatives or a professional consultant. Today, however, people daily express their opinions regarding different objects, such as books, movies, products, people, hotels and many others on review sites (Amazon[2], Booking[3], etc.) and different discussion boards. The growing availability of opinionated text has created an enormous amount of valuable information and a large repository of customer comments and reviews. On the other hand, this type of information presents different problems and challenges. Currently, search engines such as Google and others are not capable of retrieving such information for two main reasons:

1. The reviews can be mixture of opinionated and factual information, consequently the search engines are unable to distinguish between them.
2. The difficulty of detecting the polarity of the text (i.e. positive/ negative, thumb up/ thumb down).

The demand for applications and tools to accomplish sentiment classification tasks has attracted the attention of researchers in this area. Hence, sentiment analysis applications have spread to many domains, from consumer products, healthcare and financial services to political elections and social events [Liu12]. Different companies have developed sentiment analysis software, for instance SAP used BusinessObject text analysis to extract sentiments from both Web-based and internal customer feedback sources. SAS Company have also created SAS sentiment analysis, which automatically rates and classifies opinions collected from websites and social media. Other companies have performed their own built-in sentiment tools such as Twitter and Microsoft Dynamics CRM, Microsoft Live/Bing Search and Google Product Search [BgNH$^+$08].
On the other hand, many researchers have developed sentiment analysis applications in their investigations. [YLHA12] mined the online movies reviews in order to predict product sales performance. [Hu12] built a sentiment predictor for real-time Twitter sentiments related to midterm exams. Other applications were created by

[2] http://www.Amazon.com
[3] http://www.booking.com

[OBRS10]. These linked Twitter Sentiment to public opinion polls. Finally, [MY11] have tracked emotions in many types of emails, in order to distinguish between genders by taking into account the emotions expressed in emails. Women used terms from the joy-sadness axis, whereas men used terms from the fear-trust axis.

1.1.1.2 Opinion mining challenges for Arabic and English

Opinion mining has gained importance in recent years due to the wide range of applications and their use, although remains a difficult process to carry out. Therefore, the classification of polarity (positive or negative) is a challenging task. Arabic sentiment analysis faces different challenges. Many of them are common to other languages, while the other difficulties come from the complex derivational morphological system of the Arabic language itself. In fact, another important challenge is the limited number of websites containing Arabic reviews, so available resources for constructing Arabic corpora are scarce. In the following section we enumerate the most important challenges found by the researchers in sentiment analysis classification for different languages including English and Arabic:

- Not all subjective sentences have an opinion. For instance the following sentence:
  I want a laptop with very good specifications. As in Arabic:

  اريد جهازا محمولا بمواصفات جيدة
  Therefore, the previous sentence contains no opinion at all. Hence in classification tasks this sentence may be classified as expressing a positive opinion due to the positive adjective (good).

- Another challenge is that opinions may vary over a period of time; this change may depend on the mood of a given person. [BMdR06] has studied this phenomena. Opinions may also change over a period of time depending on the satisfaction of the person about a service or product.

- The informal language used in reviews provides a challenge for different reasons. For instance, the use of colloquial language, abbreviation or orthographic mistakes. In addition, reviews employ have different styles of writing.

- Another challenge faced in opinion mining is the identification of the strength of an opinion. [ES06] measured the strength of opinion using lexical resources SentiWordNet (SWN), in which each synset is associated with a score of how positive, negative or objective it is.

- Sarcasm and irony present in text make the sentiment analysis task harder. However, people tend to use irony as a way of expressing their opinion in their comments. [Fil12] generated a corpus from Amazon product reviews containing regular and sarcastic comments to improve the performance of sentiment analysis systems. This corpus was used for training sarcasm systems in order to detect opinions on two levels, a document level and sentence level.

- Mixed polarities. In many comments we find a mixture of opinions about different features of the same object. For example, a customer may evaluate the lens of a camera as positive and the battery life of the same camera as negative. Hence, this type of review makes detecting the overall opinion for the same review more challenging.

1.2 Objectives

The main objective of this study is focused on:

1. Sentiment polarity classification.
2. Arabic polarity detection.
3. Creating resources for the Arabic language.

In the following sections we discuss these objectives in more detail.

1.2.1 Sentiment polarity classification

Sentiment analysis includes several sub-tasks but the principle aim is reflect the overall opinion found in a text. A wide range of techniques and tools can be used to tackle the task of sentiment classification. For this reason, we aim to investigate the most promising ones. We select two approaches in order to apply this task of classification. The first one is machine learning and the second one is the semantic orientation approach.

Machine learning offers many algorithms designed for text classification. However, we select the one that fits best with sentiment analysis classification. Hence, we apply the SVM and NB algorithms. Moreover, we have investigated the impact of vector creation with different features (TF, TFIDF and *n-grams*). To determine the overall sentiment polarity on a multi-point scale (using, for example, the number of stars awarded) we have applied linear regression.

Semantic orientation is a basic method that uses a variety of techniques. One of these is a lexicon based approach. These types of lexicons can be constructed either manually or automatically. Several lexicons have been created for use int sentiment analysis task. Our approach aims to rely on the analysis of a textual corpus that correlates with POS features (Adjectives, Nouns, Verbs and Adverbs) with semantic orientation based on a lexicon such as (SWN).

1.2.2 Arabic polarity detection and resources

Sentiment analysis within a multilingual context is a challenging task. One of the main challenges is that statistical approaches require training data, which is normally scarce for languages other than English. Despite the fact that Arabic is one of the most important languages and one of the top 10 languages used on the Internet according to the Internet Word State[4] (see Figure 1.1), there is no reference corpus for sentiments. Henceforth, the priority for us is to create a corpus for the Arabic language in order to apply a statistical classification approach. To this end, we have collected reviews from many webs and blogs about Arabic movies, in spite of their being scarce. To implement a lexical approach for Arabic we require specific linguistic resources. However, generating such resources is very time consuming and requires manual work. So the alternative solution is to translate an Arabic corpus into English and to use the lexical recourse SWN. After doing so, we can apply the polarity classification to both languages (Arabic and English) by using different methodologies, such as the statistical and lexical approaches. In addition to the previous objectives, we want to draw a comparison between different classifiers applied to different corpora in sentiment analysis to enable us to generate another corpus. We collected new data from camera reviews from Amazon in order to determine the overall opinions contained and to analyze several features (corpus domain, corpus size and other factors) and to compare them with other corpora.

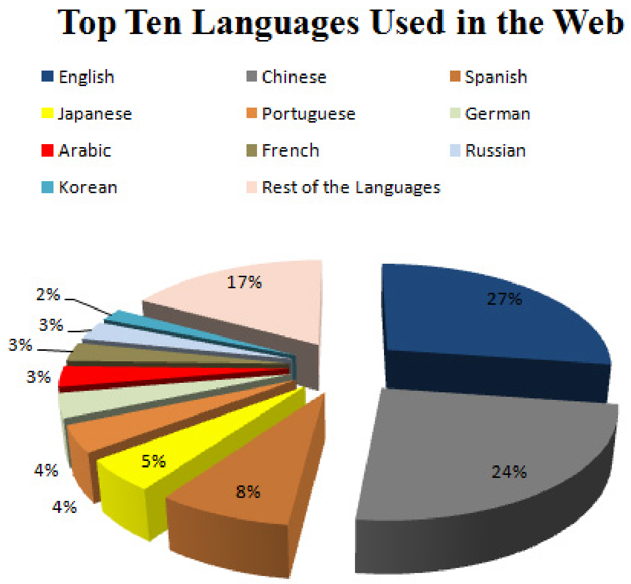


FIGURE 1.1: TOP TEN LANGUAGES IN THE INTERNET 2010-IN MILLIONS OF USERS

[4] http://www.internetworldstats.com

1.3 Related works and background

Opinion analysis has been developed by many researchers in recent years, focusing on two main research directions, i.e., machine learning approaches and the semantic orientation based approach. The target of most research into SA was applied to the English language, but other studies paid attention to other languages. A general task aimed at sentiment analysis research would be to find opinions related to a given object in any web content. In this section, we focus on the most closely related studies to both of the aforementioned approaches and summarize the techniques used to achieve the main task. Furthermore, we discuss different approaches to sentiment analysis applied to languages other than English.

1.3.1 Machine learning based approaches

[PLV02] used a training data to apply machine learning methods (Naïve Bayes, Maximum Entropy and SVM) to determine polarity. The data was collected from the Internet Movie Database (IMDb). They found that the SVM worked better than the other methods. [WA07] used a small hand labeled training corpus for feature classification performed by a supervised approach (Naïve Classier). The new feature was added manually. They tested the system with three products: Electronic dictionaries, MP3 players and Notebook PCs. The results showed that the use of a unigram was the most effective method. [MC04] apply a machine learning approach (SVM) to a movie reviews corpus downloaded from Pitchfork Media. They also used various features including the Combination of Turney value, the three text-wide Osgood values, word unigrams or lemmatized unigrams and extra features based on the movie domain. [PT09] applied SVM with combined methods to classify reviews from different corpora. One of these datasets was the same as that used by [PL04] and it included 1,000 positive and 1,000 negative samples. Several classifiers were used: the General Inquirer Based Classifier (GIBC), the Rule-Based Classifier (RBC), the Statistics Based Classifier (SBC), and SVM. They accomplished a hybrid classification, whereby if one classifier fails to classify a document, the classifier passes the document to the next classifier until the document is correctly classified or no other classifier remains. The results indicated that SBC and SVM improve their effectiveness in the hybrid classification. Another approach has been employed by [YH03], using the NB classifier with opinion-bearing features to distinguish between factual and opinionated documents collected from the Wall Street Journal. The [WBO99] approach also applied NB with other linguistic features (pronouns, adjectives, and adverbs) to classify the documents collected from the Wall Street Journal. [ZY07] developed a system to retrieve the opinions. This system was based on a three-step model. The first step was to retrieve the query-relevant documents, the second was opinion identification and the third was ranking the query-related opinions by calculating similarity scores. A sentence-based SVM model was built in a feature space of unigrams and bigrams which were selected by the Chi-square test. Finally, they classified the document as opinionated if they found at least one subjective sentence in the document. Many researchers tend to utilize the SVM algorithm for polarity detection, due to the robustness of the algorithm and its ability to tackle different types of features. [Che06] used different features in an SVM classifierin order to detect polarity in reviews from blog spots. Different linguistic features such as verbs, adjectives, pronouns and adverbs were used. The results of the classifier were promising considering the noisy nature of data of blog spots. Another pertinent approach has been accomplished by [MC04]. Here, movie review data was classified as positive or negative using several SVM models. A variety of diverse information sources were combined with SVMs to create a so-called *hybrid-SVM* classifier. Many features have been selected and tested, including unigram-style features and real-valued favorability measures (Semantic Orientation with PMI). The authors reported an improvement of 1%-3% accuracy when using their additional features with unigrams over using unigrams alone. Comparing the different machine learning algorithms in sentiment classification, it can be seen that the SVM algorithm is more efficient although it has a drawback in its computing time.

1.3.2 Semantic based approach

By contrast to the, of supervised learning approach which required training data, SO based on the unsupervised learning method does not require prior knowledge or a training data. However, SO approaches depend more on several linguistic features and the orientation of a word, phrase or sentence to accomplish the polarity classification task.

One of the earliest studies to focus on word orientation was carried out by [HM97]. They used adjectives as a good indicator for detecting text orientation. Moreover, at the phrase level, they supposed that adjectives connected with a conjunction such as "and" probably indicates the same semantic orientation, whereas, if two adjectives are connected with a conjunctive word such as "but" they are likely express opposing opinions. By using a log-linear regression and performing a clustering algorithm on all adjectives connected by conjunctions they created a set of negative and positive adjectives.

[Tur02] proposed an unsupervised learning algorithm to detect document sentiment based on selected phrases, where the phrases are chosen if they containing adjectives or adverbs. They then calculated the semantic orientation by using Pointwise Mutual Information (PMI). Finally, a label of "recommended" or "not recommended" is assigned to the reviews based on the average semantic orientation of the phrases. [HL04] proposed another technique for predicting opinion at the level of the sentence in order to summarize the costumers'reviews of a product. First, they identified features by detecting frequent words and then, defined the opinion sentences which contained both a feature and at least one adjective. The prediction of the orientation of an opinion sentence is based on the opinion word in that sentence. Hence, depending on whether most opinion words tend to be positive or negative they determined the orientation of the sentence. In cases where the number of positive and negative opinion words is the same, they take the orientation of the closest opinion sentence.

1.3.3 Sentiment prediction for non English

Some research has focused on detecting semantic orientation for non-English. The tracking of such research is challenging in itself, due to the scarce training data or even to its unavailability in different languages in order that statistical approaches might be applied.
There are two main approaches within a multilingual framework

- Lexicon-based approach.
- Corpus-based approach.

[ADY06] dealt with a multilingual framework (English, Arabic and Chinese) for financial news. First, a method for identifying financial keywords based on statistical criteria in the training data was developed. By looking at the keyword's neighborhood, statistical criteria was established while the researchers also produced a financial sentence to carry the sentiment information. Then, by using these patterns, they built a finite state automaton which is tested by hidden data. For each language in this study two data sets have been used, one for training (specific-domain) while the other was a general data set. They used the British National Corpus (BNC) as a general language for English. By contrast to the English language the general corpus was not available for Arabic, so they have built their own corpus comprising 2.6 million tokens. The recourses available for the Chinese language are fewer than for English but still more than Arabic, and in this case they have used two corpora: the TaBE (Localization for Taiwan and Big5 Encoding) Project, and LDC Chinese.
Several studies have been carried out for other languages. [vALT10] used an annotated corpus with news on the financial market in Croatia in order to apply sentiment analysis. [MVMCPOL13] presented a new method for polarity classification in a Spanish movie reviews corpus, benefiting from the parallel translation of these reviews into English. This method can be applied to different languages that lack lexical resources. The researchers have proposed a meta-classifier which combined three models generated by supervised and unsupervised learning methods. First, they created two models by applying a machine learning algorithm to the Spanish corpus and its parallel translated corpus. Then, they generated a third model for the translated English corpus using SWN. Finally, they have integrated different features of the two supervised models and of the third unsupervised model into a meta-classifier. The results were very promising using the combination techniques. [Den08] translated a German review into

English using machine translation software, and then predicted the polarity of the translated text by using three classifiers: LingPipe, SWN with classification rule, and SWN with machine learning. [ZZL$^+$09] applied sentiment classification to Chinese reviews. [GJ11] classified French movie reviews by using machine learning and SWN to assess their polarity.

1.3.4 Corpora for Arabic SA

In order to implement Arabic SA it very important to find corpora to train and test the systems. Hence, a number of researchers have paid attention to creating different corpora for this purpose. In this section we will describe some of these corpora.

[AMD12] developed a multi-genre corpus which includes documents written in two styles (Arabic-dialect and Modern Standard Arabic). This corpus was called AWATIF, in which the researchers have crawled documents from three different resources:

- The Pen Arabic Treebank. The documents were about sport, politics, finance, etc..
- Wikipedia talk pages. They have extracted 30 talk pages from different domains.
- Web forums. The data was selected from 7 web forums.

The corpus was annotated using two different procedures. First, simple annotation by an untrained annotator tagged the sentences with a positive, negative and neutral labels. The second procedure was undertaken by annotators with a linguistics background to label the sentences.

[MI12] generated a corpus of Arabic religious decrees, this domain was chosen for two reasons:

1. Religious decrees contain subjective text.
2. The text used in this domain was written in Modern Standard Arabic not in Arabic-dialect.

The data was collected from 5 Islamic sites, 77,047 decrees were downloaded. First, the researchers have carried out a simple preprocessing of the text. Then they have manually labeled the polarity of the data according two categories Halal (Allowed) and Haram (Prohibited). Also, the data was split into a question and answer in order to mine for opinions within answers.

[EAF12] released a new corpus for opinion holder extraction in the Arabic language. The researchers crawled 150 MB Arabic news articles to which they applied different preprocessing tasks including:

- Sentence segmentation.
- Morphological analysis.
- Part of speech tagging.
- Semantic analysis.
- Name entities recognition for ten classes (Person, Location, Organization, Job, Device, Car, Cell Phone, Currency, Date, and Time).
- Subjectivity analysis. They have classified the sentences from strongly to weakly subjective by manually translating the MPQA lexicon [WWH05].
- Manual annotation of 1 MB of the opinion holder corpus by three different annotators, where conflicts of tagging are resolved by using the majority voting principle.

The researchers used the Research and Development International (RDI)[5] toolkit to handle the orthographic and morphological analysis of Arabic sentences, part of speech (POS) tagging, and semantic analysis of new words. The Arabic MPQA subjective lexicon & Arabic opinion holder corpus is freely available at the Arabic Language Technology Center "ALTEC" [6].

1.4 Data sets

1.4.1 SINAI corpus

We have created the SINAI corpus[7] by crawling reviews from the Amazon website. These reviews were about cameras with different brands and series. The Amazon website is uses a 5 stars rating system. We have downloaded 1,943 documents labeled with a different number of stars (see Table 1.1). We then removed all HTML tags, in order to have a plain text while maintaining some attributes (Review title, author name, location, date, and the rating for each review). Following that we distributed the reviews to five folders from 1-5 according to the number of stars for each review (see Table 1.2).

[5] http://www.rdi-eg.com

[6] http://altec-center.org

[7] http://sinai.ujaen.es/?cat=18

TABLA 1.1: NUMBER OF REVIEWS PER PRODUCT IN THE SINAI CORPUS

Camera	Reviews
CanonA590IS	400
CanonA630	300
CanonSD1100IS	426
KodakCx7430	64
KodakV1003	95
KodakZ740	155
Nikon5700	119
Olympus1030SW	168
PentaxK10D	126
PentaxK200D	90
Total	1,942

TABLA 1.2: DISTRIBUTION OF THE REVIEWS OF THE SINAI CORPUS ACCORDING TO THE NUMBER OF STARS

Number of (*)	Number of Reviews
*	78
**	67
***	96
****	411
*****	1,290
Total	1,942

1.4.2 OCA corpus

We have collected 500 reviews, in which 250 reviews were labeled as positive and the other 250 were labeled as negative. Table 1.3 shows some statistics about this corpus which we have called OCA[8]: Opinion Corpus for Arabic. In fact, this corpus underwent different preprocessing steps in order for it to be used in our experiments. First, we have removed HTML tags and non-related characters, then manually corrected spelling mistakes and replaced the Romanization of Arabic letters with their counterpart in Arabic. Next, we have carried out for each review in the corpus different processes including tokenizing, removing Arabic

[8] http://sinai.ujaen.es/?cat=18

TABLA 1.3: STATISTICS ON THE OPINION FOR ARABIC

	Negative	Positive
Total documents	250	25
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

stop words, stemming and filtering those tokens whose length was less than two characters. Another process distinguished the reviews as positive or negative according to the rating system for each blog. For instance, some blogs had a 5 stars rating system while the other blogs used a rating scale from 10 points.

1.4.3 EVOCA corpus

We have translated the OCA corpus to English using an online translator[9] thereby obtaining a parallel corpus which we have called EVOCA[10]: the English version of OCA. The EVOCA corpus contains the same number of reviews as OCA. Table 1.4 shows some statistics for the EVOCA. We have tackled different procedures to obtain EVOCA. First, we split the text of the reviews into segments of 500 characters to fit with online translator requirements. Next, we removed a UTF-8 and invalid characters that had appeared in the text after translation. Finally, we have joined the translated segments of the text for each review.

TABLA 1.4: STATISTICS ON THE EVOCA OPINION CORPUS

	Negative	Positive
Total documents	250	25
Total tokens	122,135	153,581
Avg. tokens per review	488.54	614.32
Total sentences	5,030	3,483
Avg. sentences per review	20.12	13.93

[9] http://translation2.paralink.com/

[10] http://sinai.ujaen.es/?cat=18

1.5 Discussion and results

In this section we provide a brief discussion of the results obtained from our work. Each sub-section discusses the experiments carried out in each paper.

1.5.1 Polarity classification

1.5.1.1 Prediction of Customer Ratings on a New Corpus for Opinion Mining

The main goal of this study was to examine the strength of sentiments and to summarize the overall opinions of a customer review of a specific product. We have used the SINAI corpus to apply different experiments. The techniques used in the experiments were based on linear regression. The deviation of the regression on the real rating was within 1 and the root mean squared error obtained by this experiment was 0.638. We found that the regression model performed better on rating with a high number of sample reviews in the corpus than on those with a reduced sample. There were fewer lower rating samples in our corpus (from 1-3 stars) than samples of a higher rating (4-5 stars), due to the availability of reviews downloaded from Amazon for the selected products. Consequently, the standard deviation was higher on low rating (1-3 stars) than higher ratings (4-5 stars). We obtained promising results from our learned regression model that fit well with the real ratings, the correlation measurement was 0.802 and the relative error strict was 0.137.

1.5.1.2 Learning to Classify Neutral Examples from Positive and Negative Opinions

Data are the backbone of data mining; however, for all data mining and text mining tasks, especially in our framework opinion mining task, the biggest challenge is preparing data suitable for modeling. This study focused on how to prepare the data of different corpora in order to learn our model for classification polarity. One of the most controversial discussions among many researchers in the field of opinion mining is how to manipulate the neutral examples of data reviews. There are several varieties of rating systems on the web; for instance, Amazon uses from 1 to 5 stars while IMDb uses a numerical system from 1 to 10 points. Usually, the comments rated as above 3 stars in 5 scales will be counted as positive and the comments rated as fewer than 3 stars will be used as negative but the comments with 3 stars will be treated as *neutral*. Many researchers have suggested neglecting neutral reviews, while others have assumed that adding them to the positive examples will improve the classification. In this study we have investigated how to deal with examples of neutral reviews. We have carried out several experiments using the SVM algorithm to establish the advantages of taking neutral examples into account. We have mainly run our experiments on three corpora (SFU Review Corpus[11], SINAI Corpus and SINAI-B Corpus). The SFU corpus was collected from Epinions.com. This web site uses the 5-stars rating system. In addition it also uses "recommended" and "non-recommended". So, each review has two rating systems (number of stars and "recommended" or "non-recommended"). We have used the original documents downloaded by the researchers in order to benefit from the number of stars system tags. We have received 371 documents (some files were missing from the source) 182 of them are negative according to the "non-recommended rating system" while the other 189 are positive according to the "non-recommended rating system". We have found that 20 out of 371 of these reviews were rated as 3-stars. 14 of them were tagged as "non-recommended" and the other 6 were tagged as "recommended". We have created the SINAI corpus by crawling the documents from Amazon. As mentioned above, this web site only uses the 5-stars rating system. We have downloaded 1942 reviews: 145 reviews were labeled as 1 or 2 stars (negative), 1701 labeled as 4 or 5 stars (positive), 96 reviews labeled as 3 stars (neutral). We noticed that the corpus was unbalanced between the negative and positive reviews, so we generated a new corpus called SINAI-B by selecting only 200 positive sample reviews from the original SINAI corpus. The main goal of this study was to advantageous to use the neutral examples on the training data or to neglect them. Accordingly, we have run different experiments using the SVM algorithm. The first process in our experiments was to create different subsets of the three above corpora (SINAI, SINAI-B and SFU) according to the distributions of neutral samples. We have distributed the neutral samples in the following way:

- *SFU corpus*

1. Neglect neutral examples. This corpus only includes the positive samples rated with 4 and 5 stars and the negative ones rated with 1 and 2 stars.

2. Neutral reviews as negative. Here we have considered the neutral reviews as negative by adding 3 star reviews to 1 and 2 star reviews.

3. Neutral reviews as positive. Following the example of many other researchers we have included the neutral examples (3 stars) as positive by adding them to 4 and 5 star reviews.

4. Recommended and non-recommended review system. We have distributed the neutral examples according to the second option of the rating systems, so we have added the 3-star reviews rated as "recommended" to the positive samples and the "non-recommended" examples from among the 3 star reviews to the negative samples.

[11] http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html

- *SINAI and SINAI-B*

We have distributed the neutral reviews for SINAI and SINAI-B in the same way as above, except for the experiment taking into account the "recommended" and "non-recommended" ratings. Unfortunately, the reviews hosted on most opinion forums do not include the recommended and non-recommended information, and only the number of stars for each review is supplied. This is the case for the SINAI corpus. For this reason it was necessary to develop a method to decide the polarity of 3 star reviews. We thus generated a model using the training data from the SINAI corpus excluding the 3 star reviews and only using those with 1, 2, 4 and 5 stars. We obtained a preliminary classifier that we used to classify the 3 star reviews. With the new classifier we added the new positive 3 star reviews to the 4 and 5 stars set and the new negative 3 star reviews to the 1 and 2 stars negative set. The same procedure was followed for the SINAI-B corpus. We then distributed the neutral examples in the same way. We also generated another data set for the SFU corpus according to our model for the classification of neutral examples. Finally, after generating all of the different combinations of these corpora, we have classified them according to the new generation. The results obtained were very promising and reinforced our hypothesis of how neutral examples play a vital role in opinion mining classification. In all cases we have obtained better results when we use neutral examples in our model, except in one case in which the neutral samples were manually classified in the SFU corpus.

1.5.1.3 Experiments with SVM to classify opinions in different domains

The main purpose of this research was to test different domains of data sets and to examine several features including weighting schemes. We have used three corpora in this study. Two of them have already been used by other researchers (Pang Corpus and SFU Corpus) while the third was the SINAI Corpus. The SINAI corpus contains 1943 reviews of different camera brands and series. These reviews were labeled with different numbers of stars from 1 to 5. We have split the corpus into positive and negative documents according to the number of stars. Documents rated with 1 and 2 stars were considered as negative while the 3, 4 and 5 star reviews were selected as positive. We also have examined how a weighting scheme can influence the system classification by using the unigram, bigram and trigram weighting scheme. Figure 1.2 shows the general design of the experiments. First we have generated *n-gram* models : unigram, bigram and trigram to check their impact on the classification, while for generation word vectors we have used three different approaches: Term occurrence (TO), Binary occurrence (BO) and word frequency in the document and in the entire corpus (TF-IDF). Finally we have carried out 27 experiments, to analyze the three corpora with different combinations of *n-gram* models and the three types of word vectors (TF-IDF, TO and BO). To evaluate our system we have used 3-folds and 10-folds cross validation

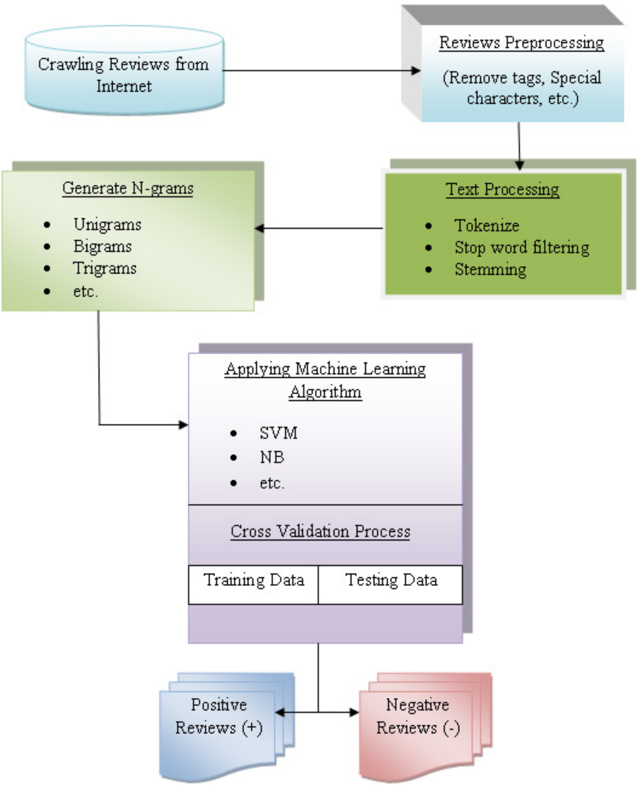


FIGURE 1.2: GENERAL DESIGN OF MACHINE LEARNING EXPERIMENT

for each corpus. The best results were obtained over all corpora by using the 10-folds cross validation. In addition, in most of the experiments using the 10-fold cross validation the accuracy obtained by using the trigram scheme dominated the other schemes (unigram and bigram), or in some cases was equivalent to bigram. In regard to the weighting schemes, TF-IDF and BO performed better than TO among the three corpora and using all the *n-gram* techniques. Comparing the results we achieved for the three corpora to the results of other researchers obtained by the Pang corpus and SFU corpus, we can confirm that our system performs well. For instance, for the Pang corpus the accuracy we achieved was 85.35% by applying SVM with trigram and BO using the 10-folds cross validation, while for the 3-folds cross validation with the same combination (trigram and BO) we obtained 84.90%. However, the accuracy obtained by Pang was 82.90% when SVM was applied with unigram, BO and the 3-folds cross validation. For the SFU corpus we achieved an accuracy of 73.25% using trigram, TF-IDF and the 10-cross validation, while the best overall accuracy obtained by SFU was 65%. Even though their methodology was based on adjectives, they have assigned weight for adjectives according to their position in the text. The experiments carried out with our corpus achieved the best result (accuracy 91.51% with TFIDF, bigrams and 10-fold cross validation).

1.5.2 Arabic polarity detection

1.5.2.1 OCA: Opinion Corpus for Arabic

The classification of Arabic reviews according to their polarity is a challenging task. Few studies have been carried out in this area due to many factors:

- Despite the importance of the Arabic language on the Internet, there are nevertheless only a few web pages that specialize in Arabic reviews.
- The available reviews for the Arabic language are limited to movie reviews.
- Web pages of Arabic reviews are not structured in a systematic way. For instance, reviews may be written in languages other than Arabic, may use a Romanization of the Arabic language, the blogs may not specialize only in text reviews but also other, non-related topics (i.e. videos, music, etc.).

The main contribution of this work was to introduce a new Arabic corpus for predicting sentiment polarity. We have therefore created the OCA corpus which consists of 500 movie reviews (Described in the Data set section). Moreover, we have applied different experiments of polarity detection over OCA using machine learning algorithms. Specifically, we have used the SVM and NB algorithms. Both of these are the most popular and powerful algorithms when applied to text classifications. In addition we used Rapid Miner software which is a good environment for machine learning and data mining processes. We have also evaluated the performance of the two algorithms (SVM and NB) using the 10-folds cross-validation. Moreover, we have applied different n-gram schemes (unigram, bigram and trigram) in order to analyze their effect on Arabic opinion mining classification. Finally we have examined the effect of weighting schemes (Term-Frequency "TF" and Term Frequency-Inverse Document Term Frequency "TF-IDF"). From the results obtained by running 24 experiments in the OCA corpus we can make the following points:

1. SVM slightly improves on the performance of NB.
2. The trigram and bigram models overcome the unigram model.
3. Applying SVM with TF and using stemming the results were identical, the same behavior was observed when we used TF-IDF but without applying stemming.
4. In the case of TF-IDF It was better not to stem the words for both NB and SVM.
5. Finally, when using TF with SVM it is recommended to use the stemmer, although when using TF with NB it is not recommended.

We have obtained promising results compared with other researchers in the same context but concerned with other languages.

1.5.2.2 Bilingual Experiments with an Arabic-English Corpus for Opinion Mining

In order to compare the experiments of polarity detection for Arabic and English we have presented a new corpus called EVOCA (Described in the Data set section). EVOCA is a parallel translated version of the OCA corpus which consists of the same number of reviews (500 in each corpus).
We have applied several experiments to both corpora (OCA and EVOCA) using machine learning algorithms (NB and SVM) to analyze the differences in classification caused by the translation process. For each experiment we have also studied the effect of using the following parameters: Stemmer, unigrams, bigrams and trigrams. The TFIDF was used in all experiments as a weighting scheme. To evaluate the classifier we have used the 10-fold cross-validation. From the results obtained from the different experiments according to the F1 measure we have observed the following:

1. SVM performed better than NB.

2. Applying stemming to the OCA corpus had a negative impact.

3. For the EVOCA corpus, the use of stemmer with SVM improved the results, while in NB experiments the results were worse.

4. The best results were obtained when we used SVM for the OCA corpus without using the stemmer the F1 was 0.9073. For the EVOCA corpus with the stemmer the best result was 0.8840.

5. Finally, the difference between SVM and NB over the OCA corpus was small. But when they applied (SVM and NB) to the EVOCA corpus, NB lost a good deal precision due to the translation process.

1.5.2.3 Comparing Machine Learning and Semantic Orientation for Polarity Detection using EVOCA

This study aimed to improve the sentiment classification system by using SO approaches. In addition to SO approaches we also applied several experiments that used machine learning techniques with different combinations of parameters. We also focused on linguistics features such as adjectives, nouns, verbs and adverbs by extracting their scores from a lexical resource; specifically we have used SWN version 3.0. Finally, the data sets used in this research was the EVOCA corpus.

In the following sections we explain how our experiments were designed using the two techniques (ML and SO), and we analyze the results achieved by the experiments.

- *Machine learning approach*

The first approach undertaken accomplished in this study was machine learning. We applied SVM and NB. For SVM we have run different types of kernels *(linear, polynomial, rbf, sigmoid or pre-computed)*, while the other parameters for each kernel of the SVM remained as default. We have also studied the behavior of different heuristics such as filtering the stop words, the use of stemming, filtering those tokens with less than four characters and the use of unigrams, bigrams and trigrams. The TF and TFIDF models were used to generate the learning vectors. After applying the combination of the former features and procedures we executed a wide range of experiments totaling 240 experiments for SVM. As the results produced were very similar we only selected the best solution and results of the 48 experiments by applying a linear kernel. For NB we have run 48 experiments as it does not have a particular configuration, although we followed the same structure as with the SVM experiments (filtering, stemming, *n-grams*).

- *Semantic orientation approach*

The second approach in our plan of this study was SO, for this we utilized different procedures. The first step was to apply the POS tagger (using the Tree Tagger tool) to extract all nouns, adjectives, verbs and adverbs from each review. The second step of preprocessing was to generate different types of the data sets based on linguistic features. Thus, we created 15 sub-corpora from EVOCA depending on the combination of the four features (nouns, adjectives, verbs and adverbs) to test the effect of each feature. Finally, we used the lexical resource SWN for each document in the sub-corpora to determine the triple-triple (positive, negative and objective). Henceforth, each review whose polarity score was larger than or equal to the negativity score was classified as positive, otherwise it was considered to be negative. From the results we obtained we can summarize the difference between the different techniques used according to the F1 measure:

1. A significant difference among the best F1 results of both approaches (ML and SO) was about 35%.

2. SVM performed better than NB, the difference was 11.78% better than NB.

3. The TFIDF weighting scheme overcomes TF (+2.74% for SVM and +5.2% for NB).

4. The application of filtering and stemming was not recommended in our case study.

5. The bigram scheme gets the best result from among the *n-grams* used in our experiments.

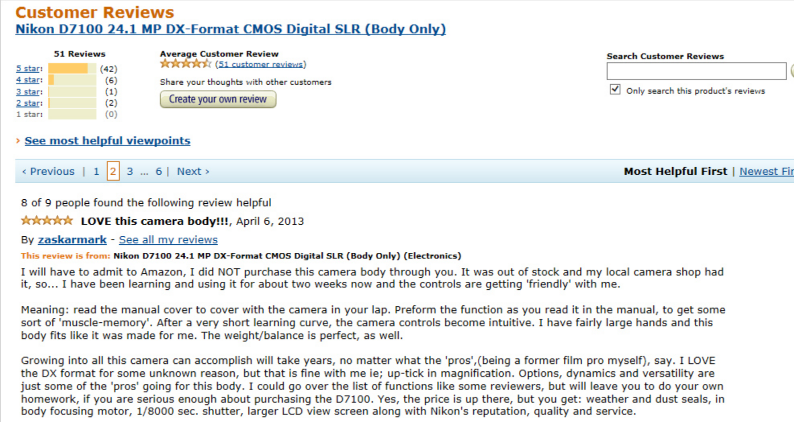
6. The results of the second approach (SO) showed that the "nouns, and adjectives" sub-corpus achieved the highest F1 score of 0.6698. The second highest score achieved was by the combination of adjectives and verbs.

1.5.3 Resource generation

In this research we have generated three new corpora in order to implement sentiment classification. Below we explain the implementation of each corpus in our investigation:

1.5.3.1 Prediction of Customer Ratings on a New Corpus for Opinion Mining

In order to predict customer ratings we have generated a new corpus which we have called the SINAI Corpus, this corpus includes the reviews of ten brands of digital cameras which have been crawled from Amazon. For the Arabic language, we have found various reviews on Amazon related to different products which have structured well (see Figur 1.3). We divided the reviews into five folders according to the rating system used by Amazon (from 1-5 stars) to fit with our experiments, each folder contains the reviews that match its rating. We then split the corpus so that 90% of the data set would be used for training and 10% for testing. Finally we applied the linear regression algorithm to predict the overall ratings.


Customer Reviews
Nikon D7100 24.1 MP DX-Format CMOS Digital SLR (Body Only)
51 Reviews
Average Customer Review
(51 customer reviews)
Share your thoughts with other customers
Create your own review
Search Customer Reviews
☑ Only search this product's reviews
› See most helpful viewpoints
‹ Previous | 1 2 3 ... 6 | Next ›
Most Helpful First | Newest Fir

5 star:	(42)
4 star:	(6)
3 star:	(1)
2 star:	(2)
1 star:	(0)

8 of 9 people found the following review helpful

★★★★★ **LOVE this camera body!!!**, April 6, 2013

By zaskarmark - See all my reviews

This review is from: **Nikon D7100 24.1 MP DX-Format CMOS Digital SLR (Body Only) (Electronics)**

I will have to admit to Amazon, I did NOT purchase this camera body through you. It was out of stock and my local camera shop had it, so... I have been learning and using it for about two weeks now and the controls are getting 'friendly' with me.

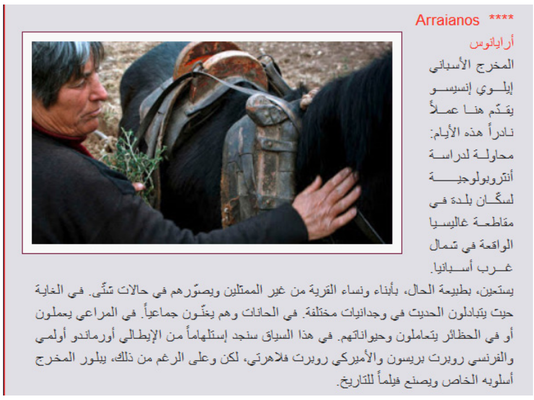
Meaning: read the manual cover to cover with the camera in your lap. Preform the function as you read it in the manual, to get some sort of 'muscle-memory'. After a very short learning curve, the camera controls become intuitive. I have fairly large hands and this body fits like it was made for me. The weight/balance is perfect, as well.

Growing into all this camera can accomplish will take years, no matter what the 'pros',(being a former film pro myself), say. I LOVE the DX format for some unknown reason, but that is fine with me ie; up-tick in magnification. Options, dynamics and versatility are just some of the 'pros' going for this body. I could go over the list of functions like some reviewers, but will leave you to do your own homework, if you are serious enough about purchasing the D7100. Yes, the price is up there, but you get: weather and dust seals, in body focusing motor, 1/8000 sec. shutter, larger LCD view screen along with Nikon's reputation, quality and service.

FIGURE 1.3: SCREEN SHOT OF A REVIEW FROM AMAZON

1.5.3.2 OCA: Opinion Corpus for Arabic

This study focused on opinion mining for the Arabic language. First, we have searched for a corpora with which to carry out different experiments related to Arabic polarity detection. Unfortunately, there were few resources for the Arabic language, which was the main reason we chose to build our corpus for the Arabic language from scratch. We built this corpus by collecting movie reviews in Arabic from different Web pages and blogs. Figure 1.4 shows a screen shot of a movie review from the Filereader blog spot[12]. We have crawled 500 reviews to create our corpus, which we called OCA. Although the Arabic language is one of the most important languages in the world, there are few studies focused on sentiment analysis for this language.



Arraianos ****

أرايانوس

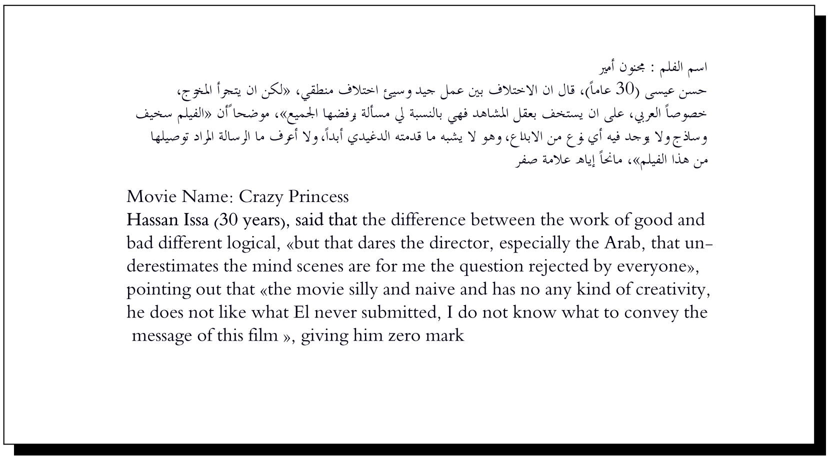
المخرج الأسباني
إيلوي إنسيسو
يقدّم هنا عملاً
نادراً هذه الأيام:
محاولة لدراسة
أنثروبولوجية
لسكّان بلدة في
مقاطعة غاليسيا
الواقعة في شمال
غرب أسبانيا.
يستعين، بطبيعة الحال، بأبناء ونساء القرية من غير الممثلين ويصوّرهم في حالات شتّى. في الغابة
حيث يتبادلون الحديث في وجدانيات مختلفة. في الحانات وهم يغنّون جماعياً. في المراعي يعملون
أو في الحظائر يتعاملون وحيواناتهم. في هذا السياق سنجد إستلهاماً من الإيطالي أورماندو أولمي
والفرنسي روبرت بريسون والأميركي روبرت فلاهرتي، لكن وعلى الرغم من ذلك، يبلور المخرج
أسلوبه الخاص ويصنع فيلماً للتاريخ.

FIGURE 1.4: AN EXAMPLE OF MOVIE REVIEWS IN ARABIC

1.5.3.3 Bilingual Experiments with an Arabic-English Corpus for Opinion Mining

In this study we have used two corpora. The first one is the OCA corpus which was presented above, while the second one was EVOCA, the English version of OCA. We have translated the OCA corpus into the English language using an online translator. Figure 1.5 shows a review in Arabic with its counterpart in English. Although the translated text does not fully give the same meaning as in the Arabic original, it still conveys the most interesting aspect of the review. Following this we carried out experiments for both corpora (OCA and EVOCA) using different machine learning algorithms (SVM and NB) to classify the texts according to their polarity in order to have a comparison of the classification. The EVOCA corpus consisted of 500 reviews of which 250 were positive and 250 were negative. The same applied for the OCA reviews. OCA and EVOCA are valuable resources for applying sentiment classification for Arabic.

[12] http://filmreader.blogspot.com



اسم الفلم : مجنون أمير
حسن عيسى (30 عاماً)، قال ان الاختلاف بين عمل جيد وسيئ اختلاف منطقي، «لكن ان يتجرأ المخرج،
خصوصاً العربي، على ان يستخف بعقل المشاهد فهي بالنسبة لي مسألة يرفضها الجميع»، موضحاً أن «الفيلم سخيف
وساذج ولا يوجد فيه أي نوع من الابداع، وهو لا يشبه ما قدمته الدغيدي أبداً، ولا أعرف ما الرسالة المراد توصيلها
من هذا الفيلم»، مانحاً إياه علامة صفر

Movie Name: Crazy Princess
Hassan Issa (30 years), said that the difference between the work of good and bad different logical, «but that dares the director, especially the Arab, that underestimates the mind scenes are for me the question rejected by everyone», pointing out that «the movie silly and naive and has no any kind of creativity, he does not like what El never submitted, I do not know what to convey the message of this film », giving him zero mark

FIGURE 1.5: SAMPLE OF TRANSLATED TEXT OF EVOCA WITH ITS COUNTERPART IN ARABIC

1.6 Conclusion

In this section we give a brief conclusion regarding our three major goals: polarity classification, Arabic polarity classification and resources.

1.6.1 Polarity classification

1.6.1.1 Prediction of Customer Ratings on a New Corpus for Opinion Mining

The problem of the increasing number of reviews that are available on the Internet without ratings, particularly those found on weblogs, encouraged us to investigate how we can come up with a prediction of customer ratings. From the above studies, we have established the following:

- The possibility of summarizing the opinions of customers in a given value closely correlated to what the customer has in mind as a ratings.
- We can implement our model in other environments where reviews are not rated.
- It is possible to summarize a whole list of comments using descriptive analysis, the average, standard deviation and other measurements from the distribution of predicted ratings.

1.6.1.2 Learning to Classify Neutral Examples from Positive and Negative Opinions

This study attempted to confirm the importance of neutral examples and how they can affect sentiment analysis classification. We can summarize that, by using the neutral samples in the training data according to a classification method as proposed in this study, will improve the classification system.

1.6.1.3 Experiments with SVM to classify opinions in different domains

Here we aimed to make a comparison of different techniques and features in sentiment classification systems. We have compared different data sets (three corpora) and classified them according to the SVM learning algorithm with different weighting systems (TF-IDF, TO and BO) and several *n-gram* techniques (unigrams, bigrams and trigrams). From the results obtained we can highlight the following:

- The sentiment classification systems are domain dependent. For example, our corpus was based on camera brands and performed best in classification, while the Pang corpus was based on movie reviews and performed second in the classification while the SFU corpus, which contains movies, music, hotels, etc. was third.
- Corpus size. We noticed that the SFU corpus was relatively small compared with the SINAI and Pang corpora.
- Learning algorithms. We confirmed that machine learning algorithms, and in particular SVM, are promising tools for sentiment classification when used in conjunction with the TF-IDF and trigrams techniques.

1.6.2 Arabic polarity classification

1.6.2.1 OCA: Opinion Corpus for Arabic

The main goal of this work was to apply Arabic sentiment classification. However, due to the difficulty of finding datasets designed for this purpose we have generated a new corpus to achieve our goal. From the different experiments we carried out we may make the following observations:

- Predicting polarity for the Arabic language is a challenging task compared with other languages such as English. The results obtained were very promising.

- A detailed study is necessary to analyze the different resources such as the "Arabic stemmer" and their effect on the classification.

1.6.2.2 Bilingual Experiments with an Arabic-English Corpus for Opinion Mining

This study investigated the classification of opinion mining using the bilingual corpora OCA and EVOCA. The effect of translation and stemmers on the classification systems was also investigated. Based on the results we achieved we can draw the following conclusions:

- Comparing the results obtained using both corpora, SVM always overcomes the NB algorithm.
- Using a stemmer for the Arabic language was not recommended in this study.

1.6.2.3 Comparing Machine Learning and Semantic Orientation for Polarity Detection using EVOCA

From both of the approaches (machine learning and semantic orientation) implemented in this article we make the following observations:

- The SVM reinforced its potency when we classified texts into their polarities and it outperformed NB.
- Machine learning algorithms are better suited to sentiment classification tasks than lexical resources such as SWN.
- The experiments carried out in this study proved that using lexical resources is a good alternative to polarity detection when we do not have an available labeled corpus.

1.6.3 Resource Generation

The three corpora we have generated (SINAI, OCA and EVOCA) were domain dependent. However, each corpus was used for a specific goal. The SINAI corpus was our contribution to testing different opinion mining systems. This corpus contains reviews related to digital cameras. We have carried out different machine learning experiments on the SINAI corpus and other corpora (SFU and Pang corpus). From the results obtained we have noticed that the corpus size and domain have influenced the system performance.

We have introduced another two important corpora to the community of opinion mining, the OCA and EVOCA corpora. The OCA corpus consists of 500 documents of Arabic movie reviews. This corpus is a valuable resource and one of the newest corpora for applying Arabic sentiment classification. The EVOCA corpus is a parallel English version of OCA, we have translated the OCA corpus using online translation in order to compare sentiment classification for the two languages. We have applied several experiments using machine learning algorithms (NB and SVM) on both corpora. The results obtained show that the loss of precision in the EVOCA corpus due to the translation process is very slight. We have therefore concluded that, due to the lack of Arabic lexical resources, it is advisable to use English resources such as SWN for opinion mining.

1.7 Future works

Finally, we present in this section our plans for further work arising from this study:

- To implement other models different to linear regression in order to identify better algorithms for rating prediction. Also, to improve our model using linguistic features in order that it might also predict product components.

- To improve the OCA corpus. We need to enlarge our data and to annotate the corpus in order to improve the classification process using some linguistics features.

- To improve the EVOCA corpus by using some resources available for the English language such as SWN. Moreover, we can test different stemmers available for the Arabic language.

- To test a hybrid technique using Machine learning and SO approaches.

Chapter 2

Publications

2.1 Polarity classification

2.1.1 Prediction of Customer Ratings on a New Corpus for Opinion Mining

- Mohammed Rushdi-Saleh, Arturo Montejo Ráez, María Teresa Martín-Valdivia, Luis Alfonso Ureña López. Prediction of Customer Ratings on a New Corpus for Opinion Mining. Proceedings Working Notes for the WOMSA 2009 Workshop. Sevilla.
    - Status: Published
    - Quality measures: The WOSMA Workshop is one of the first conferences related to OM
    - Point of interest/comments: We develop a new corpus of product reviews that is freely available for the SA research

Prediction of Customer Ratings on a New Corpus for Opinion Mining

M. Saleh, A. Montejo-Ráez, M. T. Martín-Valdivia, L. A. Ureña-López

Universidad de Jaén, Department of Computer Science, Spain

Abstract. In this paper a new corpus for opinion mining is introduced. It has been generated from Amazon customer reviews on several products. Details about its generation along with a complete description of the corpus are given. Besides, a linear regression has been applied in order to study how sort comments behave as textual information for the prediction of customer rates. Our experiments show that these texts are quite informative and that the rate is an interesting measurement on the overall opinion of the customer on the product. This technique could help to summarize opinions in other web sites where rate is not explicitly given by the user.

1 Introduction

The number of blogs in the World Wide Web has been increased over several years. In these Weblogs people can estimate a publication, such as music, movies, video games, books, or electronic products. In addition, the author may assign rating to indicate its relative merit. Different types of reviews can be found on the net: On the one hand, "consumer reviews" are written by the owner of a product or the user of a service who has experience to comment whether or not the product or service deliver on its promises. On the other hand, some reviews can be written by an expert in that field who tested several products and can identify which offers are the best according to their features and their cost. This type of reviews refers to "Expert Reviews". Opinions in these Weblogs identify the author's viewpoint about the subject rather than simply recognize the subject itself. The opinion mining in such as Weblogs gives another magnitude to search and summarization tools. The year 2001 marked the beginning of widespread of the research problems and opportunities that sentiment analysis and opinion mining raise. Both of them denote the same field of study, which itself can be considered a sub-area of subjectivity analysis [1]. Sentiment Analysis (SA) is a discipline that deals with the quantitative and qualitative analysis of text for determining opinion properties [2]. The term sentiment analysis stands for a broad area of natural language processing, computational linguistics and text mining. It aims to extract attributes and components of the object that have been commented on a document [3]. With rapid expansion of the Web and online merchants, more people buy products on the Web. In order to enhance customer satisfaction, it becomes common for customers to submit and express opinions on the products that they buy. Some products get hundreds of reviews which makes difficult read them to decide which product to choose. From this point of view an automatic mining opinion system which is able to capture the general perspective and summarize customer's viewpoints become valuable research area. Sentiment analysis classification has several characteristics [4], including various tasks, features, techniques, and application domains. SA tasks can be categorized as:

1. Classes: Positive/Negative or Objective/Subjective text.
2. Text Level: document or sentence/phrase.
3. Source/target of sentiment, if it is known or extracted.

There are four features that have been used in opinion mining domain, syntactic, semantic, link based, and stylistic. An important phase of opinion mining domain is to create a corpus of useful features that enables to categorize the textual reviews into sub-categories to identify the opinion sentences. Existing approaches often try to identify words, phrases and patterns that indicate the sentiment sentences [5]. However, the context of these patterns are playing an important rule as they can convey an ambiguity and needs more sophisticated semantic techniques. The techniques used for sentiment classification are divided into:

- Machine learning.
- Link analysis.
- Similarity score (Phrase pattern matching, frequency counts, etc.).

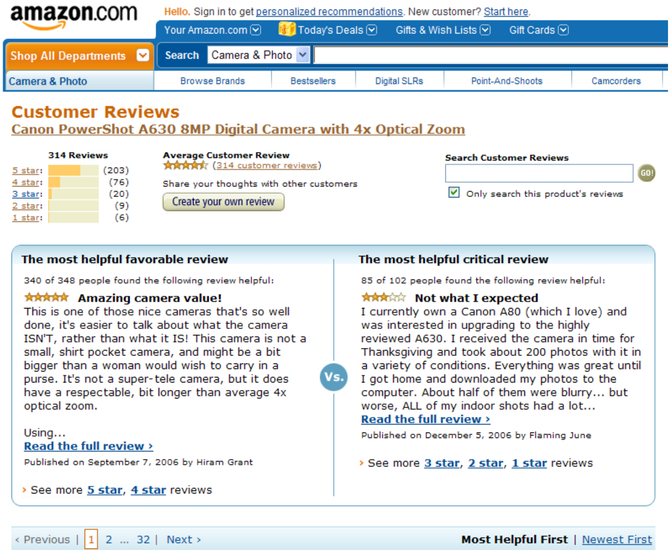
Machine learning approaches are divided into supervised and unsupervised, supervised tend to be more accurate than unsupervised approaches as needed training corpus, but no data training required for unsupervised techniques with weaker result. The last category in SA classification is the domain (for example, products, movie, music, etc.). On the other hand, manually annotated corpus is expensive to create and time consuming, and needs to be changed for different domains. It becomes more desired to use unsupervised learning machine techniques to be applied on different information systems extraction. Hu and Liu [6] [7], studied the problem of feature-based opinion summarization of customer reviews. The task was performed in two steps: firstly, identifying opinion features and ranking the features according to their frequencies, and secondly, they specified positive and negative sentences in the customer reviews.The experiment was applied on five products (two digital cameras , one cellular phone, one MP3 player and one DVD player). The first 100 reviews were piled up from amazon.com, and used in order to predict feature-based summaries. Then their system was applied to extract the products features and evaluated the discovered feature manually. Wang and Araki[8], used small hand labeled training corpus for feature classification performed by a supervised approach (a Naïve Classifier). The new feature was added manually. They tested the system with three products: Electronic dictionaries, MP3 players and Notebook PCs. Another approach done by Dave [9], trained a classifier using self-tagged reviews from major web sites, and refined the classifier using the same corpus before applied it to different broad web searches. In this paper, we give a brief summary about opinion mining approaches, and then we introduce a new corpus for opinion mining. It was generated from Amazon customer reviews on several products of a digital cameras. Details about its generation along with a complete description of the corpus will be explained later on in the following section. Then, we introduce the method (Linear regression) which has been applied, and discuss how the comments behave as textual information for the prediction of customer rates. We strongly believe this technique can help in summarizing opinions in other web sites where rate is not explicitly given by the user. Finally, we present the conclusions and further works.

2 Sentiment Analysis Approaches

The first who handle the task of opinion classification were Hatazivassiloglou et. al. 1997 [9]. To predict the semantic orientation of conjoined adjectives, they applied a Log-Linear regression model to differentiate whether the conjoined adjectives belong or not to the same orientation. Then a clustering algorithm was performed to separate the adjectives into two classes of different orientation. Finally each class was compared with the highest frequency to be labeled as positive orientation. Turney et al. 2003 [10] introduce a method for inferring the semantic orientation from association. The relation between a word and a set of positive or negative words was measured using two different statistical measures Pointwise Mutual Information(PMI) and Latent Semantic Analysis(LSA) which shows better result than PMI. Kamp et al., 2004 [11] focused on adjectives as a good clue in opinion classification. They used WordNet to define the semantic distance between the adjectives of a text and set of prepared tagged words. They relied on three major factors which can explain the variance in judgment, these factors are the evaluative factor(good/bad); the potency factor (strong/weak); and the activity factor(active/passive). They defined three functions for the three factores to measure the relative distance of a word to the two reference words(good/bad for evaluative function)then divided the difference by the distance between the two reference words, the value ranging in the interval[-1,1].-1 for words on the 'bad' side of the lexicon, 1 for words on the 'good' side of the lexicon. Pang et al., 2002 [12] applied three machine learning methods Naïve Bayes, Maximum Entropy, and Support Vector Machines SVM) to determine whether a movie reviews are positive or negative. In the results they concluded that the SVM works better than the other methods. Also they proved that the presence or absence of a word is more indicative of the content than the frequency of a word. (Hu, 2004) produces summary with positive and negative opinions product review features. First, he identifies features by detecting the frequent words and then identify opinion sentence and their orientation. The sentence which contains one feature and at least one adjective is defines as opinion sentence. He checks the adjectives in the review with a seed list of 30 adjectives, if they belonged to this list or a synonym or an antonym. The seed was expanded every time the orientation of an adjective is found. Ding et al., 2007 [13] built a system called ”opinion observer”. This system used some linguistic rules integrated with new opinion aggregation function. They computed the opinion score in a sentence to each word taking into account the distance between the features and the opinion word. The low weight to opinion words indicates that words are far away from that feature. Esuli et al., 2005 [14] deal with a new method based on the assumption that terms with similar orientation tend to have similar glosses. The method relied on the application of semi supervised learning to the task of classifying terms as belonging to positive or negative.

3 Amazon Corpus on Digital Cameras

Our corpus was split into five folders (1, 2, 3, 4, and 5) each folder contains the reviews depending on the number of stars (i.e. folder 1 contains all reviews rating by one star etc.). We selected one type of electronics products available in amazon.com that has enough data reviews with different brands Fig.(1). We decided to select the ten digital cameras that have more comments Table 1 and with rich text per each review (see Table 1). We downloaded all reviews from Amazon, then we cleaned all HTML tags, after that we generate XML files per each product that contains all reviews. Also we include the title of each review,



Customer Reviews

Canon PowerShot A630 8MP Digital Camera with 4x Optical Zoom

314 Reviews	
5 star:	(203)
4 star:	(76)
3 star:	(20)
2 star:	(9)
1 star:	(6)

Average Customer Review
(314 customer reviews)

Share your thoughts with other customers

Search Customer Reviews

☑ Only search this product's reviews

The most helpful favorable review

340 of 348 people found the following review helpful:

★★★★★ Amazing camera value!

This is one of those nice cameras that's so well done, it's easier to talk about what the camera ISN'T, rather than what it IS! This camera is not a small, shirt pocket camera, and might be a bit bigger than a woman would wish to carry in a purse. It's not a super-tele camera, but it does have a respectable, bit longer than average 4x optical zoom.

Using...

Read the full review ›

Published on September 7, 2006 by Hiram Grant

› See more 5 star, 4 star reviews

The most helpful critical review

85 of 102 people found the following review helpful:

★★★☆☆ Not what I expected

I currently own a Canon A80 (which I love) and was interested in upgrading to the highly reviewed A630. I received the camera in time for Thanksgiving and took about 200 photos with it in a variety of conditions. Everything was great until I got home and downloaded my photos to the computer. About half of them were blurry... but worse, ALL of my indoor shots had a lot...

Read the full review ›

Published on December 5, 2006 by Flaming June

› See more 3 star, 2 star, 1 star reviews

‹ Previous | 1 2 ... 32 | Next ›   Most Helpful First | Newest First

Fig. 1. Screen shoot of reviews captured from Amazon.com

author name, location, date, and rates (number of stars that were selected). Scalar ratings will be the focus on our experiment for detection the overall opinion for the products. Table 1 shows the number of reviews and the sentences per each product studied. We have selected ten different digital cameras from five different brands (Canon, Kodak, Nikon, Olympus and Pentax). We have a total of 1,943 reviews with 22,202 sentences. Each XML file was split into sentences, one sentence per each file, and then we parsed all the sentences using NLProcessor parser to have another XML file per a sentence:

```
    <?xml version=’1.0’?>
<ALL><TEXT> <P><S>([ The_DT  inside_JJ  shots_NNS
])((were_VBD))for_IN most_JJS of_IN  ([ them_PRP  ])([ grainy_NN
]),_,  on_IN([various_JJ settings_NNS]) </S></P> </TEXT></ALL>
```

Table 1. Table1 Product related statistics

Camera	Reviews	Sentences
CanonA590IS	400	4200
CanonA630	300	2945
CanonSD1100IS	426	4334
KodakCx7430	64	790
KodakV1003	95	886
KodakZ740	155	1620
Nikon5700	119	1740
Olympus1030SW	168	2654
PentaxK10D	126	1886
PentaxK200D	90	1147

4 Rate Prediction

Amazone.com and other merchant sites facilitate the evaluation process for customer’s reviews by adding stars scalar either thump up and thump down in order to estimate the products under discussion. In our corpus that introduced above we availed from this evaluation (stars rating). This rate can be seen as a summarization of the opinion of the customer. To predict this rate from the review text, leads to sentiment analysis of human natural language.

4.1 Linear Regression

In linear regression, we predict scores on one variable (dependent variable or criterion) from the values on other variables (independent variables or predictors) by defining a linear combination of independent variables. The presented corpus has been used for experimenting on this subject, learning a linear regression model from word vectors to predict customer rates. The following equation is the general form of a linear regression formula :

$$f(x_1, x_2, ..., x_n) = a_1x_1 + a_2x_2 + \ldots + a_nx_n \tag{1}$$

Where $\mathrm{x_i}$ is word weights and $\mathrm{a_j}$ are the values to be calculated in the learning process. After this, from a new review, it is possible to compute the rate once the vector model is generated. For this statistical analysis, the RapidMiner [1] tool was selected. This learning algorithm applies the Akaike Information Criterion (AIC) for selecting the best fitted model [15] . Also, attribute selection is performed applying the M5 method, which steps through the attributes removing the one with the smallest standardised coefficient until no improvement is observed in the estimate of the error given by the Akaike information criterion. In our setup, the default value of 1.0e-8 was set for the ridge parameter.

5 Generation of Word Vectors from Customer Reviews

As pointed before, each review has to be transformed into a vector of word weights. For this,the TF.IDF weighting scheme has been applied along with the following processing steps:

1. Stop words removal and stemming has been applied on words.
2. Document frequency is used as first filter to remove to rare terms (those appearing in less than 4 reviews).
3. Gain Ratio has been computed for each term to retain just those 1,000 terms with the highest value.

Gain ratio, compared to other well know information-oriented measures, compensates the problem of the information gain value which tends to prefer attributes with a large number of possible values. This prior filtering on the number of attributes lower the dimensionality from more than 4,000 to 1,000, speeding up the process of regression learning. The gain ratio has been based on each rate as a separate class. Thus, features more informative but not totally exclusive to a rate obtain higher gain ratio values.

6 Experiments and Results

The corpus was split into 90% of the data set for training, and the evaluation was performed on 10% of the data. The performance values obtained are detailed in Table 2.

As can be seen in Fig. (2), there are far more comments which rate high a product than those with low rates. It can be related to the fact that Amazon.com is a shop, and that consequently high rates tend to be at the beginning for

[1] http://rapid-i.com

Table 2. Performance measurements

Root mean squared error	0.63872
Relative error strict	0.1378
Correlation	0.802

marketing goals. Also, it is important to note that in general, the deviation of the regression on the real rate is within 1 (in fact, the root mean squared error reported a value of 0.638). This behavior can be explained by the fact that users not always agree on the number of stars even if they agree with the review on the product. As expected, for those rates with large number of samples, the regression model is better fitted than for lower rates, which had far less number of samples in the corpus. This also could explain the reason why standard deviation is higher on lower rates (from 1 to 3 stars) than in higher ones (4 and 5 stars). This effect is also visible in Fig.(3), but it is better explained from Fig.(2). In this figure we can see, for each rate, the behavior of predicted values. The parallelism between minimal, maximal and average values of predicted values, and its slope close to 1, shows a good fit of the regression model. As pointed out, the standard deviation (stddev in the graph) decreases for higher rates. In general, the learned regression model fits well on the real rates (correlation of 0.802 and relative error strict of just 0.137).

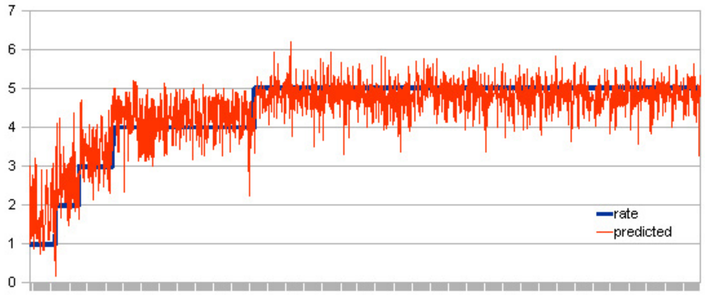


Fig. 2. Graphical view of regression values and real rates

7 Conclusions and Further Work

Our main conclusion is that it is possible to summarize the opinion of a customer in a given value, closely correlated to what the customer has in mind as ”rate”. This is an important result, as it enables us to apply similar models in other

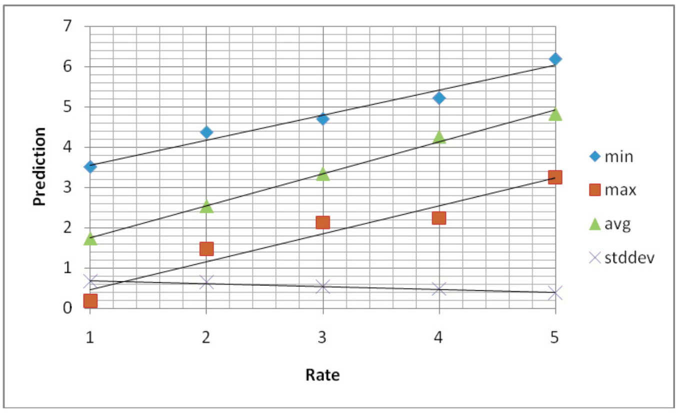


Fig. 3. Statistical analysis of predictions

environments (i.e. blogs) where comments are not rated. In this way, we could summarize a whole list of comments and study, using descriptive analysis, the average, standard deviation and other measurements from the distribution of predicted rates. As further work we plan to replicate our experiments using 10-fold cross validation to explore the effects of a learned model on non seen data and to provide statistical significant measurements on that setup. Also, other models, different to linear regression, have to be studied in order to identify possible better algorithms for rate prediction. In our opinion, deeper linguistic analysis has to be performed, as word vectors may not be best candidates as features before model learning. Actually, we are working on product features detection, so terms like optical lens or battery life would be considered as relevant attributes.

Acknowledgments

This work is partially funded by project RFC/PP208/UJA-08-16-14 granted by University of Jaén, project FFIXP06-TU2301-2007/000024, granted by Andalusian Government, and project TIMOM TIN2006-15265-C06-03, granted by Spanish Government. Also another part of this project was funded by Agencia Española de Cooperación Internacional para el Desarrollo MAEC- AECID.

References

1. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1-2) (2008) 1–135

2. Esuli, A., Sebastiani, F.: Pageranking wordnet synsets: An application to opinion mining. In: Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics. (2007) 424–431
3. Liu, B.: Opinion mining. In: Encyclopedia of Database Systems. (2004)
4. Abbasi, A., Chen, H., Salem, A.: Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. **26**(3) (2008) 1–34
5. Boiy, E., Hens, P., Deschacht, K., Moens, M.F.: Automatic analysis in on-line text. In: proceedings of ELPUB conference on Electronic Publishing. Vienna, Austria. (2007)
6. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: AAAI. (2004a) 755–760
7. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. (2004b) 168–177
8. Wang, G., Araki, K.: Oms-j: An opinion mining system for japanese weblog reviews using a combination of supervised and unsupervised approaches. In: HLT-NAACL (Demonstrations). (2007) 19–20
9. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. (1997) 174–181
10. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems **21**(4) (2003) 315–346
11. Kamps, J., Marx, M., Mokken, R., de Rijke, M.: Using wordnet to measure semantic orientations of adjectives. In: National Institute for. (2004) 1115–1118
12. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing. (2002) 79–86
13. Ding, X., Liu, B.: The utility of linguistic rules in opinion mining. In: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. (2007) 811–812
14. Esuli, A., Sebastiani, F.: Determining the semantic orientation of terms through gloss classification. In: CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management. (2005a) 617–624
15. Witten, I., Frank, E.: Data Mining: Practical machine learning tools and techniques. Volume 2. Morgan Kuffman Publishers (2005)

2.1.2 Learning to Classify Neutral Examples from Positive and Negative Opinions

- María Teresa Martín-Valdivia, Arturo Montejo Ráez, Luis Alfonso Ureña López, Mohammed Rushdi-Saleh: Learning to Classify Neutral Examples from Positive and Negative Opinions. Journal of Universal Computer Science 18(16): 2319-2333, 2012.

  - Status: Published

  - Impact Factor: 0.669

  - Category: (COMPUTER SCIENCE, SOFTWARE ENGINEERING) Ranking: 2011: 85/104; (COMPUTER SCIENCE, THEORY & METHODS) Ranking: 2011: 84/99.

  - We focus on an important problem in polarity detection task: the use of neutral examples

Journal of Universal Computer Science, vol. 18, no. 16 (2012), 2319-2333
submitted: 9/6/11, accepted: 27/8/12, appeared: 28/8/12 © J.UCS

Learning to Classify Neutral Examples from Positive and Negative Opinions

María-Teresa Martín-Valdivia
(Department of Computer Science. University of Jaén, Spain
maite@ujaen.es)

Arturo Montejo-Ráez
(Department of Computer Science. University of Jaén, Spain
amontejo@ujaen.es)

Alfonso Ureña-López
(Department of Computer Science. University of Jaén, Spain
laurena@ujaen.es)

Mohammed Rushdi Saleh
(Department of Computer Science. University of Jaén, Spain
msaleh@ujaen.es)

Abstract: Sentiment analysis is a challenging research area due to the rapid increase of subjective texts populating the web. There are several studies which focus on classifying opinions into positive or negative. Corpora are usually labeled with a star-rating scale. However, most of the studies neglect to consider neutral examples. In this paper we study the effect of using neutral sample reviews found in an opinion corpus in order to improve a sentiment polarity classification system. We have performed different experiments using several machine learning algorithms in order to demonstrate the advantage of taking the neutral examples into account. In addition we propose a model to divide neutral samples into positive and negative ones, in order to incorporate this information into the construction of the final opinion polarity classification system. Moreover, we have generated a corpus from Amazon in order to prove the convenience of the system. The results obtained are very promising and encourage us to continue researching along this line and consider neutral examples as relevant information in opinion mining tasks.

Keywords: Opinion Mining, Sentiment Polarity, Neutral Examples, NLP
Categories: I.2.7, I.7, I.2.1, H.3.3, L.3.2

1 Introduction

Recently the interest in Sentiment Analysis (SA) and Opinion Mining (OM), has grown significantly due to various different factors [Liu, 2010]. The rapid evolution of the World Wide Web has changed our view of the Internet. It has turned into a collaborative framework where technological and social trends come together, resulting in the over exploited term Web 2.0. In addition, the tremendous use of e-commerce services has been accompanied by an increase in freely available online reviews and opinions about products and services. A customer who wants to buy a product usually searches for information on the Internet trying to find other consumer analyses. In fact, web sites such as Amazon, Epinions or IMDb (Internet Movie Database), can greatly affect customer´s decisions. Moreover, opinion mining is useful not only for the individual customer but also for any company or institution as a powerful tool for understanding customer preferences. However, the huge amount of information available makes it necessary to develop new methods and strategies.

SA is becoming one of the main research areas in Natural Language Processing (NLP) and Text Mining (TM). This new discipline attempts to identify and analyze opinions and emotions [Tsytsarau and Palpanas, 2012]. It includes several subtasks such as subjectivity detection [Wiebe et al., 2001], polarity classification [Pang et al., 2002], review summarization [Somprasertsri and Lalitrojwong, 2010], humor detection [Mihalcea and Strapparava, 2006], emotion classification [Strapparava and Mihalcea, 2008] and so on. Specifically, this paper focuses on sentiment polarity classification.

Different approaches have been applied in the field of sentiment polarity classification, but there are two main trends: In the symbolic approach, which applies manually crafted rules and lexicons, the document is represented as a collection of words. Then the sentiment of each word can be determined by different methods, for example, using a web search [Hatzivassiloglou and Wiebe, 2000] or consulting lexical resources such as WordNet [Kamps et al., 2004]. The other approach relies on machine learning techniques to tackle the classification of reviews according to their orientation (positive or negative). In this approach, the document is represented by different features and a machine learning algorithm is applied. These features may include the use of n-grams or defined grammatical roles like adjectives, for instance. Commonly used machine learning algorithms include Support Vector Machines (SVM), Maximum Entropy (ME) or Naïve Bayes (NB) [Pang et al., 2002].

This paper focuses on a particular issue regarding the opinion polarity at document level: the use of neutral examples in order to classify the review as positive or negative. We train a classifier using a corpus labeled with a numerical rating for each opinion. In the first step we only use the positive and negative reviews to train the system. With this model we classify the neutral examples into positive or negative samples and then include them in the corpus in order to train the final classifier.

We use different machine learning algorithms in order to classify the polarity of reviews. Specifically we use Support Vector Machine, Logistic Regression and K Nearest Neighbors. We focus on how neutral opinions can be included in order to improve the classification of sentiment polarity. We tested different combinations of neutral examples with the positive and negative sets, and even without using any neutral review. Furthermore, we developed a method for classifying the neutral examples into positive or negative reviews. In our experiments we used different corpora labeled according to the rating of each review. The paper is organized as follows: Section 2 briefly describes previous related work on sentiment polarity classification and discusses how neutral samples can affect this challenging task. In Section 3 the data sets used in our experiments are described. We then explain the methodology used and describe the three machine learning algorithms applied in our experiments, along with the experimental framework developed. Section 5 presents the experiments carried out and discusses the main results obtained. Finally, we outline conclusions and further work.

2 Use of neutral examples in sentiment polarity

Nowadays, sentiment polarity is one of the main tasks in opinion mining. Given a subjective text, a sentiment polarity classifier must determine whether the opinion is positive or negative. In the scenario of commercial product reviews, it would be interpreted as if the customer likes (positive) or dislikes (negative) a given product overall. The opinions can be ranked into a specific ranking between 1 and 5 stars or between 1 and 10. Moreover, sentiment polarity classification can be studied at document, sentence or feature level. Document level polarity classification attempts to classify the general sentiments into reviews, news, or articles [Wiebe et al., 2001; Pang et al., 2002; Mullen and Collier, 2004], while sentence-level polarity classification tries to determine the sentiment for each sentence [Yi et al., 2003; Pang and Lee, 2005], and feature level tries to find different sentiments within one sentence [Wilson et al., 2005]. Some systems classify the opinions detected using different scales [Pang and Lee, 2008]. In some cases, the sole purpose is to identify opinions in a text and classify them into positive, negative or neutral classes. In other cases, the goal is to assign different ratings such as very bad, bad, satisfactory, good, very good, or excellent.

There are a variety of rating systems in the web and blogs which include opinions and reviews of products and services. The simplest one solely includes a binary classification of the reviews (positive or negative, thumbs up or thumbs down). Other sites use a star-based rating or numerical system (1 to 5 stars for example in Amazon, or 1 to 10 points in the IMDb).

There are different ways to treat the neutral examples in the corpus. For example, in a 5-star rating system, some studies neglected the neutral examples in the corpus. Thus, the reviews rated with 1 and 2 stars were classified as negative while 4 and 5 were labeled as positive [Turney, 2002; Pang et al., 2002; Dave et al., 2003; Yu and Hatzivassiloglou, 2003]. In this case, reviews labeled with 3 stars (i.e., neutral examples) are not included in the learning process. The information supplied by the 3 star opinions is simply disregarded. However, there are some papers showing how the use of neutral examples can help to improve the classification [Pang and Lee, 2005]. For example, [Koppel and Schler, 2006] suggest that the polarity problem might be best handled as a three-class problem with positive, negative and neutral classes. Moreover, they conclude that the use of neutral training examples in learning facilitates better distinction between positive and negative opinions.

In addition to rating systems, some web sites include other useful information about the reviewed item such as recommended and non-recommended products (for example, Epinions). Usually, 1 and 2 star reviews are labeled as non-recommended and 4 and 5 stars are labeled as recommended. However, for the 3 star reviews we can find opinions that sometimes are labeled as recommended and other reviews as non-recommended. In this type of corpus, this additional information classifying opinions as positive or negative can avoid the noise introduced by the 3 star reviews. Unfortunately, this kind of corpus is not common and usually it is necessary to decide what to do with the 3 star samples. This is a very difficult problem even for human users who must decide the polarity of neutral examples because some of them tend to be positive while others have a negative orientation. In Figure 1 and Figure 2 we can see two 3 star reviews from the Amazon site. We have underlined the positive sentences, and the negative text is in bold. The first review tends to be positive and the second one seems to be negative. However, this is only the user's subjective appraisal. Therefore, in this work we will study the effect of using neutral examples to train a classifier using a machine learning approach. Our proposal is to incorporate the information supplied by neutral samples in order to train a classifier and improve a sentiment polarity system.

> I bought this camera while i was pregnant because i Fig.d i would need a good one for when the baby came. I was really pleased with it and it did take really nice photos. **The videos werent the best** but i suppose you cant expect perfect videos from a cheap camera. When the baby came i had someone running around the delivery room snapping pictures. **Every other picture was blurry.** I dont know if it was the operator or just the camera. **I did notice that you had to wait a long time and have the perfect light for the camera to take really good pics.** After having the camera for about 2 or 3 months i had an accident involving dirty baby clothes a misplaced camera and a washing machine...needles to say the camera didnt make it out alive. I decided to go ahead and buy the same camera again. I was still pleased with it but a little bummed i couldnt find the 10mp for as cheap so instead i had to settle for the 8.2. **Anyways im an avid review reader and i had read a couple that said the camera straight up quit working after 6 months.** I decided to ignore them because most of the other comments were totally positive. I had my second camera for about 5 months and it died...on its own...no washing machine involved. **It was like the Auto Focus just completely quit working for some reason.** (really bad timing too because i was at the hospital with my friend while she was having HER baby when i found out it quit working.) Anyway i liked the camera but cant decide if i want to try it out a

Figure 1: Example of a 3 star review with positive orientation

> This is a nice camera if you're looking primarily for a camera that is small, rugged, and waterproof**. However, if you're looking for a camera that takes great pictures - keep looking. The image quality is terrible so forget the 10.1 mega pixel feature. And since the 3.6 optical zoom is hardly enough to zoom in on far away objects the poor image quality becomes a big deal. When you crop a photo in an attempt to "zoom" digitally, you can see terrible pixilation, grain, and blur. I considered sending the camera back to Amazon, but decided to keep it for taking photos in the water. If I didn't want that feature I would have definitely returned it for something else.**

Figure 2: Example of a 3 star review with negative orientation

3 Corpora description

In this paper we have used different corpora. Firstly, we performed several experiments with the Taboada corpus in order to demonstrate that the correct use of neutral examples can improve the sentiment polarity classification system. Then we trained a classifier using the 3 star samples in our SINAI corpus, demonstrating the advantages of taking the neutral examples into account. We briefly describe the two corpora in the next subsections.

#Stars	#Reviews
1	80
2	88
3	20
4	51
5	132
Total	**371**

Table 1: Review in the Taboada corpus according to the number of stars

3.1 Taboada corpus

This collection was used by [Taboada and Grieve, 2004] and by [Taboada et al., 2006] with the main goal of classifying text automatically based on subjective content. They applied a standard method for calculating semantic orientation by extracting the adjectives. This method is based on [Turney, 2002] where the combinations of adjective + noun and noun + noun were used. The corpus includes 400 opinions collected from the website Epinions.com divided into 200 reviews classified as “recommended” (positive) and 200 as “non-recommended” (negative). The texts contain opinions about products and services like movies, books, cars, cookware, phones, hotels, music and computers. The total number of categories is eight and the corpus contains 25 positive and 25 negative reviews for each category.

Although the reviews in the Epinions website use a 5-star rating system, the available Taboada corpus only includes opinions labeled with “recommended” and “non-recommended” tags, and the reviews are not rated with the number of stars. For this reason we asked the Taboada research group to supply us with the original corpus that they had crawled from the Internet in order to work with a star rating system. Hence we received 371 files because some files were missing from the source. Table 1 shows the distribution of reviews in the Taboada corpus according to the number of stars.

In this corpus all the 1 and 2 star reviews are also labeled as “non-recommended”, while 4 and 5 star opinions are tagged as “recommended”. As regards the 20 reviews with 3 stars, 14 of them are tagged as “non-recommended” and 6 “recommended”. So the whole collection includes 182 (168+14) negative samples and 189 (183+6) positive reviews.

3.2 SINAI corpus

Unfortunately most of the opinion corpora published do not include the labels “recommended” and “non-recommended”, so it is necessary to decide what to do with the neutral examples. Many authors simply neglect the 3 star reviews and only work with clearly positive and negative samples in the corpora. However, some studies show that the correct use of neutral examples significantly improves the polarity classification systems, as commented previously [Koppel and Schler, 2006]. Thus it is very interesting to study the best way to include the 3 star examples in our systems. For this reason we generated our own corpus, called SINAI, by crawling the Amazon website, and it is freely available for the scientific community through http://sinai.ujaen.es/wiki/index.php/SINAISaCorpus. SINAI stands for the name of our research group “Sistemas INteligentes de Acceso a la Información” (Intelligent Systems for Information Access). In order to build the corpus we extracted opinions about cameras of different brands and series. A total of 1,942 documents were labeled with different numbers of stars. Table 2 shows the distribution of reviews per camera model.

Camera Model	**#Reviews**
CanonA590IS	400
CanonA630	300
CanonSD1100IS	426
KodakCx7430	64
KodakV1003	95
KodakZ740	155
Nikon5700	119
Olympus1030SW	167
PentaxK10D	126
PentaxK200D	90
Total	**1,942**

Table 2: Number of reviews per product in the SINAI corpus

#Stars	**#Reviews**
1	78
2	67
3	96
4	411
5	1,290
Total	**1,942**

Table 3: Reviews in the SINAI corpus according to the number of stars

The opinions in Amazon are rated using a 5-star scale, but they do not include additional information about recommended or non-recommended items. Table 3 shows the distribution of reviews according to the number of stars for the SINAI corpus. In the same way as in the Taboada corpus we used 1 and 2 stars as negative samples and 4 and 5 stars as positive reviews. However, the 3-star reviews must be treated in a different way.

The original SINAI corpus is also extremely unbalanced, with the number of positive reviews (rated with 4 and 5 stars) clearly higher than the number of negative reviews (rated with 1 and 2 stars). So we randomly chose 200 positive examples from the total of positive reviews. The new corpus also contains the 145 negative reviews and the 96 neutral examples. This corpus has been called SINAI-B (SINAI Balanced corpus) and was built with the sole purpose of testing the effect of neutral examples on a balanced corpus that does not include the “recommended” and “not recommended” information for each review.

4 Methodology

In this section we describe the framework followed in our experiments, mainly based on the training of different classifiers in order to determine the polarity of reviews in an opinion corpus. Specifically, we applied the Support Vector Machine (SVM), Logistic Regression (LR) and K-Nearest Neighbor (KNN).

4.1 Machine Learning Algorithms

The SVM algorithm [Vapnik, 1995] has been applied successfully in many text classification tasks due to these features [Joachims, 1998]: first, it is robust in high dimensional spaces; second, any feature is relevant; third, it works well when there is a sparse set of samples; finally, most text categorization problems are linearly separable. In addition, SVM has achieved good results in opinion mining, and this algorithm has surpassed other machine learning techniques [O’Keefe and Koprinska, 2009].

Logistic Regression (LR) is a mathematical modeling approach in which the best-fitting, yet least-restrictive model is desired to describe the relationship between several independent explanatory variables and a dependent dichotomous response variable. Some studies have been successful applying this model in the area of sentiment analysis [Martínez-Cámara et al., 2011].

K-Nearest Neighbors (KNN) is a case-based learning method, which keeps all the training data for classification. KNN is very simple; for each new item to be classified KNN seeks the k closest items in the training set, and then it returns the major class in the “neighbors” set. KNN has been used in other opinion mining studies, obtaining good results [Tan and Zhang, 2008].

4.2 Experimental framework

We have used the Rapid Miner software with its text mining plug-in which contains different tools designed to assist in the preparation of text documents for mining tasks (tokenization, stop word removal and stemming, among others). Rapid Miner is an environment for machine learning and data mining processes that is freely available from htpp://rapid-i.com.

As regards the document model, we used the Vector Space Model (VSM) in order to generate the bag of words for each document. The English Porter stemming algorithm was applied in order to reduce words to their common root or stem. We also removed some tokens using a stop word list. However, we did preserve some useful sentiment information such as “ok” and “not”.

For SVM, we implemented our experiments using the libsvm learner by [Chang and Lin, 2001], which is integrated into Rapid Miner as one of the available functions. In our experiments we applied a Linear SVM with the default configuration set by the tool (C-SVC type, RBF kernel, epsilon equal to 0.001 and shrinking heuristics enabled). For LR we used the kernel type Anova available in Rapid Miner with the default values for the other parameters. Finally, for KNN we used the Euclidean distance (1-NN) because it is the configuration with the best results.

4.3 Experiments

Our experiments were run on the Taboada corpus and SINAI corpus. They are different in domain and size. The Taboada corpus contains eight categories with different domains, while the SINAI corpus includes nine different models of cameras (thus, only one domain). In order to train the classifier the corpus is divided into positive and negative samples. For both corpora we considered reviews with 1 and 2 stars as negative samples and reviews with 4 and 5 stars as positive ones. However, for the 3 star reviews we performed different partitions, and thus several training corpora were generated:

- N12P45: the 3 star reviews were ignored
- N123P45: the 3 star examples are considered as negative reviews
- N12P345: the 3 star examples are considered as positive reviews

In addition, the Taboada corpus includes information about recommended and non-recommended items, so we can use this important information to train the classifier. Thus, we included the 3 stars labeled “non-recommended” in the negative set and the 3 stars tagged with “recommended” in the positive samples (N12NR3P45R3). Unfortunately, reviews expressed in most of the opinion forums do not include the recommended and non-recommended information, and only the number of stars for each review is supplied. This is the case with the SINAI corpus. In this situation, it is necessary to develop a method to decide about the polarity of 3 star reviews. Thus we generated a model using the training data from the SINAI corpus but excluding the 3 star reviews and only using 1, 2, 4 and 5 star opinions. We obtained a preliminary classifier C1 which we used to classify the 3 star examples. We used this new classification and we added the new positive 3 star reviews to the 4 and 5 star set and the new negative 3 star opinions to the 1 and 2 star negative set. This new corpus including the neutral examples was used to generate a completely new classifier C2. Figure 3 shows the process followed to generate our classifier. The experiments performed following this strategy have been called N12CNR3P45CRP3.

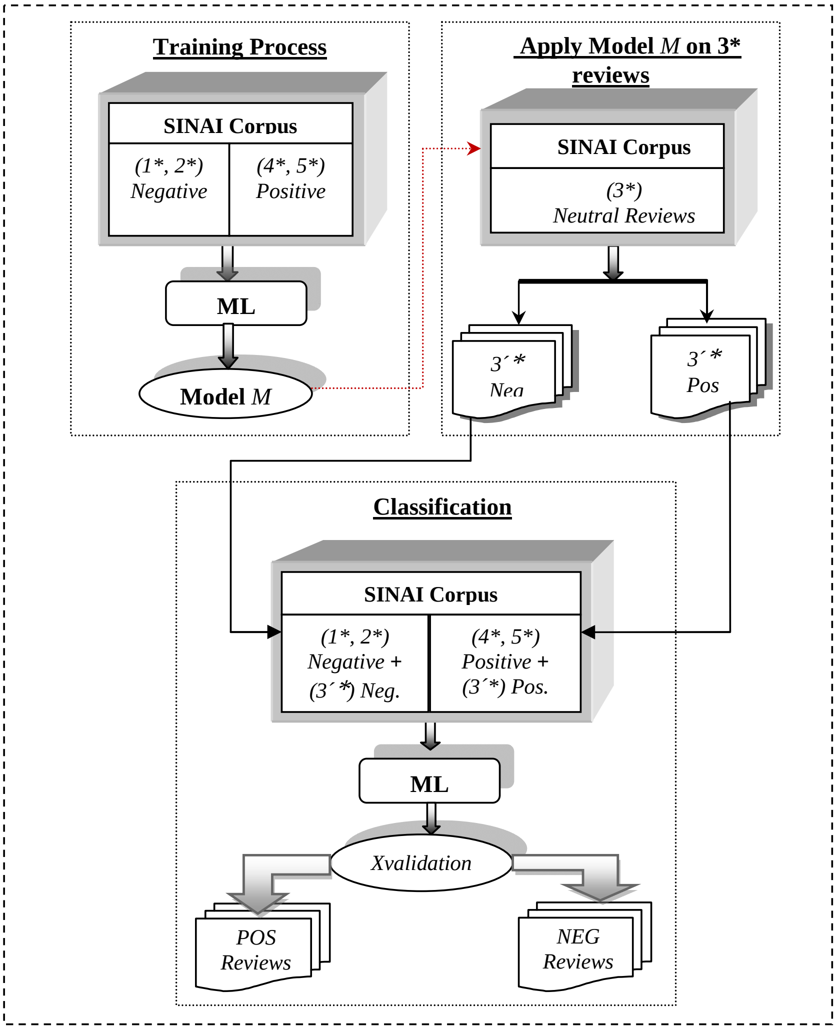


Figure 3: Process followed to build the final classifier C2

4.4 Evaluation

The system has been evaluated by applying 10-fold cross validation on each corpus, and measuring performance according to the indicators given below:

$$Precision\ (P) = \frac{tp}{tp + fp} \quad (1)$$

$$Recall\ (R) = \frac{tp}{tp + fn} \quad (2)$$

$$Accuracy = \frac{tp + tn}{tp + tn + fn + fp} \quad (3)$$

$$k = \frac{\Pr(a) + \Pr(e)}{1 - \Pr(e)} \quad (5)$$

$$Pr(a) = \frac{tp + tn}{tp + tn + fn + fp} \quad (6)$$

$$Pr(e) = \frac{(tp + fp)(fn + tn)(tp + fn)(fp + tn)}{(tp + tn + fn + fp)^2} \quad (7)$$

where *tp* (True Positives) are those assessments where the system and human expert agree on a label assignment, *fp* (False Positives) are those labels assigned by the system which do not agree with the expert assignment, *fn* (False Negatives) are those labels that the system failed to assign as they were given by the human expert, and *tn* (True Negatives) are those non assigned labels that were also discarded by the expert [see Tab. 4]. The precision tells us how well the labels are assigned by our system (the fraction of assigned labels that are correct). The recall measures the fraction of expert labels found by the system. Finally, accuracy combines both precision and recall, calculating the proportion of true results (both true positives and true negatives). *k*: Kappa; *Pr(a)* is the relative observed agreement among raters; *Pr(e)* is the hypothetical probability of chance agreement [Sebastiani, 2002].

	True Yes	True No
Predicted Yes	*tp*	*fp*
Predicted No	*fn*	*tn*

Table 4: Contingency table

5 Results and discussion

The experiments were divided into three parts: the first one was run on the Taboada corpus, the second one on the original SINAI corpus and the third part on the SINAI-B corpus. For each corpus we performed experiments for each partition using different combinations of neutral examples. As a reminder, the corpora that have been used are:

- N12P45: the 3 star reviews were ignored
- N123P45: the 3 star examples are considered as negative reviews
- N12P345: the 3 star examples are considered as positive reviews
- N12NR3P45R3 (only applicable to Taboada corpus): the 3 stars labeled "non-recommended" are included in the negative set and the 3 stars tagged with "recommended" are considered as positive samples.
- N12CNR3P45CRP3: the corpus includes the 3 star examples tagged as "recommended" by the C1 classifier into the positive samples and the 3 star examples tagged as "non-recommended" by the C1 classifier into the negative samples.

In addition, these experiments were run using the three machine learning algorithms SVM, LR and KNN.

The experiments accomplished with the Taboada corpus are shown in Table 5. As presumed, the best result was obtained when recommended and non-recommended information in the 3 star reviews (N12NR3P45R3) was taken into account. The 20 reviews labeled with 3 stars were distributed between 6 as positive (recommended) and 14 as negative (non-recommended). However, it is very interesting to note that the second best results were obtained when we applied the approach described in Figure 3 for all the algorithms (N12CNR3P45CRP3). According to the machine learning algorithm, LR clearly overcomes the other two algorithms. In addition, the Kappa measure is also bigger for LR than for SVM and KNN.

Regarding the SINAI corpus, we performed almost the same experiments as with the Taboada corpus, except for the case where the "recommended" and "non-recommended" information was used. Table 6 shows the results obtained. Although the best results were achieved with the new model proposed in Figure 3, the improvement is not as significant as the one obtained with Taboada corpus. We think the main reason for this is the high accuracy already obtained with the baseline case. This makes it very difficult to improve the final results. In fact, the best improvement is obtained with KNN, the algorithm with the worst accuracy. Nevertheless, the experiments reinforce our hypothesis about the improvement when neutral examples are used.

Algorithm	**Corpus**	**Precision**	**Recall**	**Accuracy**	**Kappa**
SVM	N12P45	78.62%	87.37%	80.38%	0.603
	N123P45	79.05%	85.88%	81.66%	0.634
	N12P345	76.25%	85.71%	77.10%	0.531
	N12NR3P45R3	81.84%	91.49%	85.16%	0.702
	N12CNR3P45CRP3	80.74%	88.05%	**82.74%**	0.653
LR	N12P45	87.39%	86.20%	86.32%	0.725
	N123P45	86.22%	85.69%	85.70%	0.714
	N12P345	85.25%	84.03%	84.39%	0.639
	N12NR3P45R3	87.52%	87.31%	87.33%	0.746
	N12CNR3P45CRP3	88.24%	87.57%	**87.61%**	0.751
KNN	N12P45	72.34%	71.27%	71.53%	0.427
	N123P45	79.05%	75.88%	71.66%	0.634
	N12P345	71.67%	70.00%	70.88%	0.405
	N12NR3P45R3	74.91%	69.57%	72.90%	0.393
	N12CNR3P45CRP3	74.21%	72.19%	**72.80%**	0.449

Table 5: Taboada corpus with different distribution of 3 star reviews

Algorithm	**Corpus**	**Precision**	**Recall**	**Accuracy**	**Kappa**
SVM	N12P45	94.38%	99.65%	94.20%	0.421
	N123P45	92.02%	98.77%	91.41%	0.489
	N12P345	94.70%	99.28%	94.19%	0.413
	N12CNR3P45CRP3	94.64%	99.61%	**94.44%**	0.466
LR	N12P45	93.68%	68.26%	94.80%	0.497
	N123P45	88.89%	68.47%	91.50%	0.481
	N12P345	91.28%	63.26%	94.24%	0.374
	N12CNR3P45CRP3	95.45%	69.87%	**95.01%**	0.536
KNN	N12P45	63.66%	64.11%	80.38%	0.273
	N123P45	63.80%	65.20%	84.19%	0.286
	N12P345	63.41%	64.49%	80.49%	0.276
	N12CNR3P45CRP3	68.10%	70.23%	**89.03%**	0.377

Table 6: SINAI corpus with different distribution of 3 star reviews

As imbalance may affect classifier behavior, a drawback of the original SINAI corpus is the great difference between the number of positive and negative examples. So we performed the same experiments with the SINAI-B corpus. The results are shown in Table 7 and as we can see the results obtained when we consider the neutral examples are better than when we neglect them, although in this case the improvement is slightly lower than with the original SINAI corpus. However, these experiments highlight the advantage of using the neutral examples in an appropriate way.

6 Conclusions

This paper focuses on the importance of neutral examples in reviews used in sentiment polarity classification tasks. We have applied several machine learning algorithms on different corpora in order to classify the sentiment polarity of subjective documents. We proposed a model to divide the neutral examples of a corpus into positive and negative samples. This information was then incorporated into the original corpus in order to regenerate and improve the model.

Algorithm	Corpus	Precision	Recall	Accuracy	Kappa
SVM	N12P45	84.69%	90.50%	84.92%	0.687
	N123P45	79.62%	77.47%	80.72%	0.610
	N12P345	80.18%	89.89%	78.23%	0.474
	N12CNR3P45CRP3	87.54%	83.40%	**86.17%**	0.722
LR	N12P45	93.30%	92.67%	93.04%	0.857
	N123P45	87.81%	87.52%	87.73%	0.752
	N12P345	83.60%	78.47%	83.47%	0.604
	N12CNR3P45CRP3	93.33%	93.17%	**93.20%**	0.862
KNN	N12P45	74.18%	72.88%	73.94%	0.460
	N123P45	70.75%	70.40%	70.52%	0.406
	N12P345	63.66%	63.48%	67.35%	0.267
	N12CNR3P45CRP3	75.19%	75.31%	**75.28%**	0.502

Table 7: SINAI-B corpus with different distribution of 3 star reviews

The results obtained encourage us to continue working along this line. Thus, in future work we will include more information on neutral examples in order to improve the classification, for example, using external resources like SentiWordNet [Baccianella et al., 2010]. In addition, we will apply the classifier developed to other corpora such as the Pang corpus on movie reviews [Pang and Lee, 2008].

Acknowledgments

This work has been partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. It has been also partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER) through the TEXT-COOL 2.0 project (TIN2009-13391-C04-02) from the Spanish Government. Also another part of this project was funded by Agencia Española de Cooperación Internacional para el Desarrollo MAEC-AECID.

References

[Baccianella et al., 2010] Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of LREC-10, 7th Conference on Language Resources and Evaluation, Valletta, MT, 2010, pages 2200-2204.

[Chang and Lin, 2001] Chang, C. C., Lin, & C. J. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. (last accessed 01/02/2011)

[Dave et al., 2003] Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In WWW ’03: Proceedings of the 12th international conference on World Wide Web. ACM, New York, NY, USA, pp. 519–528.

[Hatzivassiloglou and Wiebe, 2000] Hatzivassiloglou, V., & Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In: COLING-00: Proceedings International Conference on Computational Linguistics. pp. 299–305.

[Joachims, 1998] Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In: N´edellec, C., Rouveirol, C. (Eds.), Proceedings of ECML-98, 10th European Conference on Machine Learning. No. 1398. Springer Verlag, Heidelberg, DE, Chemnitz, DE, pp. 137–142.

[Kamps et al., 2004] Kamps, J., Marx, M., Mokken, R. J., & Rijke, M. D. (2004). Using wordnet to measure semantic orientation of adjectives. In Proceedings of LREC-04, Conference on Language Resources and Evaluation, pp. 1115–1118.

[Koppel and Schler, 2006] Koppel, M., & Schler, J. (2006). The importance of neutral examples for learning sentiment. Computational Intelligence 22 (2), 100–109.

[Liu, 2010] Liu, B. (2010). Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing, Second Edition.

[Martínez-Cámara et al., 2011] Martínez-Cámara, E., Martín-Valdivia, M.T.,. Ureña-López, L.A. (2011) Opinion Classification Techniques Applied to a Spanish Corpus. Proceedings of Natural Language Processing and Information Systems, pp. 169-176.

[Mihalcea and Strapparava, 2006] Mihalcea, R.,Strapparava, C.: Learnin to Laugh (automatically): Computational Models for Humor Recognition, Journal of Computational Intelligence, Vol. 22, 2006, 126-142

[Mullen and Collier, 2004] Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 412–418.

[O’Keefe and Koprinska, 2009] O’Keefe, T., & Koprinska, I. (2009). Feature selection and weighting methods in sentiment analysis. In: Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia.

[Pang and Lee, 2004] Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL’04 Association for Computational Linguistics. pp. 271–278.

[Pang and Lee, 2005] Pang, B., Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the ACL’05 Association for Computational Linguistics. pp. 115–124.

[Pang and Lee, 2008] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Found. Trends Information Retrieval 2 (1-2), 1–135.

[Pang et al., 2002] Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 79–86.

[Sebastiani, 2002]. Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), 1.

[Somprasertsri and Lalitrojwong, 2010] Somprasertsri, G., Lalitrojwong, P.: Mining Feature-Opinion in Online Custumer Reviews for Opinion Summarization, Journal of Universal Computer Science, vol. 16, 2010, 938-955.

[Strapparava and Mihalcea, 2008] Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing* (SAC '08). ACM, 2008, 1556-1560

[Taboada et al., 2006] Taboada, M., Anthony, C., & Voll, K. (2006). Methods for creating semantic orientation dictionaries. In Proceedings of 5th International Conference on Language Resources and Evaluation (LREC).

[Taboada and Grieve, 2004] Taboada, M., & Grieve, J. (2004). Analyzing appraisal automatically. In: In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications. pp. 158–161.

[Tan and Zhang, 2008] Tan, S., Zhang, J. (2008) .An empirical study of sentiment analysis for Chinese documents. Expert System with Applications 34, 2622–2629

[Tsytsarau and Palpanas, 2012] Tsytsarau, M. and Palpanas, T. (2012) Survey on mining subjective data on the web Data Mining and Knowledge Discovery. Vol. 24 (3) pp. 478-514. Doi: 10.1007/s10618-011-0238-6

[Turney, 2002] Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: ACL ’02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, USA, pp. 417–424.

[Vapnik, 1995] Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer, New York.

[Wiebe et al., 2001] Wiebe, J., Wilson, T., Bell, M. (2001). Identifying collocations for recognizing opinions. In: Proceedings of the ACL’01 Association for Computational Linguistics Workshop on Collocation: Computational Extraction, Analysis, and Exploitation. pp. 24–31.

[Wilson et al., 2005] Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In: HLT ’05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Morristown, NJ, USA, pp. 347–354.

[Yi et al., 2003] Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: IEEE Intl. Conf. on Data Mining (ICDM). pp. 427–434.

[Yu and Hatzivassiloglou, 2003] Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics, Morristown, NJ, USA, pp. 129–136.

2.1.3 Experiments with SVM to classify opinions in different domains

- Mohammed Rushdi-Saleh, María Teresa Martín-Valdivia, Arturo Montejo Ráez, Luis Alfonso Ureña López: Experiments with SVM to classify opinions in different domains. Expert Systems With Applications 38(12): 14799-14804, 2011.
  - Status: Published
  - Impact Factor: 2.203
  - Category: (COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE) Ranking: 2011: 22/111; (ENGINEERING, ELECTRICAL & ELECTRONIC) Ranking: 2011: 41/245; (OPERATIONS RESEARCH & MANAGEMENT SCIENCE) Ranking: 2011: 5/77.
  - Number of cites: 9

Expert Systems with Applications 38 (2011) 14799–14804

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Experiments with SVM to classify opinions in different domains

M. Rushdi Saleh *, M.T. Martín-Valdivia, A. Montejo-Ráez, L.A. Ureña-López

SINAI Research Group, Department of Computer Science, University of Jaén, Campus Las Lagunillas, E-23071 Jaén, Spain

ARTICLE INFO

Keywords:
Opinion mining
Machine learning
SVM
Corpora

ABSTRACT

Recently, opinion mining is receiving more attention due to the abundance of forums, blogs, e-commerce web sites, news reports and additional web sources where people tend to express their opinions. Opinion mining is the task of identifying whether the opinion expressed in a document is positive or negative about a given topic. In this paper we explore this new research area applying Support Vector Machines (SVM) for testing different domains of data sets and using several weighting schemes. We have accomplished experiments with different features on three corpora. Two of them have already been used in several works. The last one has been built from Amazon.com specifically for this paper in order to prove the feasibility of the SVM for different domains.

1. Introduction

Nowadays, the web is the most important place for expressing sentiments, evaluations, and reviews. Lots of people are tending to give their opinions in forums, blogs or wikis. However, with the rapid growth of e-commerce activity, the number of reviews and opinions about products has increased exponentially and this source of information is becoming unworkable. A customer who wants to buy a product usually searches information on the Internet trying to find analyses of other customers. In fact, web sites such as Amazon,[1] Epinions[2] or IMDb[3] can affect the customer decision.

Nevertheless, it is becoming an impossible task to read all of these reviews and opinions in different forums or blogs. On the other hand, it is also very difficult for the companies to track this amount of evaluations about their products or services. Therefore, it is necessary to develop new methods that can improve the access to this kind of information.

The automatic processing of documents to detect opinion expressed therein, as a unitary body of research, has been denominated opinion mining. Most work on this area has been carried out on highly subjective text types such as articles in blogs or product reviews. Authors of such type of documents mostly express their opinions quite freely. In general, an opinion is a message expressing a belief about something, the expression of a belief that is held with confidence but not substantiated by positive knowledge or proof.

Opinion mining is a recent research area in the field of the Text Mining (TM) that has been designated by different terms like subjectivity analysis, sentiment analysis or sentiment orientation. There are lots of definitions for each one. Pang and Lee (2008) captured different definitions about these terms based on applications done in this field. For example, Subjectivity Analysis is defined as the recognition of opinion-oriented language in order to distinguish it from objective language. Sentiment Analysis classifies reviews according to their polarity (positive or negative). Henceforth, all these terms refer to the same field of study.

Some tasks in opinion mining try to classify the detected opinion using different scales. In a number of cases, the purpose is to identify opinions in a text and classify them into positive, negative or neutral classes. In other occasions, the goal is to assign different rates, such as “very bad”, “bad”, “satisfactory”, “good”, “very good”, or “excellent”. The sentiments can be ranked into a range of one to five stars. Other systems use the “thumb up” or “thumb down” notation.

Although sentiment analysis may seem an easier task than text categorization, opinion mining includes several challenges which make researchers focus on this stimulating topic. These challenges can be found in review texts, since it is not a quality text such as the content found in newspapers or scientific journals. This text can contain many orthographic mistakes, abbreviations, colloquial expressions, idiomatic expressions or ironic sentences. Another important issue is the time influence. The opinions may change over time due to product improvement (Balog, Mishne, & de Rijke, 2006). In addition, sentiment analysis systems are highly domain dependant. The extraction of features for a corpus about movie reviews is different from one about electronic products. The results can vary significantly from a domain to another. All of these factors make the opinion mining a very interesting and challenging task.

* Corresponding author. Tel.: +34 953 212898; fax: +34 953 212472.
E-mail address: msaleh@ujaen.es (M. Rushdi Saleh).

[1] http://www.amazon.com.

[2] http://www.epinions.com.

[3] http://www.imdb.com.

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.05.070

Different approaches have been applied in the field of sentiment analysis but mainly we can distinguish two main methodologies used in opinion mining. On the one hand, there is a lot of work based on the symbolic approach, which applies manually crafted rules and lexicons. The document in this approach is represented as a collection of words. Then, the sentiment of each word can be determined by different methods, for example, using a web search (Hatzivassiloglou & Wiebe, 2000) or consulting a dictionary like WordNet[4] (Kamps, Marx, Mokken, & Rijke, 2004). On the other hand, machine learning techniques are very extended in order to attack the classification of reviews according to their orientation (positive or negative). In this approach the document is represented by different features for classification task. Then, a machine learning algorithm is applied. These features may include the use of *n*-grams or defined grammatical roles like, for instance, adjectives. Machine learning algorithms commonly used are Support Vector Machines, Maximum Entropy or Naïve Bayes. Of course, there are several researches that combine both approaches (symbolic and machine learning).

In this work, we have applied a supervised machine learning method in order to classify reviews. Specifically, we have used Support Vector Machines (SVM) on three datasets with different sizes and domains. One is the Pang and Lee (2004) one about movie reviews; the second corpus is the one prepared by Taboada and Grieve (2004) about several topics like computers, hotels or music; and finally, we have generated one last corpus by crawling opinions about digital cameras from the Amazon website.

We chase several goals. First, we compare the results obtained with these corpora in order to characterize our own corpus feasibility. Secondly, we test the behavior of the system when we use *n*-grams. Finally, we check our model applied over several corpora with different sizes, domains and number of positive and negatives samples.

The paper is organized as follows. Next section comments some related work and approaches in sentiment analysis. The datasets used in our experiments are described in Section 3. Section 4 presents the method applied and the experiments carried out. Results obtained are discussed in Section 5. Finally, in Section 6, the main conclusions and proposals for further work are expounded.

2. Related work

In the recent years, relevant research has been developed in the area of opinion mining. Hatzivassiloglou and McKeown (1997) used adjectives as good clues to determinate the text orientation. They also studied phrases where adjectives are connected with conjunction words like "and" or "but". Then, they applied log-linear regression to check whether the two adjectives had the same polarity. They performed clustering in order to separate the adjectives in two classes and they assumed the class with highest frequency to be as a positive class.

Wiebe (2000) classified objective and subjective sentences using a corpus labeled with subjective adjectives.

Turney (2002) introduced an unsupervised learning algorithm for classifying a review as "recommended" (*thump up*) or "not recommended" (*thump down*). First, they extracted phrases containing adjectives or adverbs. Secondly, they calculated the semantic orientation using Pointwise Mutual Information (PMI). Finally, they classified the reviews based on the average semantic orientation of the phrase. Turney and Littman (2003) also introduced a method for inferring the semantic orientation from associations. The relation between a word and a set of positive or negative words was measured using two different statistical measures: PMI and Latent Semantic Analysis (LSA). The results showed that PMI works better than LSA. Kamps et al. (2004) based their work on the paper of Turney (Turney, 2002) but they used the semantic network WordNet and a set of tagged words.

Hu and Liu (2004) made summaries with positive and negative opinions about the features of product reviews. First, they identified features by detecting frequent words and then, they defined the opinion sentences which contained both a feature and at least one adjective. The adjective is checked using a list of 30 predefined adjectives. If the adjective did not belong to this list and it was neither a synonym nor an antonym, then the adjective was included in the list.

Esuli and Sebastiani (2005) dealt with a new method based on the assumption that terms with similar orientation tend to have similar glosses. The method used a semi-supervised learning in order to classify terms as positives or negatives.

Ding and Liu (2007) improved the system proposed by Hu by assigning a score to opinion words located near to the feature. The score depends on the distance between the opinion word and the feature. Low score was given to the opinion words that were far from the feature.

A common approach to sentiment analysis is to employ supervised machine-learning methods to acquire prominent features of sentiment. However, the success of these methods depends on the domain, topic and time-period represented by the training data.

Pang, Pang, and Lee (2002) applied machine learning methods (Naïve Bayes, Maximum Entropy and SVM) on movie reviews to determine the polarity. The data was downloaded from Internet Movie Database (IMDb). They used 700 negative reviews and 700 positive reviews. In order to apply machine learning algorithms on the documents, they used the standard bag of features framework and a predefined set of features that can appear in a document. They also treated the effect of the negation by adding the negation prefix *NOT_*. The word position and the part-of-speech (POS) was also taken into account. They accomplished several experiments using different *n*-grams techniques and the results showed that the use of unigram was the most effective method.

Mullen and Collier (2004) worked on the same dataset used by Pang et al. (2002). They calculated the average rating for the whole collection. Then, the reviews under the average rating were classified as negatives and those above the average rating were classified as positives. They investigated various features including Combination of Turney value, the three text-wide Osgood values, word unigrams or lemmatized unigrams. In addition, they accomplished experiments over a movie reviews corpus downloaded from the Pitchfork Media.[5] In this case, they extracted the same features and extra features based on the movie domain. The machine learning algorithm used was the SVM. They concluded that the combination of unigrams and lemmatized unigrams outperforms the models which do not use this kind of information.

Prabowo and Thelwall (2009) applied SVM with combined methods to classify reviews from different corpora. One of these datasets was downloaded from Pang and Lee (2004) and it includes 1,000 positive and 1,000 negative samples. Several classifiers were used: General Inquirer Based Classifier (GIBC), Rule-Based Classifier (RBC), Statistics Based Classifier (SBC) and SVM. They accomplished a hybrid classification, where if one classifier fails to classify a document, the classifier passes the document onto the next classifier until the document is correctly classified or no other classifier remains. The results indicated that SBC and SVM improve their effectiveness in the hybrid classification.

[4] http://wordnet.princeton.edu.

[5] http://www.pitchforkmedia.com.

Table 1
Number of reviews per product in the SINAI corpus.

Camera	Reviews
CanonA590IS	400
CanonA630	300
CanonSD1100IS	426
KodakCx7430	64
KodakV1003	95
KodakZ740	155
Nikon5700	119
Olympus1030SW	168
PentaxK10D	126
PentaxK200D	90

3. Corpora description

One of our goals in this paper is to apply SVM on several datasets with different sizes and domains. For this reason we have used three different corpora: the corpus used by Pang and Lee (2004), the corpus prepared by Taboada and Grieve (2004) and a new corpus that we have generated by crawling from Amazon.com. A detailed description of the three corpora is given below.

3.1. Pang corpus

This corpus[6] was prepared by Pang and Lee (2004) in order to classify movie reviews collected from the IMDb.com (Internet Movie Database). They examined the data manually to ensure the quality of the collection. They generated several corpora from the same website using different rating systems and different number of samples. In our experiments we have used the collection with 2,000 reviews (1,000 positive samples and 1,000 negative samples). All these documents have been written before 2002, with a cap of 20 reviews per author (312 authors total) per category. The reviews are classified according to the rating systems either stars or numbers. In a five-star system (or any compatible numbering system), four stars or more are considered positive, while two stars or less are considered negative. In a four-star system (or any other compatible numbering system), three stars or more are considered positive, while one star or less are considered negative.

3.2. Taboada corpus

This collection[7] was used by Taboada and Grieve (2004) and Taboada et al. (2006) with the main goal of classifying text automatically based on subjective content. They applied a standard method for calculating semantic orientation which is based on Turney (2002), and also they applied linguistic classification of appraisal. The corpus includes 400 opinions collected from the website Epinions.com divided into 200 reviews classified as "recommended" (positive) and 200 as "not recommended" (negative). The texts contain opinions about products and services like movies, books, cars, cookware, phones, hotels, music and computers. The total number of categories is eight and the corpus contains 25 positive and 25 negative reviews per each category.

3.3. SINAI corpus[8]

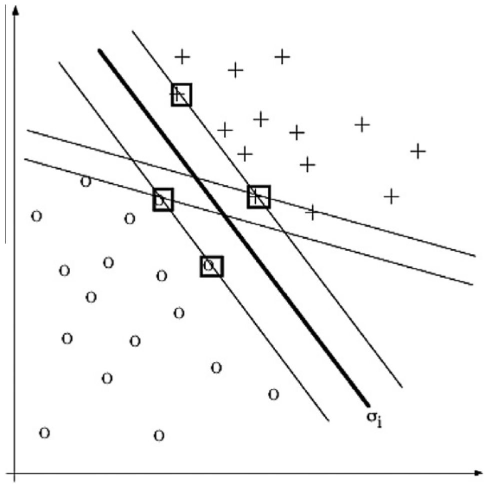


Fig. 1. Support vectors delimiting the widest margin between classes.

Finally, we have created a new corpus by crawling the Amazon website. We have extracted opinions about cameras with different brands and series. A total of 1,943 documents were labeled with different number of starts. The reviews were rated using the number of stars. In order to select the positive and negative examples, the reviews ranked with 3, 4 and 5 stars are classified as positive opinions (1,798 text reviews). The documents ranked with 1 and 2 stars are considered as negative reviews (145 text reviews). Table 1 shows the distribution of reviews per each camera model.

4. Methodology

4.1. Support Vector Machines

In this work, Support Vector Machines have been applied in order to classify a set of opinions as positives or negatives. SVM is a product of applied complexity theory developed by Vapnik (1995). Some years ago, Joachims (1998) proposed SVM for text categorization tasks, to profit from its robustness in high dimensional spaces. The name of the algorithm taken from the idea behind it: find those samples (support vectors) that delimit the widest frontier between positive and negative samples in the feature space Fig. 1. The width of such border is known as the margin hyperplane, and SVM tries to find the maximal margin by applying constraint quadratic optimization.

Support Vector Machines have been applied successfully in many text classification tasks due to their principal advantages: first, they are robust in high dimensional spaces; second, any feature is relevant; third, they are robust when there is a sparsely set of samples; finally, most text categorization problems are linearly separable. In addition, SVM has achieved good results in opinion mining and this algorithm has overcome other machine learning techniques (O'Keefe & Koprinska, 2009).

4.2. Experimental framework

We have used the Rapid Miner[9] software with its text mining plug-in which contains different tools designed to assist on the preparation of text documents for mining tasks (tokenization, stop word removal and stemming, among others). Rapid Miner is an environment for machine learning and data mining processes. We have implemented our experiments using the *libsvm*[10] learner by Chang

[6] The dataset is freely available and can be downloaded from the URL www.cs.cornell.edu/people/pabo/movie-review-data.

[7] http://www.sfu.ca/mtaboada/research/SFU_Review_Corpus.html.

[8] SINAI stands for the name of our research group "Sistemas INteligentes de Acceso a la Información" (Intelligence Systems for Information Access).

[9] http://rapid-i.com/.

[10] http://www.csie.ntu.edu.tw/cjlin/libsvm/.

Table 2
Pang corpus 10-fold cross-validation results.

	TFIDF			BO			TO		
	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)
Precision	82.54	83.72	84.01	84.93	85.83	86.19	85.74	85.37	85.29
Recall	84.30	85.70	85.80	83.90	84.50	84.50	77.10	78.70	78.60
F1	83.36	84.61	84.79	84.36	85.06	85.22	81.11	81.82	81.71
Accuracy	83.20	84.45	84.65	84.45	85.15	85.35	82.05	82.50	82.40
Kappa	66.40	68.90	69.30	68.90	70.30	70.70	64.10	65.00	64.80

Table 3
Taboada corpus 10-fold cross-validation results.

	TFIDF			BO			TO		
	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)
Precision	71.79	72.93	72.35	70.89	69.97	71.00	68.86	70.33	70.33
Recall	70.00	73.50	75.00	59.00	59.50	58.50	32.00	32.50	32.50
F1	70.58	72.95	73.37	63.89	64.05	63.95	43.27	43.99	43.99
Accuracy	71.00	73.00	73.25	66.75	66.75	67.25	58.50	59.00	59.00
Kappa	42.00	46.00	46.50	33.50	33.50	34.50	17.00	18.00	18.00

Table 4
SINAI corpus 10-fold cross-validation results.

	TFIDF			BO			TO		
	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)
Precision	92.06	92.17	92.02	90.37	90.58	90.42	87.97	88.07	88.07
Recall	98.59	98.71	98.77	99.24	99.35	99.30	99.77	99.77	99.77
F1	95.20	95.32	95.27	94.59	94.76	94.65	93.50	93.55	93.55
Accuracy	91.30	91.51	91.41	90.07	90.38	90.17	87.85	87.96	87.96
Kappa	48.80	50.00	48.90	34.90	37.30	35.80	06.40	07.70	07.70

Table 5
Pang corpus 3-fold cross-validation results.

	TFIDF			BO			TO		
	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)
Precision	81.06	82.69	82.86	84.62	84.76	85.36	83.87	84.62	84.81
Recall	84.10	85.90	85.61	84.50	83.60	84.30	75.41	76.71	76.71
F1	82.52	84.24	84.17	84.48	84.12	84.78	79.33	80.37	80.44
Accuracy	82.20	83.95	83.95	84.50	84.25	84.90	80.45	81.35	81.45
Kappa	64.40	67.90	67.90	69.00	68.50	69.80	60.90	62.70	62.90

and Lin (2001), which is integrated into Rapid Miner as one of the available operators.

We have used the Vector Space Model (VSM) in order to generate the bag of words for each document. The English Porter stemming algorithm was applied in order to reduce words to their common root or stem. We have also removed some tokens using a stop words list. However, we have preserved some useful sentiment information such as “ok” and “not”.

On the other hand, one of our main goals is to compare the influence of using different *n*-gram schemes. For this reason, we have applied several *n*-gram models: unigrams, bigrams and trigrams. Finally, for each *n*-gram scheme, we have used three different approaches to generate the word vectors: word frequency in document and in the entire corpus (TFIDF), Binary Occurrence (BO) and Term Occurrence (TO). Different experiments were carried out as result of the possible combinations of three factors: corpus, weighting scheme and *n*-gram model. Thus, it resulted in a total of 3 × 3 × 3 = 27 experiments.

5. Results and discussion

In order to test our system, we have applied 3-fold and 10-fold cross validation for each corpus. Analyzing the results according to the *n*-grams scheme, we can notice that, in general, the trigram model slightly overcomes unigram and bigram models. For example, in 10-fold cross validation, only for the SINAI corpus with TFIDF and BO and the Pang corpus with TO the results are vaguely lower. However, the differences are insignificant and the results obtained for the different corpora and techniques are comparable.

As regards the weighting scheme, it seems that the TO is the worst option for all the corpora and all the *n*-gram techniques. However, TFIDF and BO obtain a similar result.

Table 6
Taboada corpus 3-fold cross-validation results.

	TFIDF			BO			TO		
	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)
Precision	67.67	68.03	66.87	71.35	69.55	67.10	67.09	66.63	66.21
Recall	71.99	72.00	72.49	59.52	53.51	51.48	31.97	31.47	30.97
F1	69.75	69.92	69.55	64.52	60.26	58.16	43.27	42.72	42.16
Accuracy	68.75	69.00	68.25	67.24	64.75	63.00	58.24	57.99	57.74
Kappa	37.50	38.00	36.50	34.50	29.50	26.00	16.50	16.00	15.50

Table 7
SINAI corpus 3-fold cross-validation results.

	TFIDF			BO			TO		
	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)	Unigram (%)	Bigram (%)	Trigram (%)
Precision	91.71	91.12	91.13	89.86	89.92	89.73	87.80	87.89	87.89
Recall	98.82	98.94	99.06	99.47	99.47	99.53	99.82	99.82	99.82
F1	95.13	94.87	94.93	94.42	94.45	94.37	93.43	93.48	93.48
Accuracy	91.15	90.63	90.74	89.71	89.76	89.60	87.70	87.80	87.80
Kappa	47.00	42.10	42.40	30.00	30.30	28.40	03.90	05.30	05.30

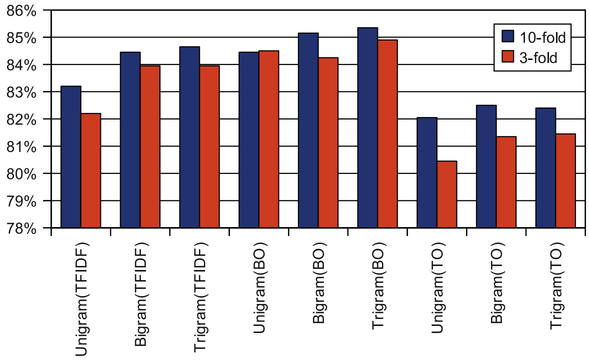


Fig. 2. Accuracy achieved by SVM with different features on Pang corpus.

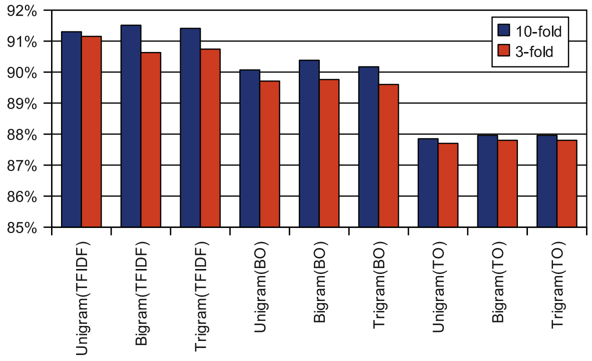


Fig. 4. Accuracy achieved by SVM with different features on SINAI corpus.

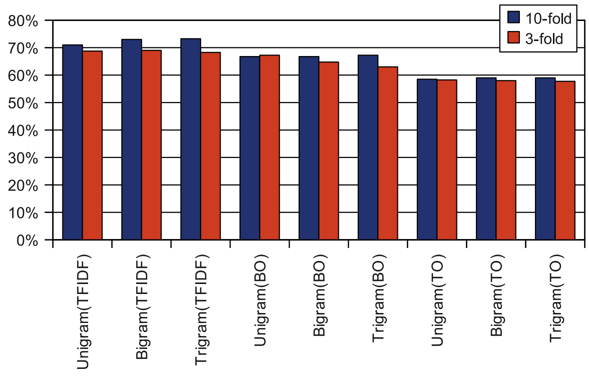


Fig. 3. Accuracy achieved by SVM with different features on Taboada corpus.

Tables 2, and 5 show the results obtained over the Pang corpus. If we compare our results with other similar works on the same dataset and with the similar techniques, we can find that our results are promising. For example, Pang et al. (2002) applied machine learning techniques on a collection of 1,400 samples (700 positives and 700 negatives) about movie reviews incorporated with different features including part of speech and word position. They found that SVM outperforms Naïve Bayes and Maximum Entropy. However, they obtained a maximum accuracy of 82.90% by applying SVM using unigram and feature presence with 3-fold cross validation. In our experiments, only for TO the results are lower than this value and for the TDFIF with unigrams (82.20%) accuracy is slightly smaller. The rest of the experiments (included all of 10-fold cross validation) overcomes the best results obtained for Pang et al. (2002). The best result obtained with the Pang corpus is 85.35% using trigram and BO features and 10-fold cross validation, while applying 3-fold cross validation is 84.90% with the same features.

As regards Taboada corpus, the results are lower than the other collections. The best accuracy obtained is 73.25% using TFIDF and trigram word vector with 10-fold cross validation. The range of accuracy in other experiments on this same data set ranged between 59% and 73%. Equally, for 3-fold cross validation the results vary between 57% and 69%. From our point of view, the accuracy is affected by the divergent domain due to the fact that the Taboada collection contains different topics (movies, music, hotels, etc.) and, also, due to the number of samples because the corpus size is relatively small.

Comparing our results with the research on the papers of Taboada and Grieve (2004); Taboada, Anthony, and Voll (2006), it can be noticed that our system works better. For example, in Taboada et al. (2006) the best overall accuracy they achieved was 56.75%.

This value is lower than any experiment accomplished by our models. Another approach by Taboada and Grieve (2004) on the same dataset achieve better results. The methodology is based on the adjectives, where they assigned a weight for adjectives according to their position in the text. They supposed that opinions tend to be expressed in the middle and at the end part of the text. Although the overall accuracy using this technique was 65%, this value is smaller than most of those in our experiments on the Taboada corpus.

Finally, the experiments carried out with our corpus achieved the best result (accuracy 91.51% with TFIDF, bigrams and 10-fold cross validation). We believe the main reason is that we have crawled enough data. In addition, this result can be affected by the data domain. All reviews were about one product (digital cameras) with different brands, i.e., Canon, Kodak, Nikon, etc. Thus, most of review comments specify features easily identifiable. On the contrary, Pang corpus includes movie reviews which it is a challenged domain (Turney, 2002) because a recommended movie often contains unpleasant scenes which reduce the average semantic orientation. On the other hand, as we have already commented, the overall accuracy in Taboada corpus is the worst obtained due to the variance of the domain.

TO weighting scheme seems to be the worst. However, we obtain similar results for TFIDF and BO although the first one works better for Taboada corpus and SINAI corpus and BO obtain a high result for Pang corpus (see Tables 2–7).

As regard the *n*-grams techniques, trigram is visibly superior for Pang corpus and Taboada corpus while bigram achieves better results for SINAI corpus. Thus, it seems clear that unigram is not a good option for our system.

Figs. 2–4 summarize the results obtained for accuracy applying 3-fold and 10-fold cross-validation over Pang, Taboada and SINAI corpora respectively (see Tables 2–7).

6. Conclusions

The main goal of this paper is to compare different corpora available for scientific research in opinion mining. In addition, we have introduced a new corpus that includes reviews about digital cameras. This corpus constitutes a valuable resource to test opinion mining systems. We have also applied a machine learning algorithm (SVM) with different features in order to test how the sentiment classification is affected. We have used different weighting schemes (TFIDF, BO, TO) and several *n*-grams techniques (unigrams, bigrams and trigrams).

We have noticed that the corpus size and the corpus domain have an effect on the system performance.

Besides, we have confirmed that SVM is a promising tool to deal with sentiment orientation classification.

For further work we would like to examine how the different rating reviews can affect the results. We will accomplish experiments changing the number of stars to be considered as positive or negative samples. Moreover, we would like to investigate the integration of external knowledge like SentiWordNet (Esuli & Sebastiani, 2006).

Acknowledgments

This paper has been partially supported by a Grant from the Spanish Government, project TEXT-COOL 2.0 (TIN2009-13391-C04-02), a Grant from the Andalusian Government, project GeOasis (P08-TIC-41999), and a Grant from the Fondo Europeo de Desarrollo Regional (FEDER). Also another part of this Project was funded by Agencia Española de Cooperación Internacional para el Desarrollo MAEC-AECID.

References

Balog, K., Mishne, G., & de Rijke, M. (2006). Why are they excited?: Identifying and explaining spikes in blog mood levels. *EACL '06: Proceedings of the 11th conference of the European chapter of the association for computational linguistics: Posters and demonstrations* (pp. 207–210). Morristown, NJ, USA: Association for Computational Linguistics.

Chang, C. C., & Lin, C. J. (2001). LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Ding, X., & Liu, B. (2007). The utility of linguistic rules in opinion mining. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 811–812).

Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. In *CIKM '05: Proceedings of the 14th ACM international conference on information and knowledge management* (pp. 617–624).

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th conference on language resources and evaluation (LREC06)* (pp. 417–422).

Hatzivassiloglou, V., McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 8th conference on European chapter of the association for computational linguistics* (pp. 174–181).

Hatzivassiloglou, V., & Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *COLING* (pp. 299–305).

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In: *KDD '04: Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 168–177).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Proceedings of ECML-98. 10th European conference on machine learning* (Vol. 1398, pp. 137–142). Heidelberg, DE, Chemnitz, DE: Springer-Verlag.

Kamps, J., Marx, M., Mokken, R.J., & Rijke, M.D. (2004). Using wordnet to measure semantic orientation of adjectives. In *Conference on language resources and evaluation (LREC)* (pp. 1115–1118).

Mullen, T., Collier, N., 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* (pp. 412–418).

O'Keefe, T., & Koprinska, I. (2009). Feature selection and weighting methods in sentiment analysis. In *Proceedings of the 14th Australasian document computing symposium*, Sydney, Australia.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL* (pp. 271–278).

Pang, T. B., Pang, B., & Lee, L. (2002). Thumbs up? sentiment classification using machine learning. In *Proceedings of EMNLP* (pp. 79–86).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrival, 2*(1–2), 1–135.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics, 3*(2), 143–157.

Taboada, M., & Grieve, J. (2004). Analyzing appraisal automatically. In *Proceedings of the AAAI spring symposium on exploring attitude and affect in text: Theories and applications* (pp. 158–161).

Taboada, M., Anthony, C., & Voll, K. (2006). Methods for creating semantic orientation dictionaries. In *Proceedings of 5th international conference on language resources and evaluation (LREC)*.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *ACL '02: Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424). Morristown, NJ, USA: Association for Computational Linguistics.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems, 21*(4), 315–346.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.

Wiebe, J. (2000). Learning subjective adjectives from corpora. In *Proceedings of the 17th national conference on artificial intelligence and twelfth conference on innovative applications of artificial intelligence* (pp. 735–740). AAAI Press/ The MIT Press.

2.2 Arabic polarity classification

2.2.1 OCA: Opinion Corpus for Arabic

- Mohammed Rushdi-Saleh, María Teresa Martín-Valdivia, Luis Alfonso Ureña López, José M. Perea-Ortega: OCA: Opinion corpus for Arabic. Journal of the American Society for Information Science and Technology 62(10): 2045-2054 (2011)

  - Status: Published
  - Impact Factor: 2.081
  - Category: (Information Science & Library Science) Ranking: 2011: 10/83; (Computer Science Information Systems) Ranking: 2011: 21/135.
  - Number of cites: 3
  - Comment: We have developed a corpus of movie reviews written in Arabic. This is a very Valuable resources that is made freely available for the research community.

OCA: Opinion Corpus for Arabic

Mohammed Rushdi-Saleh, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, and José M. Perea-Ortega
SINAI Research Group, Computer Science Department, University of Jaén, 23071, Spain.
E-mail: msaleh@ujaen.es; maite@ujaen.es; laurena@ujaen.es; jmperea@ujaen.es

Sentiment analysis is a challenging new task related to text mining and natural language processing. Although there are, at present, several studies related to this theme, most of these focus mainly on English texts. The resources available for opinion mining (OM) in other languages are still limited. In this article, we present a new Arabic corpus for the OM task that has been made available to the scientific community for research purposes. The corpus contains 500 movie reviews collected from different web pages and blogs in Arabic, 250 of them considered as positive reviews, and the other 250 as negative opinions. Furthermore, different experiments have been carried out on this corpus, using machine learning algorithms such as support vector machines and Naïve Bayes. The results obtained are very promising and we are encouraged to continue this line of research.

Introduction

The proliferation in the use of the World Wide Web and the rise of blogs and forums have paved the way for increased exposure of individual comments and sentiments. The growth of participation in the Internet fortifies the importance of public opinion as well as the use of public polls for different topics that many websites already employ. These opinions can be about different issues such as electronic products, politics, movies, books, cars, and many others. The idea of processing these comments or reviews has automatically attracted many researchers in the field of text mining, the aim being to be able to extract a general opinion about one item or theme among the huge unstructured data available in the Internet. This new task of analyzing and detecting the orientation of some data is given different names: opinion mining (OM), sentiment analysis, subjectivity analysis, or sentiment orientation.

On the other hand, the rapid growth of e-commerce has increased the number of reviews enormously. Nowadays, it is possible to find a variety of reviews for almost all the products in several merchants websites such as Amazon[1] or CNET[2]. When customers need to purchase laptops, cameras, cars, etc., they usually consult comments about that product and learn from other people's experiences. Summarized opinions could facilitate the task of Internet users and help them make the best choice by giving them a general idea about a product, without the need to explore the crowd data. These opinions are interesting not only for customers but also producers, who can obtain feedback through these reviews to more effectively adapt their products to customers' needs.

The tracking of the many reviews posted on different web pages is a challenging task for researchers. However, although comments in the web are expressed in any language, especially after the explosion of the Web 2.0 and the social web, most research in this field has focused on English texts (Pang & Lee, 2008), mainly because of the lack of resources in other languages. For example, despite the fact that Arabic is one of the top 10 languages most used on the Internet, according to the Internet World State[3] rank (see Figure 1) and is spoken by hundreds of millions of people, there is no reference corpus with sentiments or opinions. This is the main reason that has motivated the generation of an opinion corpus for Arabic in this work.

The Arabic language is becoming very interesting for many researchers in the field of text mining and information retrieval (Ahmed & Nürnberger, 2009; Kanaan, Al-Shalabi, Ghwanmeh, & Al-Ma'adeed, 2009). Several studies have been realized in this context, and there are different corpora, resources, and tools available for testing and implementing applications like text classification (Duwairi, 2006; Duwairi, Al-Refai, & Khasawneh, 2009) or name entity recognition (Shaalan & Raza, 2009). However, Arabic resources that focus on analyzing and mining opinions and sentiments are very difficult to find.

In this article, we present a new opinion corpus for Arabic (OCA) collected from a variety of web pages about movie

Received February 7, 2011; revised May 24, 2011; accepted May 24, 2011

© 2011 ASIS&T • Published online 15 July 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.21598

[1] http://www.amazon.com
[2] http://www.cnet.com
[3] http://www.internetworldstats.com

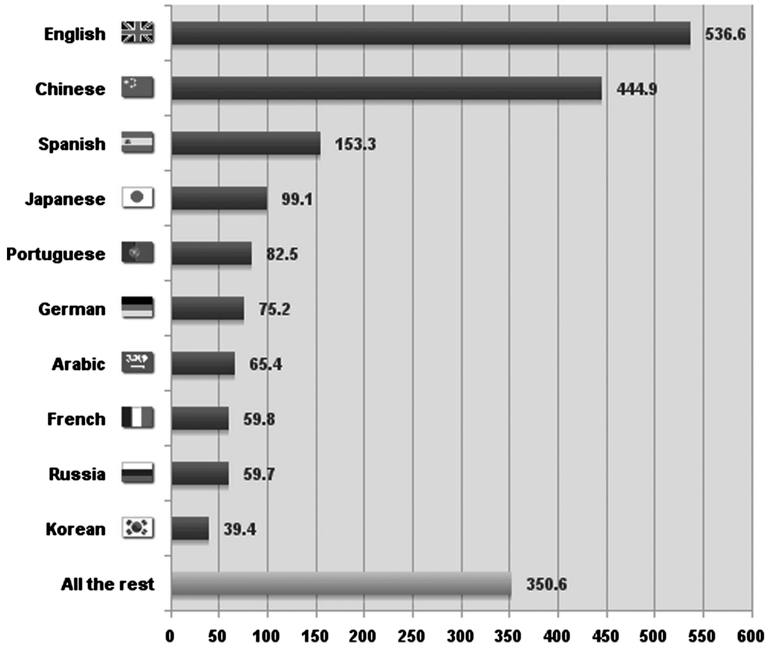


FIG. 1. Top 10 languages on the Internet in 2010 (in millions of users).

reviews in the Arabic language. In addition, we have carried out some experiments on the corpus, using machine learning algorithms to train an opinion classifier. Specifically, we have used the support vector machine (SVM) and Naïve Bayes (NB) algorithms to determine the opinion polarity of the reviews.

Background: Related Work

OM is a discipline that involves several interesting tasks. For example, opinion extraction, a specialization of information extraction, can be considered a specialization of the information extraction task. Its aim is to detect expressions denoting the key components of an opinion within a sentence or document. Another popular OM task focuses on detecting the subjectivity in a document, i.e., whether the document or part of the document is subjective or objective (informative). One of the most widely studied tasks is that of determining the polarity of a document, sentence, or feature (positive or negative) and measuring the degree of the polarity expressed in it. In this article, we train a classifier using SVM to determine whether an Arabic review is positive or negative. Next, we present an overview of the most important research and methods used in this area. In addition, we present a summary of the main work related to OM using non-English languages.

Related Work on Polarity Classification

Different approaches have been applied in the field of polarity or sentiment classification. Two main methodologies can be distinguished in this domain: On the one hand, there is a lot of work based on the semantic orientation approach, which represents the document as a collection of words. Then the sentiment of each word can be determined by different methods, for example, using a web search (Hatzivassiloglou & Wiebe, 2000) or consulting a lexical database like WordNet[4] (Kamps, Marx, Mokken, & Rijke, 2004). On the other hand, machine learning techniques are more extensively used for the classification of reviews. With this approach, the document is represented by different features that may include the use of n-grams or defined grammatical roles like, for instance, adjectives or other linguistic feature combinations, and then a machine learning algorithm is applied. Machine learning algorithms commonly used are SVMs, maximum entropy (ME), or NB.

Regarding methods that consider some linguistic features such as adjectives and adverbs, we can find many studies in the literature (Hatzivassiloglou & McKeown, 1997; Wiebe, 2000; Turney, 2002; Kamps et al., 2004; Hu & Liu, 2004; Ding & Liu, 2007). Another interesting approach is that of Esuli and Sebastiani (2005). They propose a new method based on the assumption that terms with similar orientation tend to have similar glosses. They use a semi-supervised learning algorithm to classify terms as positive or negative. In another study, Ding and Liu improved the previous system proposed by Hu and Liu by assigning a score to opinion words located near the feature. The score depends on the distance between the opinion word and the feature, with a low

[4] http://wordnet.princeton.edu

score given to the opinion words far from the feature. A common approach in sentiment analysis is to employ supervised machine learning methods to acquire prominent features of sentiment. However, the success of these methods depends on the domain, topic, and time-period represented by the training data.

On the other hand, Pang, Lee, and Vaithyanathan (2002) applied machine learning methods such as NB, ME, and SVM on movie reviews to determine their polarity. The data were downloaded from the Internet Movie Database (IMDb)[5]. They used 700 negative and 700 positive reviews. To apply machine learning algorithms on the documents, the standard bag of features framework was used in this work, predefining a set of features that could appear in a document. They also treated the effect of the negation by adding the negation prefix "*not.*" The word position and the part-of-speech (POS) were also taken into account. They performed several experiments using different n-grams techniques, and the results showed that the use of unigram was the most effective method. In addition, they found that SVM outperforms NB and ME algorithms.

Mullen and Collier (2004) worked on the same dataset used by Pang et al. (2002). They calculated the average rating for the whole collection; the reviews under this average rating were classified as negative and those above the average rating were classified as positives. They investigated several features including various combinations of the Turney value, the three text-wide Osgood values (Osgood, Suci, & Tannenbaum, 1957), word unigrams, or lemmatized unigrams. In addition, they performed experiments on a movie reviews corpus downloaded from the Pitchfork Media[6]. In this case, they extracted the same features and extra features based on the movie domain. The machine learning algorithm used was SVM. They concluded that the combination of unigrams and lemmatized unigrams outperforms the models that do not use this kind of information.

Finally, Prabowo and Thelwall (2009) applied SVM with combined methods to classify reviews from different corpora. One of these datasets was the same as that used by Pang and Lee (2004) and it included 1,000 positive and 1,000 negative samples. Several classifiers were used: General Inquirer Based Classifier (GIBC), Rule-Based Classifier (RBC), Statistics Based Classifier (SBC), and SVM. They accomplished a hybrid classification, whereby if one classifier fails to classify a document, then the classifier passes the document unto the next classifier until the document is correctly classified or no other classifier remains. The results indicated that SBC and SVM improve their effectiveness in the hybrid classification.

Non-English Sentiment Analysis

Most research in OM has focused on English texts, and there is little work using other languages. The main reason for this is the lack of resources oriented to analysis sentiments in other idioms. Generating these resources is very time-consuming and labor-consuming. However, the number of comments, opinions and reviews in all languages is increasing exponentially on the Internet.

According to Mihalcea, Banea, and Wiebe (2007), there are two main approaches in the context of multilingual sentiment analysis:

- Lexicon-based approach, in which a target-language subjectivity classifier is generated by translating an existing lexicon into another idiom.
- Corpus-based approach, in which a subjectivity-annotated corpus for the target language is built through projection, training a statistical classifier on the resulting corpus.

There are some interesting papers that have studied the problem using non-English collections. For example, Denecke (2008) worked on German comments collected from Amazon. These reviews were translated into English using standard machine translation software, and then the translated reviews were classified as positive or negative, using three different classifiers: LingPipe[7], SentiWordNet (Esuli & Sebastiani, 2006b) with classification rule, and SentiWordNet with machine learning. Denecke worked on three different corpora to compare the results:

- The multiperspective question answering (MPQA) corpus[8], in English.
- 1,000 positive and 1,000 negative reviews in English from IMDb.
- 100 positive and 100 negative reviews in German from Amazon.

The experiments carried out for German language were based on translating the reviews into English and then classifying them. They used the IMDb corpus as training data and the dataset translated into English as testing data. Zhang, Zeng, Li, Wang, and Zuo (2009) applied Chinese sentiment analysis on two datasets. In the first one, euthanasia reviews were collected from different websites, while in the second dataset, six product categories were collected from Amazon (Chinese reviews). The euthanasia dataset was manually reviewed and classified into 502 positive and 349 negative articles for training. All the articles were used for testing sentiment analysis approaches, and the standard 10-fold cross-validation was chosen for evaluation. The Amazon dataset was distributed as 310,390 positive and 29,540 negative opinions for the six products. They randomly selected 200 positive and 200 negative reviews for each product to balance the distribution of two classes (positive/negative) for the training dataset. From the remaining comments, 500 positive and 500 negative reviews from each category were randomly selected for testing. The experiments were run using rule-based and machine learning approaches (SVM, NB, and decision tree). Ghorbel and Jacot (2010) used a corpus with movie reviews

[5] http://www.imdb.com

[6] http://www.pitchforkmedia.com

[7] http://alias-i.com/lingpipe

[8] http://www.cs.pitt.edu/mpqa/databaserelease

DOI: 10.1002/asi

in French. They applied a supervised classification combined with SentiWordNet to determinate the polarity of the reviews. Agić, Ljubešić, and Tadić (2010) presented a manually annotated corpus with news on the financial market in Croatia.

Regarding the OM in a multilingual framework using several languages, Ahmad, Cheng, and Almas (2006) performed a local grammar approach for three idioms using financial news: Arabic, Chinese, and English. They selected and compared the distribution of words in a domain-specific document with the distribution of words in a general corpus. Abbasi, Chen, and Salem (2008) accomplished a study for sentiment classification on English and Arabic inappropriate content. Specifically, they applied their methodologies on a U.S. supremacist forum for English and a Middle Eastern extremist group for Arabic language. Boldrini, Balahur, Martínez-Barco, and Montoyo (2009) aimed to build up a corpus with a fine-gained annotation scheme for the detection of subjective elements. The data were collected manually from 300 blogs in three different languages: Spanish, Italian, and English. Text was collected on three different topics, gathering 100 texts for each topic, with a total of 30,000 words approximately for each language.

OCA

In this article, we present OCA, a new Arabic resource made available to the scientific community that can be used in sentiment analysis[9]. First, we explain the difficulty of finding Arabic opinions because of the lack of websites that include reviews and comments using this language. Second, the process followed to generate the OCA corpus is expounded.

Difficulty in Arabic Websites

Despite the importance of the Arabic language on the Internet, there are very few web pages that specialize in Arabic reviews. In fact, our first attempt to build an Arabic corpus aimed at obtaining opinions for typical objects such as electronic products or cars, but, unfortunately, we had little success because of the lack of websites like Amazon or Booking[10] using Arabic. The most common Arabic opinion sites on the Internet are related to movies and films, although these blogs also present several obstacles to their being used in sentiment analysis tasks. Some of these difficulties are stated below:

- Nonsense and nonrelated comments. Many reviews in different web pages are not related to the topic. People attempt to comment on anything, even with unrelated words or nonsense. For instance, instead of comment an item, the user just types a word:

  Thaaaaaaanks = مشكوووووور

TABLE 1. Different variants of Roman alphabet transcriptions.

English	*Qatar is a great country*
Arabic	قطر دولة عظيمة
Roman alphabet 1	*Qatar dawla athema*
Roman alphabet 2	*Qatr dawlah 3 athema*
Roman alphabet 3	*9atar dawlah 3 athemah*

- Romanization of Arabic. Many comments use the Roman alphabet. Each phoneme in Arabic can be replaced by its counterpart in the Roman alphabet. This can be because of nonuse of Arabic keyboards for people who comment on Arabic topics from abroad. For instance, Table 1 shows a fragment explaining the problem of commenting on a topic using the Roman alphabet. There are also possible variants in the case of Romanization of Arabic for the above example, taking into account the diacritics in the Arabic language. However, a native speaker could still understand this sentence.
- Comments in different languages. It is also possible to find international languages in Arabic web pages, so you could read comments in English, Spanish, or French mixed with Arabic sentences.

Corpus Generation

To generate the OCA we have extracted the reviews from different web pages about movies. OCA comprises 500 reviews in Arabic, of which 250 are considered as positive reviews and the other 250 as negative opinions. This process involved collecting reviews from several Arabic blog sites and web pages using a simple bash script for crawling. Then, we removed HTML tags and special characters, and spelling mistakes were corrected manually. Next, a processing of each review was carried out, which involved tokenizing, removing Arabic stop words, and stemming and filtering those tokens whose length was less than two characters. Specifically, we have used the Arabic stemmer from the Rapid Miner[11] software. Rapid Miner includes two implementations of Arabic stemming: the basic Arabic stemmer, which is based on Khoja Arabic stemmer (Khoja & Garside, 1999), and the light Arabic stemmer developed by Larkey, Ballesteros, and Connell (2007). In our experiments, we have used only the basic Arabic stemmer of Rapid Miner and the Arabic stop word list provided by the same software. Finally, three different n-gram schemes are generated (unigrams, bigrams, and trigrams) and cross validation is applied to evaluate the corpus. Figure 2 shows the different steps followed in our approach. Table 2 shows an example of generation of unigram, bigrams, and trigrams for a fragment from an original review of the OCA corpus, using the Rapid Miner software and removing the stop words previously with the same tool.

Table 3 presents the number of reviews according to negative or positive classification from each web page, the name of the web page, and the highest score used in the rating system. On the other hand, Figure 3 shows an excerpt from a

[9] The OCA corpus is freely available at the SINAI website http://sinai.ujaen.es/wiki/index.php/OCA_Corpus_(English_version)

[10] http://www.booking.com

[11] http://rapid-i.com

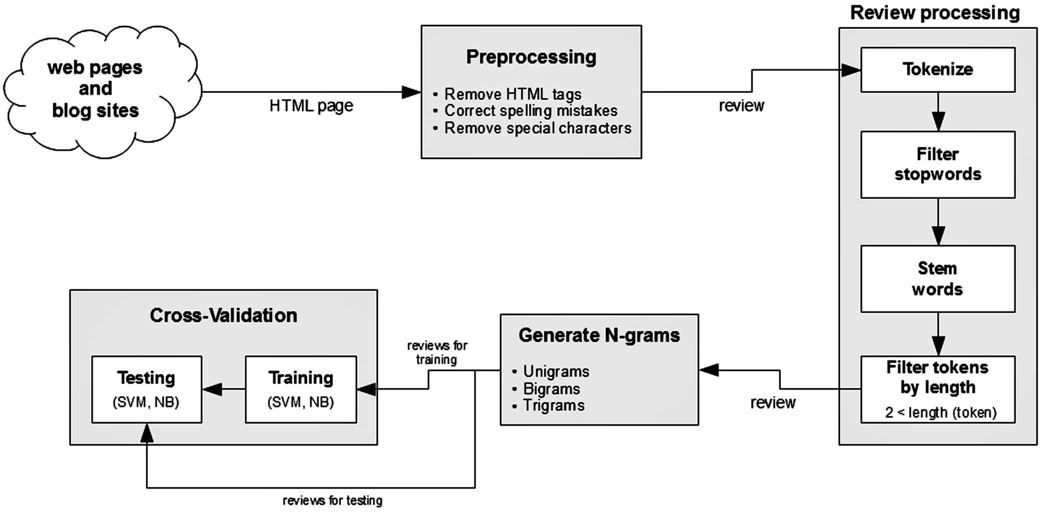


FIG. 2. Steps followed in the generation and validation of the OCA corpus.

TABLE 2. Examples of generation of unigram, bigrams, and trigrams for a fragment from an original review of the opinion corpus for Arabic.

Fragment from an original review	أداء الممثلين كان رائعا.. من قامت بدور هايدي تميزت جدا وأبدعت، ومن قامت بدور دون أبدعت أيضا، مع أنه لوحظ على أدائها التكلف، لكنه هامشي جدا إذا ما قورن بعمر الطفلة.
Unigram	أداء الممثلين رائعا قامت بدور هايدي تميزت وأبدعت قامت بدور أبدعت لوحظ أدائها التكلف هامشي قورن بعمر الطفلة.
Bigrams	وأساسها_أداء أداء أداء_الممثلين الممثلين الممثلين_رائعا رائعا رائعا_قامت قامت قامت_بدور بدور بدور_هايدي هايدي هايدي_تميزت تميزت تميزت_وأبدعت وأبدعت وأبدعت_قامت قامت قامت_بدور بدور بدور_أبدعت أبدعت أبدعت_لوحظ لوحظ لوحظ_أدائها أدائها أدائها_التكلف التكلف التكلف_هامشي هامشي هامشي_قورن قورن قورن_بعمر بعمر بعمر_الطفلة الطفلة الطفلة_الموسيقى.
Trigrams	العائلة_وأساسها العائلة_وأساسها_أداء وأساسها وأساسها_أداء وأساسها_أداء_الممثلين أداء أداء_الممثلين أداء_الممثلين_رائعا الممثلين الممثلين_رائعا الممثلين_رائعا_قامت رائعا رائعا_قامت رائعا_قامت_بدور قامت قامت_بدور قامت_بدور_هايدي بدور بدور_هايدي بدور_هايدي_تميزت هايدي هايدي_تميزت هايدي_تميزت_وأبدعت تميزت تميزت_وأبدعت تميزت_وأبدعت_قامت وأبدعت وأبدعت_قامت وأبدعت_قامت_بدور قامت قامت_بدور قامت_بدور_أبدعت بدور بدور_أبدعت بدور_أبدعت_لوحظ أبدعت أبدعت_لوحظ أبدعت_لوحظ_أدائها لوحظ لوحظ_أدائها لوحظ_أدائها_التكلف أدائها أدائها_التكلف أدائها_التكلف_هامشي التكلف التكلف_هامشي التكلف_هامشي_قورن هامشي هامشي_قورن هامشي_قورن_بعمر قورن قورن_بعمر قورن_بعمر_الطفلة بعمر بعمر_الطفلة بعمر_الطفلة_الموسيقى الطفلة الطفلة_الموسيقى الطفلة_الموسيقى_والموسيقى.

DOI: 10.1002/asi

TABLE 3. Distribution of reviews crawled from different web pages.

	Name	Web page	Rating system	Positive reviews	Negative reviews
1	Cinema Al Rasid	http://cinema.al-rasid.com	10	36	1
2	Film Reader	http://filmreader.blogspot.com	5	0	92
3	Hot Movie Reviews	http://hotmoviews.blogspot.com	5	45	4
4	Elcinema	http://www.elcinema.com	10	0	56
5	Grind House	http://grindh.com	10	38	0
6	Mzyondubai	http://www.mzyondubai.com	10	0	15
7	Aflamee	http://aflamee.com	5	0	1
8	Grind Film	http://grindfilm.blogspot.com	10	0	8
9	Cinema Gate	http://www.cingate.net	bad/good	0	1
10	Emad Ozery Blog	http://emadozery.blogspot.com	10	0	1
11	Fil Fan	http://www.filfan.com	5	81	20
12	Sport4Ever	http://sport4ever.maktoob.com	10	0	1
13	DVD4ArabPos	http://dvd4arab.maktoob.com	10	11	0
14	Gamraii	http://www.gamraii.com	10	39	0
15	Shadows and Phantoms	http://shadowsandphantoms.blogspot.com	10	0	50
Total				250	250

ليس هناك الكثير من الإهتمام الذي يستطيع الفيلم بثّه لمشاهديه. إنه مثل مقال حول
خطاب مهم مكتوب بلغة لا تجسّد تلك الأهمية. مليء بالمشاهد الدالّة لكنها غير
المؤثرة خصوصاً وأن الفيلم لا يُشيد زوجين سعيدين فعلاً من البداية، فيبقى الحب
بينهما مسألة نظرية او افتراضية.

FIG. 3. Example of an excerpt from a comment of the OCA corpus.

comment of the OCA corpus, which could be translated as follows:

> There is not much of interest in the film, which can be broadcasted for viewers. It is like an article on an important speech written in a language that does not reflect that importance. The movie is filled with scenes, but it is not influential, especially since the film does not describe a happy couple from the beginning, and love remains between them a theoretical or hypothetical issue.

The selection of the web pages was based on the quality of the language used, because many sites use slang, making understanding difficult for many Arabic speakers. Most of Arabic dialects can be understood in different Arabic countries except some specific cases such as some Moroccan dialects. Therefore, for generating the OCA corpus, we have used the reviews provided by the web pages shown in Table 3, without discarding or filtering any comment from them. However, previously, we carried out an in-depth analysis of these blogs to ensure that the dialects used in all comments were understandable by Arabic native speakers. On the other hand, there are important issues that must be taken into account in these blogs:

- Rating system. We found that there is no common system of rating among these blogs. Some of them use a rating scale of 10 points, so reviews with less than five points are classified as negative, while those with a rating between 5 and 10 points are classified as positive. Other blogs use a 5-rating scale. In these cases, we considered the movies with three, four and five points as positive, while those with less than three points were classified as negative. This classification was based on a deep study of the reviews that were rated as neutral. Finally, we also found binary classifications such as *good* or *bad*.
- Cultural and political emotions. We noticed that the culture in Arabic countries could also affect the behavior of the reviewers. For instance, an "Antichrist" movie is rated with 1 point out of 10 in one of the Arabic blogs (clearly, a negative opinion), while the same movie on the IMDb is rated at 6.7 out of 10.
- Movie and actor names in English. There are different ways of naming movies and actors in the reviews. In some cases, the names are translated into Arabic, while others keep the names in English and the reviews in Arabic.

Finally, another important factor in preparing this corpus was the richness of the text. We tried to select reviews that have more tokens than short text reviews. Table 4 shows some statistics on the OCA corpus.

Experimental Study Using OCA

Several experiments have been accomplished to evaluate the OCA corpus. We have used cross-validation to compare the performance of two of the most widely used learning algorithms: SVM and NB. Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a

TABLE 4. Statistics on the opinion corpus for Arabic.

	Negative	Positive
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

model and the other used to validate the model (Manning & Schutze, 1999). The basic form of cross-validation is k-fold cross-validation. In k-fold cross-validation, the data are first partitioned into k equally sized segments or folds. Subsequently, k iterations of training and validation are performed so that within each iteration a different fold of the data is held out for validation, while the remaining k-1 folds are used for learning. In our experiments, the 10-fold cross-validation (k=10) has been used to evaluate the classifiers.

On the other hand, evaluation has been carried out on three main measures: precision (P), recall (R), and accuracy (Acc):

$$precision(P) = \frac{TP}{TP + FP}$$

$$recall(R) = \frac{TP}{TP + FN}$$

$$accuracy(Acc) = \frac{TP + TN}{TP + FP + FN + TN}$$

where TP (true positives) are those assessments in which system and human expert agree for a label assignment, FP (false positives) are those labels assigned by the system that does not agree with the expert assignment, FN (false negatives) are those labels that the system failed to assign as they were given by the human expert, and TN (true negatives) are those nonassigned labels that were also discarded by the expert. The precision tells us how well the labels are assigned by our system (the fraction of assigned labels that are correct). The recall measures the fraction of expert labels found by the system. Finally, accuracy combines both precision and recall, calculating the proportion of true results (both true positives and true negatives; Sebastiani, 2002).

Machine Learning Algorithms

In our experiments, we used two different machine learning algorithms: NB and SVM.

NB is a method of classification based on the Bayes theorem. The major idea of the NB is to use the assumption that predictor variables are independent random variables. This assumption makes it possible to compute probabilities required by the Bayes formula from a relatively small training set. Despite its simplicity and the fact that its conditional independence assumption clearly does not hold in real-world situations, NB-based text categorization still tends to perform surprisingly well (Lewis, 1998). Indeed, Pazzani and Domingos (1997) show that NB is optimal for certain problem classes with highly dependent features. Esuli and Sebastiani (2006a) used NB to determine term subjectivity and term orientation for OM. They also applied other learning algorithms such as SVM or Rocchio, but better results were obtained using NB.

On the other hand, SVM have been shown to be highly effective in traditional text categorization, generally outperforming NB (Joachims, 1998). SVM have been applied successfully in many text classification tasks because of their principal advantages: First, they are robust in high dimensional spaces; second, any feature is relevant; third, they are robust when there is a sparse set of samples; and, finally, most text categorization problems are linearly separable. In addition, SVM have achieved good results in OM and this algorithm has overcome other machine learning techniques (O'Keefe & Koprinska, 2009).

Experiments and Results

For the experiments, we used the Rapid Miner[11] software with its text mining plug-in, which contains different tools designed to assist in the preparation of text documents for mining tasks (tokenization, stop word removal, and stemming, among others). Rapid Miner is an environment for machine learning and data mining processes that includes a cross-validation process to estimate the performance of several learning operators such as SVM or NB. As mentioned above, the 10-fold cross-validation was used to test the classifiers. We applied the Arabic stemming algorithm included in Rapid Miner to reduce words to their common root or stem. The Arabic stop words list included in Rapid Miner was also applied to the texts of the corpus to remove those words without relevant meaning.

On the other hand, a study of different n-gram schemes was also carried out to analyze its influence on the corpus generated. For this reason, we applied several n-gram models (unigram, bigrams, and trigrams) for each learning algorithm in the cross-validation process. In addition, we have evaluated the use of two different weighting schemes in the validation process: tf–idf (term frequency–inverse document frequency) and tf (term frequency). These schemes are often used in information retrieval and text mining. The impact of using stemming in the text preprocessing was also analyzed. Therefore, a total of 24 experiments were carried out on OCA corpus, 12 experiments using tf–idf as weighting scheme and the other ones using tf:

- Unigram, bigrams, and trigrams using SVM or NB as learning algorithms with stemmer,
- Unigram, bigrams, and trigrams using SVM or NB as learning algorithm without stemmer.

Table 5 and Table 6 show the results obtained in the validation process using tf–idf and tf weighting schemes respectively. Comparing the two learning algorithms used in the cross-validation process, SVM slightly improves on the performance of NB. The improvement between the best

TABLE 5. Ten-fold cross-validation results using term frequency–inverse document frequency as weighting scheme.

		Precision		Recall		Accuracy	
n-gram model	Stemming	SVM	NB	SVM	NB	SVM	NB
Unigram	Yes	0.8614	0.8106	0.8800	0.8880	0.8680	0.8380
	No	0.8699	0.8274	0.9480	**0.9520**	0.9020	0.8740
Bigrams	Yes	0.8685	0.8353	0.9080	0.9040	0.8840	0.8600
	No	0.8738	0.8525	0.9520	0.9480	0.9060	0.8900
Trigrams	Yes	0.8721	0.8361	0.9120	0.9080	0.8880	0.8620
	No	**0.8738**	**0.8525**	**0.9520**	0.9480	**0.9060**	**0.8900**

Note. SVM = support vector machine; NB = Naïve Bayes.

TABLE 6. Ten-fold cross-validation results using term frequency as weighting scheme.

		Precision		Recall		Accuracy	
n-gram model	Stemming	SVM	NB	SVM	NB	SVM	NB
Unigram	Yes	0.8701	0.7999	0.9440	0.8560	0.9000	0.8180
	No	0.8690	0.8104	0.9320	**0.9360**	0.8940	0.8560
Bigrams	Yes	0.8710	0.8275	0.9520	0.8880	0.9040	0.8460
	No	0.8690	0.8404	0.9320	0.9240	0.8940	0.8720
Trigrams	Yes	**0.8710**	0.8275	**0.9520**	0.8880	**0.9040**	0.8460
	No	0.8535	**0.8434**	0.9360	0.9240	0.8860	**0.8740**

Note. SVM = support vector machine; NB = Naïve Bayes.

accuracy results of both models is 1.8% for SVM using tf–idf as weighting scheme and 3.43% using tf. This behavior is similar to that obtained by Pang et al. (2002). Regarding the n-gram model, we can note clearly that trigram and bigram models overcome the unigram model. According to the SVM results, it should be noted that for bigram and trigram models there are no differences using stemming and the tf weighting scheme. Identical behavior is observed when we use tf–idf but without applying stemming. The use of a stemmer in the preprocessing phase will depend on the weighting scheme used. For tf–idf, it is clear that the best solution is not to stem the words. However, for tf, it depends on the learning algorithm selected. If we use SVM, we will always achieve better results by applying stemming, while if we use NB, then the best option is not to use stemming. Finally, the comparison between both weighting schemes is not relevant. tf–idf slightly improves the best result achieved by tf regarding accuracy measure (0.22%). On the other hand, the high values obtained for accuracy during the validation process show the good quality of the corpus proposed (0.90 using both weighting schemes and SVM with trigram model).

According to Kanaan et al. (2009), the results of applying different text classification techniques using Arabic language are comparable to the results obtained for English and other languages. To contrast the results obtained with OCA, we have compared them with similar experiments using the corpus generated by Pang et al. (2002). This corpus is also a collection of 1,400 samples (700 positive and 700 negative) of movie reviews. Table 7 shows the results obtained with Pang's corpus using 10-fold cross-validation and SVM, compared with our best results obtained with OCA using tf–idf, SVM and without applying stemmer in the preprocessing phase.

TABLE 7. Pang corpus 10-fold cross-validation results compared to OCA corpus best results (using tf–idf, SVM, and without stemming).

Corpus	n-gram model	Precision	Recall	Accuracy
Pang	Unigram	0.8493	0.8390	0.8445
	Bigrams	0.8583	0.8450	0.8515
	Trigrams	0.8619	0.8450	0.8535
OCA	Unigram	0.8699	0.9480	0.9020
	Bigrams	0.8738	0.9520	0.9060
	Trigrams	**0.8738**	**0.9520**	**0.9060**

Note. OCF = opinion corpus for Arabic; tf–idf = term frequency–inverse document frequency; SVM = support vector machine.

Analyzing the best results obtained with both corpus, related to the accuracy measure and 10-fold cross-validation, we can observe that the best result (0.90) using SVM over the OCA improves on the best result obtained with the Pang corpus (0.8535), using trigrams to generate the word vectors. This improvement is 5.45%. Moreover, it should be noted that for both corpora, the use of the trigram and bigram models overcomes the use of unigram model.

Conclusions and Further Work

In this work, we have generated a new Arabic corpus for predicting sentiment polarity. Nowadays, it is difficult to find a corpus designed for implementing sentiment analysis application and, more specifically, for the Arabic language.

Few blogs are oriented to expressing opinions in Arabic. Finding web pages in Arabic about topics such as electronic products, books, or cars is almost impossible. The data for the proposed corpus were collected from several blogs of movies reviews, obtaining a total of 500 comments (250 positive and 250 negative). Some experiments were also carried out on the proposed corpus to evaluate classifiers trained for determining the polarity of a review. The results obtained were very promising.

For further work, we will continue in this line of research by improving our corpus using techniques such as enlarging or fine-grained annotation. Moreover, we will focus on some linguistic features (adjectives, nouns, etc.) using WordNet for Arabic along with English resources like SentiWordNet. Furthermore, it would be worthwhile to translate this corpus into English using standard machine translation software and evaluate it with SVM and NB to analyze the results.

Acknowledgments

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), project TEXT-COOL 2.0 (TIN2009-13391-C04-02) from the Spanish Government, a grant from the Andalusian Government, project GeOasis (P08-TIC-41999) and Geocaching Urbano research project (RFC/IEG2010). Also, another part of this project was funded by Agencia Española de Cooperación Internacional para el Desarrollo MAEC-AECID.

References

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

Agic, Z., Ljubešic, N., & Tadić, M. (2010). Towards sentiment analysis of financial texts in croatian. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, … D. Tapias (Eds.), Language resources and evaluation (LREC). Paris: European Language Resources Association.

Ahmad, K., Cheng, D., & Almas, Y. (2006). Multi-lingual sentiment analysis of financial news streams. Proceedings of Science (GRID '06). Retrieved from http://cdsweb.cern.ch/record/964964/files/001GRID2006_001.pdf

Ahmed, F., & Nürnberger, A. (2009). Evaluation of n-gram conflation approaches for Arabic text retrieval. Journal of the American Society for Information Science and Technology, 9(2), 1448–1465.

Boldrini, E., Balahur, A., Martínez-Barco, P., & Montoyo, A. (2009). Emotiblog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. In R. Stahlbock, S.F. Crone & S. Lessmann (Eds.), DMIN (pp. 491–497). Las Vegas, NV: CSREA Press.

Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis. ICDE Workshops (pp. 507–512). Washington, DC: IEEE Computer Society.

Ding, X., & Liu, B. (2007). The utility of linguistic rules in opinion mining. In W. Kraaij, A.P. de Vries, C.L.A. Clarke, N. Fuhr, & N. Kando (Eds.), Proceedings of the 30th International Conference on Research and Development in Information Retrieval (ACM SIGIR '07) (pp. 811–812). New York: ACM Press.

Duwairi, R.M. (2006). Machine learning for Arabic text categorization. Journal of the American Society for Information Science and Technology, 57(8), 1005–1010.

Duwairi, R., Al-Refai, M.N., & Khasawneh, N. (2009). Feature reduction techniques for Arabic text categorization. Journal of the American Society for Information Science and Technology, 60(11), 2347–2352.

Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. In O. Herzog, H. Schek, N. Fuhr, A. Chowdhury, & W. Teiken (Eds.), Proceedings of the 14th ACM International Conference on Information and Knowledge Management (pp. 617–624). New York: ACM Press.

Esuli, A., & Sebastiani, F. (2006a). Determining term subjectivity and term orientation for opinion mining. In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL '06) (pp. 193–200). East Stroudsburg, PA: Association for Computational Linguistics.

Esuli, A., & Sebastiani, F. (2006b). SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC '06) (pp. 417–422). Paris: European Language Resources Association (ELRA).

Ghorbel, H., & Jacot, D. (2010, June). Sentiment analysis of French movie reviews. Paper presented at the Fourth international Workshop on Distributed Agent-based Retrieval Tools (DART '10), Geneva, Switzerland.

Hatzivassiloglou, V., & McKeown, K.R. (1997). Predicting the semantic orientation of adjectives. In Proceedings of the Joint ACL/EACL Conference (pp. 174–181). Morristown, NJ: Association for Computational Linguistics.

Hatzivassiloglou, V., & Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING '00) (pp. 299–305). Morristown, NJ: Association for Computational Linguistics.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 168–177). New York: ACM Press.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. European Conference on Machine Learning (ECML '98) (pp. 137–142). London: Springer-Verlag.

Kamps, J., Marx, M., Mokken, R.J., & Rijke, M.D. (2004). Using WordNet to measure semantic orientation of adjectives. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC '04) (pp. 1115–1118). Paris: European Language Resource Association.

Kanaan, G., Al-Shalabi, R., Ghwanmeh, S.H., & Al-Ma'adeed, H. (2009). A comparison of text-classification techniques applied to Arabic text. Journal of the American Society for Information Science and Technology, 60(9), 1836–1844.

Khoja, S., & Garside, R. (1999). Stemming Arabic text (Tech. rep.). Computer Department, Lancaster University, Lancaster.

Larkey, L., Ballesteros, L., & Connell, M. (2007). Light stemming for Arabic information retrieval. In A. Soudi, A. Van den Bosch, & G. Neumann (Eds.), Arabic computational morphology (Vol. 38, pp. 221–243). Heidelberg, Germany: Springer.

Lewis, D.D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nedellec & C. Rouveirol (Eds.), Proceedings of the 10th European Conference on Machine Learning (ECML '98) (pp. 4–15). Heidelberg, Germany Springer-Verlag.

Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. Cambridge, MA: MIT Press.

Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In Proceedings of the Association for Computational Linguistics (ACL '07) (pp. 976–983). East Stroudsburg, PA: Association for Computational Linguistics.

Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '04) (pp. 412–418). Morristown, NJ: Association for Computational Linguistics.

O'Keefe, T., & Koprinska, I. (2009, December). Feature selection and weighting methods in sentiment analysis. Paper presented at the 14th Australasian Document Computing Symposium, Sydney, Australia.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). The measurement of meaning. Urbana, IL: University of Illinois Press.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (pp. 271–278). Morristown, NJ: Association for Computational Linguistics.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), pp. 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '02) (pp. 79–86). Morristown, NJ: Association for Computational Linguistics.

Pazzani, M., & Domingos, P. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29 (2–3), 103–130.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3(2), 143–157.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1.

Shaalan, K.F., & Raza, H. (2009). NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology, 60(8), 1652–1663.

Turney, P.D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02) (pp. 417–424). Morristown, NJ: Association for Computational Linguistics.

Wiebe, J. (2000). Learning subjective adjectives from corpora. Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence (AAAI '00) (pp. 735–740). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Zhang, C., Zeng, D., Li, J., Wang, F.-Y., & Zuo, W. (2009). Sentiment analysis of Chinese documents: From sentence to document level. Journal of the American Society for Information Science and Technology, 60(12), 2474–2487.

2.2.2 Bilingual Experiments with an Arabic-English Corpus for Opinion Mining

- Mohammed Rushdi-Saleh, María Teresa Martín-Valdivia, Luis Alfonso Ureña López, José M. Perea-Ortega: Bilingual Experiments with an Arabic-English Corpus for Opinion Mining. Recent Advances in Natural Language Processing, RANLP 2011: 740-745.

  - Status: Published

  - Comment: We have introduced a translated corpus of Arabic movie reviews. This corpus is very important for researchers who interesting in Multilingual Sentiment Classification.

  - Number of cites: 5

Bilingual Experiments with an Arabic-English Corpus for Opinion Mining

Mohammed Rushdi-Saleh
SINAI research group
University of Jaén
msaleh@ujaen.es

M. Teresa Martín-Valdivia
SINAI research group
University of Jaén
maite@ujaen.es

L. Alfonso Ureña-López
SINAI research group
University of Jaén
laurena@ujaen.es

José M. Perea-Ortega
SINAI research group
University of Jaén
jmperea@ujaen.es

Abstract

Recently, Opinion Mining (OM) is receiving more attention due to the abundance of forums, blogs, e-commerce web sites, news reports and additional web sources where people tend to express their opinions. There are a number of works about Sentiment Analysis (SA) studying the task of identifying the polarity, whether the opinion expressed in a text is positive or negative about a given topic. However, most of research is focused on English texts and there are very few resources for other languages. In this work we present an Opinion Corpus for Arabic (OCA) composed of Arabic reviews extracted from specialized web pages related to movies and films using this language. Moreover, we have translated the OCA corpus into English, generating the EVOCA corpus (English Version of OCA). In the experiments carried out in this work we have used different machine learning algorithms to classify the polarity in these corpora showing that, although the experiments with EVOCA are worse than OCA, the results are comparable with other English experiments, since the loss of precision due to the translation is very slight.

1. Introduction

Nowadays, the interest in Opinion Mining (OM) has grown significantly due to different factors. On the one hand, the rapid evolution of the World Wide Web has changed our view of the Internet. It has turned into a collaborative framework where technological and social trends come together, resulting in the over exploited term Web 2.0. On the other hand, the tremendous use of e-commerce services has been accompanied by an increase in freely available online reviews and opinions about products and services. A customer who wants to buy a product usually searches information on the Internet trying to find other consumer analyses. In fact, web sites such as Amazon[1], Epinions[2] or IMDb[3], can affect the customer decision. Moreover, the automatic Sentiment Analysis (SA) is useful not only for individual customer but also for any company or institution. However, the huge amount of information makes necessary to accomplish new methods and strategies to tackle the problem.

Thus, SA is becoming one of the main research areas that combines Natural Language Processing (NLP) and Text Mining (TM). This new discipline attempts to identify and analyze opinions and emotions. It includes several subtasks such as subjectivity detection, polarity classification, review summarization, humor detection, emotion classification, sentiment transfer, and so on [9]. However, most of works related to OM are oriented to use English language. Perhaps due to the novelty of the task, there are very few papers analyzing the opinions using other languages different to English. In this paper, we present the experiments accomplished with an Opinion Corpus for Arabic (OCA) collected from different web pages with comments about movies. In addition, we have used automatic machine translation tools to translate OCA corpus into English. We have generated different classifiers using Support Vector Machine and Naïve Bayes in order to determinate the polarity of the opinions. The experiments carried out with the English Version of OCA (EVOCA) show that, although we lost precision in the translation, the results are comparable to other works using English texts. So, we can use this procedure in order to determine the polarity of an Arabic corpus by using English translation. This is important because most of resources are in English and we can take advantage of this situation.

The paper is organized as following: Next section presents some papers about OM using non-English language. Section 3 and Section 4 describe the OCA

[1] http://www.amazon.com
[2] http://www.epinions.com
[3] http://www.imdb.com

Proceedings of Recent Advances in Natural Language Processing, pages 740–745,
Hissar, Bulgaria, 12-14 September 2011.

corpus and its English version (EVOCA), respectively. In Section 5, accomplished experiments are showed and results are analyzed. Finally, conclusion and future work is presented.

2. Related works

Although opinions and comments in the Internet are expressed in any language, most of research in OM is focused on English texts. However, languages such as Chinese, Spanish or Arabic, are ever more present on the web[4]. Thus, it is important to develop resources for helping researcher to work with these languages.

There are some interesting papers that have studied the problem using non-English collections. For example, Denecke [5] worked on German comments collected from Amazon. These reviews were translated into English using standard machine translation software. Then the translated reviews were classified as positive or negative, using three different classifiers: LingPipe7, SentiWordNet [6] with classification rule, and SentiWordNet with machine learning.

Zhang et al. [12] applied Chinese sentiment analysis on two datasets. In the first one euthanasia reviews were collected from different web sites, while the second dataset was about six product categories collected from Amazon (Chinese reviews).

Ghorbel and Jacot [7] used a corpus with movie reviews in French. They applied a supervised classification combined with SentiWordNet in order to determinate the polarity of the reviews.

Agić et al. [2] presented a manually annotated corpus with news on the financial market in Croatia.

Boldrini et al. [4] aimed to build up a corpus with a fine-gained annotation scheme for the detection of subjective elements. The data were collected manually from 300 blogs in three different languages: Spanish, Italian and English.

Regarding opinion mining for Arabic language, Ahmad et al. [3] performed a local grammar approach for three languages: Arabic, Chinese and English using financial news. They selected and compared the distribution of words in a domain-specific document to the distribution of words in a general corpus.

Finally, Abbasi et al. [1] accomplished a study for sentiment classification on English and Arabic inappropriate content. Specifically, they applied their methodologies on a U.S. supremacist forum for English and a Middle Eastern extremist group for Arabic language.

3. OCA: Opinion Corpus for Arabic

Despite the importance of the Arabic language on the Internet, there are very few web pages which specialize in Arabic reviews. The most common Arabic opinion sites in the Internet are related to movies and films, although these blogs also present several obstacles to their being used in sentiment analysis tasks. Some of these difficulties are stated below:

- **Nonsense and non related comments**. Many reviews in different web pages are not related to the topic. People attempt to comment on anything, even with unrelated words or nonsense. For instance, instead of comment an item, the user just types a word:

  Thaaaaaaanks= **مشكوووووور**

- **Romanization of Arabic**. Many comments use the Roman alphabet. Each phoneme in Arabic can be replaced by its counterpart in the Roman alphabet. This can be due to non-use of Arabic keyboards for people who comment on Arabic topics from abroad. For instance, Table 1 shows a fragment explaining the problem of commenting on a topic using the Roman alphabet. There are also possible variants in the case of Romanization of Arabic for the above example, taking into account the diacritics in the Arabic language. However, a native speaker could still understand this sentence.

Table 1. Different variants of Roman alphabet transcriptions

English	*Qatar is a great country*
Arabic	قطر دولة عظيمة
Roman alphabet 1	*Qatar dawla athema*
Roman alphabet2	*Qatr dawlah 3athema*
Roman alphabet3	*9atar dawlah 3athemah*

- **Comments in different languages.** It is also possible to find international languages in Arabic web pages, so you could read comments in English, Spanish or French mixed with Arabic sentences.

In order to generate the Opinion Corpus for Arabic we have extracted the reviews from different web pages about movies. OCA consists of 500 reviews in Arabic, of which 250 are considered as positive reviews and the other 250 as negative opinions. This process has consisted of collecting reviews from several Arabic blog sites and web pages. Table 2 presents the number of reviews according to negative or positive classification from each web page, the name of the web page and the highest score used in the rating system.

[4] http://www.internetworldstats.com

Table 2. Distribution of reviews crawled from different web pages

	Name	web page	Rating system	PR	NR
1	Cinema Al Rasid	http://cinema.al-rasid.com	10	36	1
2	Film Reader	http://filmreader.blogspot.com	5	0	92
3	Hot Movie Reviews	http://hotmoviews.blogspot.com	5	45	4
4	Elcinema	http://www.elcinema.com	10	0	56
5	Grind House	http://grindh.com	10	38	0
6	Mzyondubai	http://www.mzyondubai.com	10	0	15
7	Aflamee	http://aflamee.com	5	0	1
8	Grind Film	http://grindfilm.blogspot.com	10	0	8
9	Cinema Gate	http://www.cingate.net	bad/good	0	1
10	Emad Ozery Blog	http://emadozery.blogspot.com	10	0	1
11	Fil Fan	http://www.filfan.com	5	81	20
12	Sport4Ever	http://sport4ever.maktoob.com	10	0	1
13	DVD4ArabPos	http://dvd4arab.maktoob.com	10	11	0
14	Gamraii	http://www.gamraii.com	10	39	0
15	Shadows and Phantoms	http://shadowsandphantoms.blogspot.com	10	0	50
			Total	250	250

We have removed HTML tags and special characters as well as spelling mistakes were corrected manually. Next, a processing of each review was carried out which consisted of tokenizing, removing Arabic stop words, stemming and filtering those tokens whose length was less than two characters. Figure 1 shows the different steps followed in our approach in order to generate the OCA corpus and Table 3 shows some statistics on such corpus.

On the other hand, there are important issues that must be taken into account in these blogs:

- **Rating system**. We found that there is no common system of rating among these blogs. Some of them use a rating scale of 10 points, so reviews with less than five points are classified as negative while those with a rating between five and 10 points are classified as positive. Other blogs use a 5-rating scale. In these cases, we considered the movies with three, four and five points as positive, while those with less than three points were classified as negative. This classification was based on a deep study of the reviews which were rated as neutral. Finally, we also found binary classifications such as *good* or *bad*.

Table 3. Statistics on the OCA opinion corpus

	Negative	Positive
Total **documents**	250	250
Total **tokens**	94,556	121,392
Total **sentences**	4,881	3,137

- **Cultural and political emotions**. Culture in Arabic countries can also affect the behavior of the reviewers. For instance, an “Antichrist” movie is rated with 1 point from 10 in one of the Arabic blogs, while the same movie on IMDb is rated at 6.7 out of 10.

- **Movie and actor names in English**. There are different ways of naming movies and actors in the reviews. In some cases, the names are translated into Arabic, while others keep the names in English and the reviews in Arabic.

4. EVOCA: English Version of OCA

In order to compare the experiment for Arabic and English, we have translated OCA into English using an automatic Machine Translation (MT) tool freely available. Specifically, we have used the online translator provided by PROMT[5].

The processing followed to carry out the translation consisted of splitting the text of the reviews in blocks of 500 characters to fit with the maximum length allowed by the online translator. Secondly, after the translation, extra UTF-8 invalid characters were removed and, finally, the translated reviews were generated from the blocks belonging to each of them. Figure 2 summarizes the processing followed to generate the EVOCA corpus.

The new corpus EVOCA contains the same number of positive and negative reviews that OCA corpus, with a total of 500 reviews. Table 4 shows some statistics for the EVOCA corpus.

Table 4. Statistics on the EVOCA opinion corpus

	Negative	Positive
Total **documents**	250	250
Total **tokens**	122,135	153,581
Avg. tokens per review	488.54	614.32
Total **sentences**	5,030	3,483
Avg. sentences per review	20.12	13.93

[5] Available at http://translation2.paralink.com

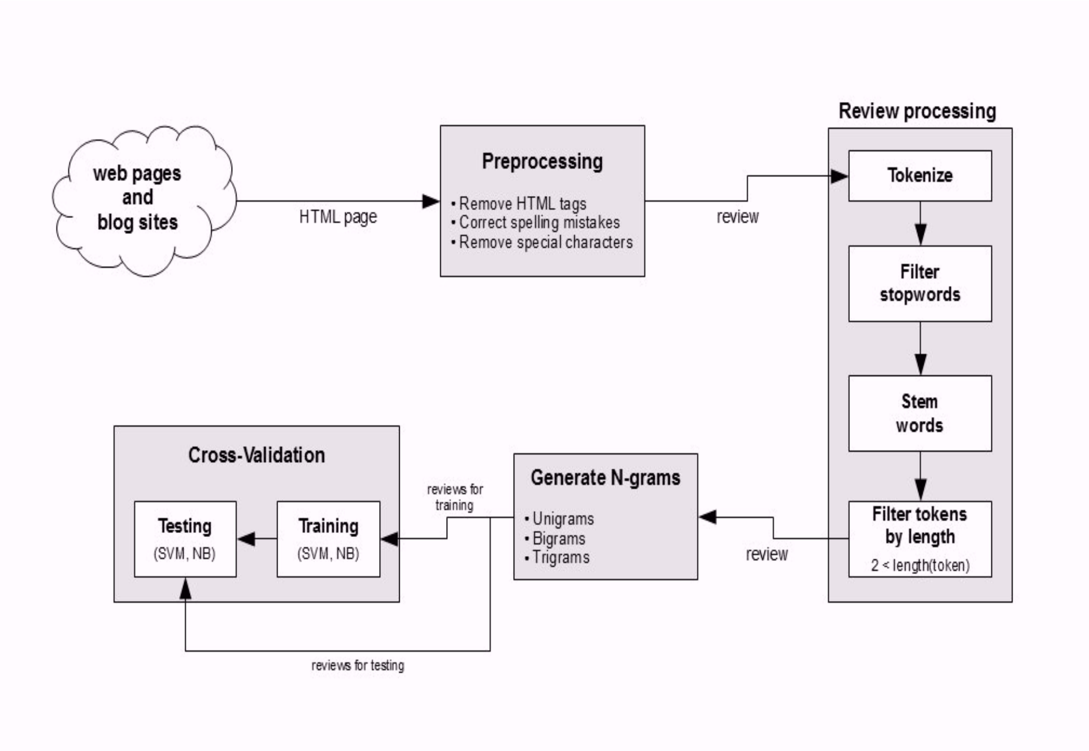


Figure 1. Steps followed in the generation and validation of the OCA corpus

5. Experiments and Results

For the experiments, we have used the Rapid Miner[6] software with its text mining plug-in which contains different tools designed to assist in the preparation of text documents for mining tasks (tokenization, stop word removal and stemming, among others). Rapid Miner is an environment for machine learning and data mining processes.

We have applied two of the most used classifiers: Support Vector Machines (SVM) and Naïve Bayes (NB).

SVM [11] is based on the structural risk minimization principle from the computational learning theory, and seek a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set.

On the other hand, NB algorithm [8] is based on the Bayes theorem. Due to its complex calculation, the algorithm has to make two main assumptions: first, it considers the Bayes denominator invariant, and second, it assumes that the input variables are conditional independence.

In our experiments, the 10-fold cross-validation has been used in order to evaluate the classifier. This evaluation has been carried out on three main measures: precision (P), recall (R) and F1 measure [10].

Moreover, for each machine learning algorithm, we have analyzed how the use of stemmer affects the experiments. TF·IDF has been used as weighting scheme. We have also accomplished several experiments using different n-grams models. However, the obtained results with bi-grams and trigrams were very similar to unigrams. For this reason we have only shown the best results obtained with unigrams. Results for SVM and NB are shown in Table 5 and Table 6, respectively.

As we can see, taking into account the F1 measure, all the experiments with OCA overcome EVOCA except when we use SVM and stemmer. In fact, this is the only case where stemmer obtains a better result although the improvement is very slight (+1.54%). Anyway, the best result is achieved using SVM without stemmer over the OCA corpus with 0.9073 of F1 measure.

[6] http://rapid-i.com

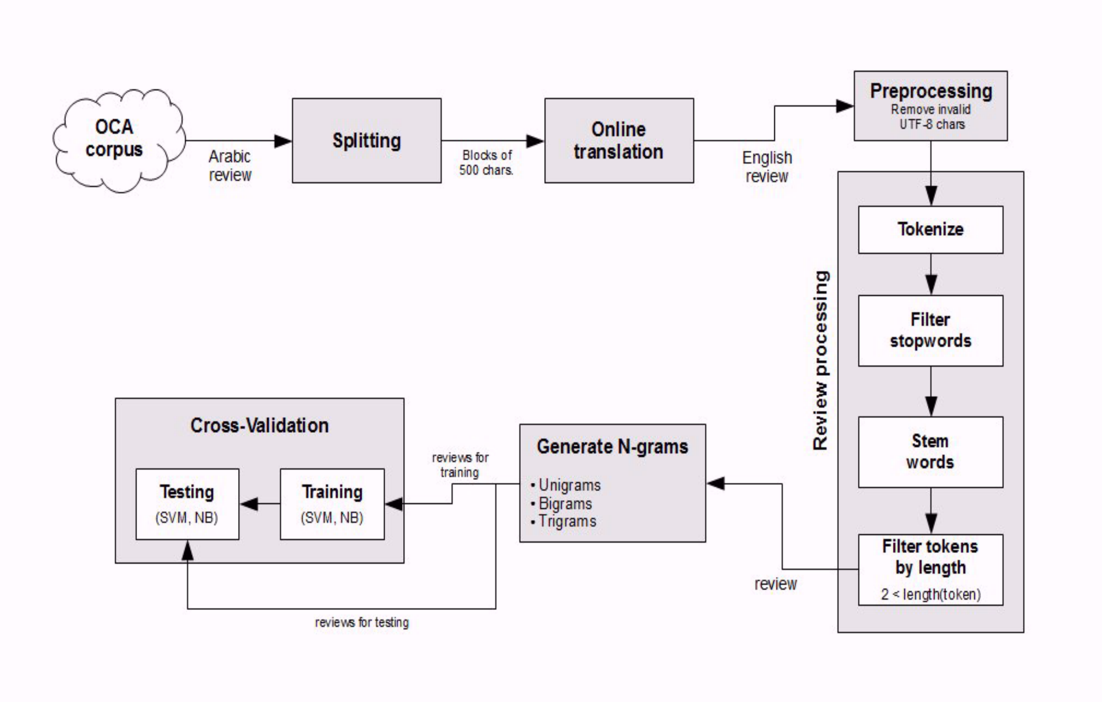


Figure 2. Processing followed to generate and validate the EVOCA corpus

However, it is interesting to note that, in the SVM experiments, the loss of precision due to the translation is very little. The highest difference is 4.31% when we do not apply stemmer, while it is 1.54% when the stemmer is applied. In general, the results with EVOCA, near to 90%, are very good comparing them with other works using SVM and English corpora [9].

Table 5. Results with SVM

	Stem	P	R	F1
OCA	Yes	0.8614	0.8800	0.8706
	No	0.8699	0.9480	**0.9073**
EVOCA	Yes	0.9007	0.8680	0.8840
	No	0.8561	0.8840	0.8698

Table 6. Results with NB

	Stem	P	R	F1
OCA	Yes	0.8106	0.8880	0.8475
	No	0.8274	0.9520	**0.8853**
EVOCA	Yes	0.7100	0.8320	0.7662
	No	0.7323	0.8640	0.7927

As regard the machine learning algorithm, it is clear that SVM works better in all cases. Taking into account the best results on the OCA corpus, SVM improves 2.49% the result obtained with NB (both without applying stemmer). On the EVOCA corpus the difference is higher for SVM +15.37% and +9.73%, using stemmer and without using it, respectively. Although the differences between SVM and NB over the OCA corpus are small, when they are applied over EVOCA, NB loses too much precision. In this case, the translation is affecting highly the results.

Finally, we have analyzed the impact of the stemmer in the experiments. As can be observed in both Table 5 and Table 6, in all cases the stemming process gets worse results except when we use SVM on the EVOCA corpus (+1.63% for stemming). For the OCA corpus, not use the stemmer always improves the results when we apply it (+4.22% using SVM and +4.46% using NB), while we obtain an improvement of 3.46% on the EVOCA corpus using NB.

6. Conclusion

In this paper we have presented an Arabic corpus for opinion mining along with its English translation. OCA and EVOCA corpora are freely available for the research community[7]. The OCA corpus is composed of Arabic reviews obtained from specialized Arabic web pages related to movies and films. Then, we have generated the EVOCA corpus, which is the English translation of the OCA corpus using an automatic machine translation tool. Both corpora include a total of 500 reviews, 250 positives and 250 negatives. In

[7] OCA and EVOCA corpora are freely available at http://sinai.ujaen.es/wiki/index.php/Recursos

addition, we have accomplished several experiments over the corpora using two different machine learning algorithms (SVM and Naïve Bayes) and applying a stemming process. The results obtained show that, although the precision with the EVOCA are lower, they are comparable with other sentiment analysis researches using English texts. This loss of precision due to the translation is very slight (-4.31% when stemmer is not applied) and therefore it is very interesting for the future because we could apply English resources for opinion mining such as SentiWorNet in order to improve the results. On the other hand, we have shown that the use of the stemming process is not recommended to work with these corpora.

7. Acknowledgments

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), project TEXT-COOL 2.0 (TIN2009-13391-C04-02) from the Spanish Government, a grant from the Andalusian Government, project GeOasis (P08-TIC-41999), and a grant from the Instituto de Estudios Giennenses, project Geocaching Urbano (RFC/IEG2010). Also, another part of this project was funded by Agencia Española de Cooperación Internacional para el Desarrollo MAEC-AECID.

8. References

[1] Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Trans. Inf. Syst. 26 (3).

[2] Agić, Z., Ljubešić, N., & Tadić, M. (2010). Towards Sentiment Analysis of Financial Texts in Croatian. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), Language Resources and Evaluation (LREC). European Language Resources Association.

[3] Ahmad, K., Cheng, D., & Almas, Y. (2006). Multilingual sentiment analysis of financial news streams. Proceedings of Science (GRID2006).

[4] Boldrini, E., Balahur, A., Martínez-Barco, P., & Montoyo, A. (2009). Emotiblog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. In R. Stahlbock, S.F. Crone & S. Lessmann (Eds.), DMIN (pp. 491-497). CSREA Press.

[5] Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis. ICDE Workshops (pp. 507–512). IEEE Computer Society.

[6] Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of the 5th Conference on Language Resources and Evaluation (LREC) (pp. 417-422).

[7] Ghorbel, H., & Jacot, D. (2010). Sentiment analysis of French movie reviews. Proceedings of the 4th international Workshop on Distributed Agent-based Retrieval Tools (DART 2010). Geneva, Italy.

[8] Mitchell, T. (1997). Machine Learning. McGraw-Hill.

[9] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2 (1-2) (pp. 1-135).

[10] Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), 1.

[11] Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, New York.

[12] Zhang, C., Zeng, D., Li, J., Wang, F.-Y., & Zuo, W. (2009). Sentiment analysis of Chinese documents: From sentence to document level. Journal of the American Society for Information Science and Technology (JASIST), 60(12), 2474–2487.

2.2.3 Comparing Machine Learning and Semantic Orientation for Polarity Detection using EVOCA

- María Teresa Martín-Valdivia, M. Perea-Ortega, Luis Alfonso Ureña López, Mohammed Rushdi-Saleh: Comparing Machine Learning and Semantic Orientation for Polarity Detection using EVOCA. Data & Knowledge Engineering.

  - Status: Submitted/ under review

  - Impact Factor: 1.422

  - Category: (COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE) Ranking: 2011: 46/111; (COMPUTER SCIENCE, INFORMATION SYSTEMS) Ranking: 2011: 41/135.

Elsevier Editorial System(tm) for Data & Knowledge Engineering
Manuscript Draft

Manuscript Number: DATAK-D-12-00039

Title: Comparing Machine Learning and Semantic Orientation for Polarity Detection using EVOCA

Article Type: Full Length Article

Keywords: Sentiment Analysis; Opinion Mining; Parallel corpora; Classification; Information Retrieval

Corresponding Author: Mr. José M. Perea-Ortega, Ph.D.

Corresponding Author's Institution: University of Sevilla

First Author: M. Teresa Martín-Valdivia

Order of Authors: M. Teresa Martín-Valdivia; José M. Perea-Ortega, Ph.D.; L. Alfonso Ureña-López; Mohammed Rushdi-Saleh

Comparing Machine Learning and Semantic Orientation for Polarity Detection using EVOCA

Abstract. Currently, there is a growing interest in Opinion Mining (OM) due to the large number of blogs, forums and social networks where people discuss many different topics. There are several works that analyze different tasks, namely polarity classification, subjectivity detection, humor appraisal and the like. Most of these papers only deal with English despite the fact that there are many other languages in which the explosion of Web 2.0 has had a spectacular increase. However, OM resources for languages other than English are scarce. In this paper, we present the EVOCA corpus: the English Version of the Opinion Corpus for Arabic (OCA). This corpus is the English translation of the OCA corpus that was presented in a previous work. Our main goal now is to compare the two main approaches to tackle the polarity detection problem: Machine Learning (ML) and Semantic Orientation (SO). For ML, Support Vector Machines (SVM) and Naïve Bayes (NB) have been applied with different configurations. For SO, we have built several corpora from the original EVOCA corpus including different combinations of adjectives, nouns, adverbs and verbs. Then, SentiWordNet (SWN) has been used in order to determine the polarity of the review. In addition, we have carried out a comparison with the best results obtained using the OCA corpus, for the machine learning experiments. The results show that although ML overcomes SO, the use of SWN could be a good alternative for polarity detection when a training corpus is not available.

Keywords: Sentiment Analysis; Opinion Mining; Parallel corpora; Classification; Information Retrieval

1. Introduction

Nowadays, the World Wide Web is the most important place to express reviews, evaluations, and sentiments. People tend to disclose their opinions and sentiments in Internet forums. For example, they protest, organize and plan their activities in Facebook, Twitter and other blogs. Their opinions on various topics can be expressed in unstructured documents, reviews, posts, comments, etc. Tackling and tracking this huge unstructured information in order to detect its polarity is attracting many researchers in the field of text mining.

The Web is a vast information resource, in which we can find two main types of information: facts and opinions. Although there are lots of issues to be resolved, the management of factual information has been extensively studied. However, the automatic processing of textual opinions is a new task closely related to text mining, which has just started to be studied. This is a challenging task known as Opinion Mining (OM), sometimes also called Sentiment Analysis (SA) [1]. This new discipline aims to identify and analyze opinions and emotions. It includes several subtasks such as subjectivity detection [2], polarity classification [3], review summarization [4], humor detection [5] or emotion classification [6] among others. Specifically, sentiment classification or polarity detection is an opinion mining activity oriented to determine which is the overall sentiment-orientation of the opinions contained within a given document. The document is supposed to contain subjective information such as product reviews or opinionated posts in blogs.

Although different approaches have been applied to the field of sentiment-polarity classification, the mainstream basically consists of two major methodologies. On the one hand, the Machine Learning (ML) approach is based on using a collection of data to train the classifiers [3]. On the other hand, the approach based on Semantic Orientation (SO) does not need prior training, but it takes into account the orientation of words, positive or negative [7]. Both methodologies have their advantages and drawbacks. For example, the ML approach requires training data, which in many cases are impossible or difficult to achieve, partially due to the novelty of the task. In opposition, the SO approach requires having lots of linguistic resources which generally depend on the language.

In this paper we have carried out experiments with both methodologies in order to check their effectiveness in an English corpus automatically translated from the Arabic corpus OCA [8]. In a previous work, we generated the OCA corpus and different ML algorithms were applied. The results were very promising, but we could not compare them to the SO approach due to the lack of Arabic resources for OM.

Although the best results are usually obtained with the ML approach using supervised learning, this methodology requires a training corpus labeled with the correct classes. The problem is that such resources are very difficult to achieve mainly due to the novelty of the task. Furthermore, most of the existing resources are oriented towards managing English texts, perhaps because of their greater availability. Examples of these resources are General Inquire[1] [9], WordNet Affect [10] or SentiWordNet[2] [11]. We consider that these resources are very valuable sources of information that we should take advantage of.

On the other hand, the proliferation of opinions in several languages different to English is exponentially increasing. In fact, less than 50% of Internet users speak English[3], and consequently, the management and study of subjectivity and sentiment analysis in languages other than English is a growing need. For example, Chinese, Spanish or Arabic are becoming very important for business and economy. From our point of view, there are two ways to address the problem when we want to apply sentiment analysis to languages aside from English:

- Generating resources for the target language. For example, in our previous work, we generated the Opinion Corpus for Arabic (OCA corpus). This corpus is a collection of 500 movie reviews, out of which 250 are labeled as positive reviews and the other 250 are considered as negative.

[1] http://www.wjh.harvard.edu/~inquirer
[2] http://sentiwordnet.isti.cnr.it
[3] http://www.internetworldstats.com/stats.htm

- Extracting information in the other language and translating it into English. Then, this information can be managed using the English resources. In our case, we have translated the OCA corpus into English generating the EVOCA corpus (English Version of OCA). Then, SWN has been applied in order to determine the opinion polarity.

The first methodology is a very promising approach as we have already proved in [8]. However, the generation of resources is a very difficult and time-consuming task. The second one is easier but usually the results are not comparable. Our main goal in this paper is to demonstrate that the translation of a non-English OM corpus and the subsequent application of English resources could be a good alternative when a labeled corpus is not available for training purposes.

In summary, this paper presents the EVOCA corpus that has been automatically generated by translating the Arabic OCA corpus into English. Then, two different kinds of experiments have been accomplished. Firstly, we have applied ML algorithms such as SVM and NB with different parameters. Secondly, we have used Semantic Orientation based on the use of the linguistic resource SWN on the EVOCA corpus. We have carried out several experiments using different corpora taking into account combinations of adjectives, nouns, adverbs and verbs. The obtained results show that ML clearly overcomes SO. However, the experiments using SWN also reveal that the results obtained applying SO could be competitive when a labeled training corpus is not available. In addition, we have carried out a comparison with the best results obtained using the OCA corpus, for the machine learning experiments.

The paper is organized as follows: the following section deals with some related works and different approaches in sentiment analysis including papers that make use of SWN and works about OM in other languages different to English. The EVOCA corpus used in our experiments is presented in Section 3, along with a brief description of the original Arabic corpus OCA. Section 4 describes the different experiments carried out applying ML and SO approaches. Finally, the results obtained and the comparison with the OCA corpus are discussed in Section 5 and main conclusions and further work are expounded in Section 6.

2. Background: Related Work

In this section we analyze some relevant works related to our paper. Firstly, we present some basic references for OM and papers about the two main approaches: ML and SO. Then, we also point out some papers that applied the lexical resource SWN. Finally, we present some works that manage corpora in languages other than English.

Although OM is a relatively new discipline, there is a considerable number of researches on this area. A good review of Opinion Mining and Sentiment Analysis can be found in [1]. This work describes some useful resources and tools for OM and also comments the main contributions in this field.

Out of the two main approaches in polarity detection, ML has been studied in more detail perhaps because it usually obtains better results. The most commonly used algorithm is Support Vector Machines (SVM) [12], although a wide range of methods has been applied. The work of Pang et al. [3] made a comparison between three algorithms (SVM, Maximum Entropy and Naïve Bayes) on a movie review corpus showing that the SVM obtained the best results although the performance was similar for all of them. These algorithms have also been successfully applied to other text mining tasks like document categorization, text summarization or information retrieval. Specifically, the SVM and NB algorithms are used both in the earlier OCA-based work and in the present EVOCA-based paper.

In the Semantic Orientation (SO) approach, which applies manually crafted rules and lexicon, the document is represented as a collection of words. Then, the sentiment of each word can be determined by different methods, for example, using a list of opinionated words [13], applying web search [14], making use of annotated terms in dictionaries [15], or lexical resources such as General Inquirer [16] or WordNet [17]. Moreover, some authors have applied a hybrid architecture which combines both approaches in order to improve the classification effectiveness [18].

In this paper we have used SWN in order to apply the SO approach. Different researchers have focused on integrating this lexical resource with polarity detection. For example, Devitt and Ahmad [19] applied SWN along with WordNet in order to determine the polarity of financial news. Chaumartin [20] combined SWN and WordNet-Affect to enrich a news headline corpus developing a rule-based system. In film reviews, sentiment polarity integrating SWN was studied by Ohana and Tierney [21], also considering the negation. Saggion and Funk [22] make use of SWN for opinion classification comparing the results between short and large textual reviews in business.

On the other hand, although most of the works carried out in this area use a set of data, chiefly English texts, there are also some researches studying the use of other languages. For example, Kim and Hovy [23] applied different OM techniques on a German email corpus. Zhang, Zeng, Li, Wang, and Zuo [24] applied Chinese SA on two datasets. The experiments were run using rule-based and machine learning approaches (SVM, Naïve Bayes, and Decision Tree). Ghorbel and Jacot [25] used a corpus with movie reviews in French. They applied a supervised classification combined with SWN in order to determine the polarity of the reviews. Martínez-Cámara, Martín-Valdivia and Ureña-López [26] presented several experiments using ML algorithms (SVM, NB, BBR, KNN, C4.5) on a Spanish corpus of movie reviews. Finally, in our previous work [8], we generated an Arabic corpus with opinions extracted from several websites devoted to movie reviews. The results obtained applying different ML techniques were very promising.

As regards the multilingual research in OM, there are also some significant examples. Banea, Mihalcea, Wiebe and Hassan [27] show that automatic translation is a viable alternative for the construction of resources and tools for subjectivity analysis in a new target language. Ahmad, Cheng and Almas [28] performed a local grammar approach for three languages: Arabic, Chinese and English using financial news. They selected and compared the distribution of words in a domain-specific document to the distribution of words in a general corpus. Denecke [29] used a German corpus with Amazon product reviews to train a classifier in order to determine the polarity of the opinions. Denecke uses translation software to translate the comments from German into English and then applies SWN. The results are compared with the work of Kim and

Hovy [23] obtaining a slightly higher performance. Abbasi, Chen and Salem [30] accomplished a study for sentiment classification on English and Arabic inappropriate content. Specifically, they applied their methodologies on a U.S. supremacist forum for English and a Middle Eastern extremist group for Arabic language. Boldrini, Balahur, Martínez-Barco and Montoyo [31] built up a corpus with a fine-gained annotation scheme for the detection of subjective elements. The data were collected manually from 300 blogs in three different languages: Spanish, Italian and English. The texts were collected on three different topics, gathering 100 texts for each topic, with a total of 30,000 words approximately for each language.

3. The EVOCA corpus: a translated corpus from OCA

We have performed our experiments on a corpus which was translated from Arabic into English. The Arabic corpus is called OCA (Opinion Corpus for Arabic) and was prepared by our XXX group[4]. A detailed description of this corpus can be found in [8]. The Arabic reviews contained in OCA were crawled from several movie blogs. The lack of specialized webs in Arabic language for opinions in different domains was a real problem when we generated the corpus. Eventually, we collected 500 reviews from fifteen different web pages, consisting of 250 positive and 250 negative reviews.

Important issues had to be taken into account when we generated the OCA corpus:

- Rating system. We found different rating systems in the blogs we used to extract the opinions. Some of them used a rating scale of 10 points, other blogs used a 5-rating scale and even we also found binary classifications such as good or bad.

- Cultural and political emotions. Depending on the country of the blog, we found the same movie rated with very different scores.

- Movie and actor names in English. There were different ways of naming movies and actors in the reviews. In some cases, the names were translated into Arabic, while others kept the names in English and the reviews in Arabic.

The whole OCA corpus has been automatically translated into English using the PROMT-Online translator[5]. This new corpus has been called EVOCA (English Version of the OCA corpus) and it is freely available from our web page besides the OCA corpus[6]. Table 1 shows some statistics of the OCA and the EVOCA corpora.

[4] XXXX
[5] http://translation2.paralink.com
[6] XXXX

	Negative	**Positive**
Total **documents**	250	250
OCA Total **tokens**	94,556	121,392
OCA Total **sentences**	4,881	3,137
EVOCA Total **tokens**	122,135	153,581
EVOCA Total **sentences**	5,030	3,483

Table 1. Statistics of the OCA and the EVOCA corpora

3.1. Translation issues

The main difficulties found during the generation of the EVOCA corpus are expounded in the following lines. Although the process of translation from the OCA corpus was carried out using the PROMT-online translator, there were different facts that affected the polarity classification and they must be discussed.

Firstly, we found some limitations in the use of the online translator. For example, it only allowed translating 500 characters at a time. In this case, sometimes we had to divide the reviews into two or more parts, which affected the coherence of the translated text.

Another issue related to the influence of the translation in the polarity classification was the difference found in the results obtained from the Part Of Speech (POS) tagging after the translation process. For instance, in one of the text reviews of the OCA corpus, we can find the sentence in Arabic “ ”, which means “stigma” in English. If we translate this sentence into English word by word it must be:

: stain (Noun)
: shame (Noun)

While the real parsing in Arabic for this sentence is:

: (Noun)
: (Adjective)

In addition, the phrase “ ” in Arabic can be translated as “stigma” in English, where “stigma” is parsed as a noun. In summary, if we use the POS as feature in order to classify a text into positive or negative, we must take into account the different POS tags when we translate Arabic texts.

Finally, it is important to note some problems when we manage ironic or figurate text. Take for example the following text extracted from the OCA corpus:

“ ”

It was translated into English as:

“*Advice from me, respected, even if the mentality of the viewer onion*”

In the above sentence the reviewer used an ironic way to advice the producer of the movie to respect the mentality of people who watch the movie using the word “   ” which is translated as “onion”. In the context of the Arabic review, this is not the real meaning of the word and so the English translation cannot depict the same meaning of what the reviewer of the movie wanted to express. As a consequence, the translation of the sentence was completely wrong.

4. Experimental Framework

In this section, we explain the configuration followed for the experiments carried out in this work. On the one hand, we have applied a machine learning approach to the EVOCA corpus. Specifically, we have used different configurations for SVM and NB. On the other hand, we have applied a semantic orientation approach. For this, the information provided by the lexical resource SWN version 3.0 [32] has been integrated into the EVOCA corpus.

In order to evaluate the different approaches, we have used the traditional measures employed in text classification: precision (P), recall (R), accuracy (Acc) and F1:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1 = \frac{2PR}{P + R}$$

where TP (True Positives) are those assessments where the system and a human expert agree on a label, FP (False Positives) are those labels assigned by the system that do not agree with the expert assignment, FN (False Negatives) are those labels that the system failed to assign as they were given by the human expert, and TN (True Negatives) are those non-assigned labels that were also discarded by the expert [33].

4.1. Machine Learning applied to EVOCA

To carry out the experiments applying machine learning algorithms, we have used the Rapid Miner[7] software with its text mining plug-in, which contains different tools designed to assist in the preparation of text documents for mining tasks (namely tokenization, stop word removal and stemming, among others). Rapid Miner is an environment for machine learning and data mining processes that includes learning operators such as SVM or NB.

[7] http://rapid-i.com

The evaluation for the machine learning approach has been carried out applying the cross-validation method using SVM and NB as learning algorithms. Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model [34]. The basic form of cross-validation is k-fold cross-validation. In k-fold, cross-validation the data are first partitioned into k equally sized segments or folds. Subsequently, k iterations of training and validation are performed so that within each iteration a different fold of the data is held-out for validation while the remaining k-1 folds are used for learning. In our experiments, the 10-fold cross-validation has been used (k=10).

SVM [35] and NB [36] have been chosen as learning algorithms in this study because they are the most commonly applied to opinion mining tasks. Moreover, both of them have proved to be highly effective in traditional text categorization and have been applied successfully in many opinion mining tasks overcoming other machine learning techniques [37, 38].

For the SVM experiments, different configurations have been applied. The Rapid Miner software allows the use of several SVM kernels such as *linear*, *polynomial*, *rbf*, *sigmoid* or *precomputed*. We have tested each of them. Moreover, each kernel can be configured using different parameters. In the SVM experiments, we have used the default parameters established by Rapid Miner for each kernel:

- *Linear: C = 0.0, epsilon = 0.0010*
- *Polynomial: degree = 3, gamma = 0.0, coef0 = 0.0, C = 0.0, epsilon = 0.0010*
- *Rbf: gamma = 0.0, C = 0.0, epsilon = 0.0010*
- *Sigmoid: gamma = 0.0, C = 0.0, epsilon = 0.0010*
- *Precomputed: coef0 = 0.0, C = 0.0, epsilon = 0.0010*

For the NB experiments, only the Laplace correction has been activated for the learning algorithm. This is also the default parameter when the NB algorithm is used in the Rapid Miner software.

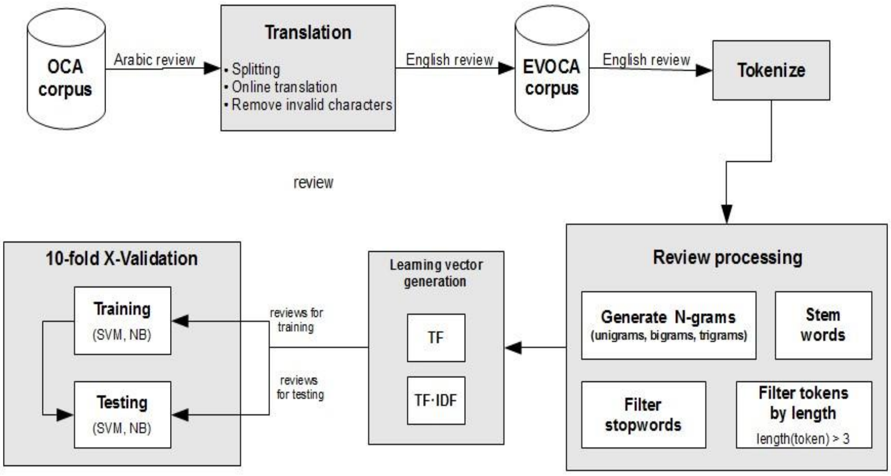


Figure 1. Overview of the machine learning approach

For both machine learning experiments (SVM and NB), we have also studied the impact of different heuristics, such as the elimination of stop words, the use of stemming, filtering of those words of less than four letters or using n-grams (unigram, bigrams and trigrams). Finally, we have analyzed the results obtained after using two different weighting schemes in order to generate the learning vectors: TF and TF·IDF. The combination of these heuristics along with the different kernels applied for the SVM has resulted in a total of 24 experiments for each kernel, i.e. 120 experiments using TF·IDF as weighting scheme and 120 experiments using TF (a total of 240 experiments for SVM). However, for NB, only 24 experiments have been carried out using TF·IDF and 24 experiments using TF (a total of 48 experiments for NB). Figure 1 shows an overview of the procedure followed for the machine learning approach.

4.2. Semantic Orientation applied to EVOCA

For our semantic orientation experiments we have included the knowledge extracted from SentiWordNet version 3.0 [32] into the EVOCA corpus. SentiWordNet is a publicly available lexical resource for opinion mining which assigns three sentiment scores to each synset of WordNet[8]: positivity, negativity and objectivity. In fact, we have used the latest version, SWN 3.0, which includes a total of 82,115 nouns, 18,156 adjectives, 13,767 verbs and 3,621 adverbs with their respective scores. In order to achieve our goal of extracting the sentiment scores from SWN, we have used nouns, adjectives, verbs and adverbs as linguistic features.

[8] http://wordnet.princeton.edu. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.

In a first step, the English documents from EVOCA were processed applying a POS tagger: TreeTagger[9] [39]. The aim of this process was to obtain all the nouns, adjectives, verbs and adverbs of each review. Figure 2 shows a fragment of a tagged text from the EVOCA corpus.

The second step after tagging the EVOCA corpus was to generate a total of 15 sub-corpora from EVOCA by making a combination among the four possibilities: nouns, adjectives, verbs and adverbs. In this way, we generated the following sub-corpora in order to analyze the impact of each type of word:

- *only-noun, only-adj, only-verb, only-adv*
- *adj+noun, adj+verb, adj+adv, noun+verb, noun+adv, verb+adv*
- *adj+noun+verb, adj+noun+adv, noun+verb+adv, adj+verb+adv*
- *adj+noun+verb+adv*

For the purpose of generating the sub-corpora, we have taken into account the blanks found within each token. For example, if the POS tagger recognizes "*movie name"* as a noun, then we have added two nouns to the corresponding generated sub-corpus: *movie#n* and *name#n*.

```
the_DT first_JJ moments_NNS of_IN the_DT film_NN his_PRP$ music_NN
soundtracks_NNS  quiet_JJ  loss_NN  of_IN  function_NN  ,_,  to_TO
follow_VB  the_DT  first_JJ  quarter_NN  of_IN  the_DT  film_NN  -_:
Canadian_NNP American_NNP -_: rhythm_NN is_VBZ monotonous_JJ ,_,
unites_VBZ us_PRP with_IN the_DT main_JJ character_NN of_IN the_DT
film_NN  ,_,  and_CC  every_DT  now_RB  and_CC  then_RB  breaking_VBG
director_NN   boredom_NN   leaked_VBN   to_TO   the_DT   viewer_NN
through_IN  one_CD  of_IN  the_DT  surprises_NNS  the_DT  new_JJ  ,_,
then_RB we_PRP define_VBP the_DT lives_NNS of_IN Allen_NNP 's_POS
marriage_NN ,_, or_CC the_DT new_JJ neighbor_NN ._.
```

Figure 2. Example of a POS tagged text from the EVOCA corpus

Finally, we calculate the SWN score for each review or document of EVOCA in order to classify them as positive or negative. In this case, the SWN score of a document can be seen as the polarity score of such document. This score is obtained following the procedure proposed by Denecke [29] based on the calculation of a triplet of positivity, negativity and objectivity scores:

- For each token *A* with *n* synsets found in SWN, we calculate the average of its positivity score ($score_{pos}$) and the average of its negativity score ($score_{neg}$) by means of:

$$score_{pos}(A) = \frac{1}{n}\sum_{i=1}^{n} score_{pos}(i)$$

[9] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

$$score_{neg}(A) = \frac{1}{n}\sum_{i=1}^{n} score_{neg}(i)$$

- Then, we obtain the objectivity score ($score_{obj}$) for each token:

$$score_{obj}(A) = 1 - (score_{pos}(A) + score_{neg}(A))$$

- Finally, we determine the score-triplet for a document from summing up the score-triplet of each term and dividing each score by the number of considered terms in such document.

In order to classify a review as “positive” or “negative”, we have applied a classification rule according to which each review whose positivity score is larger than or equal to the negativity score is classified as “positive”; otherwise, it is considered “negative”. Figure 3 shows an overview of the procedure followed for the semantic orientation approach.

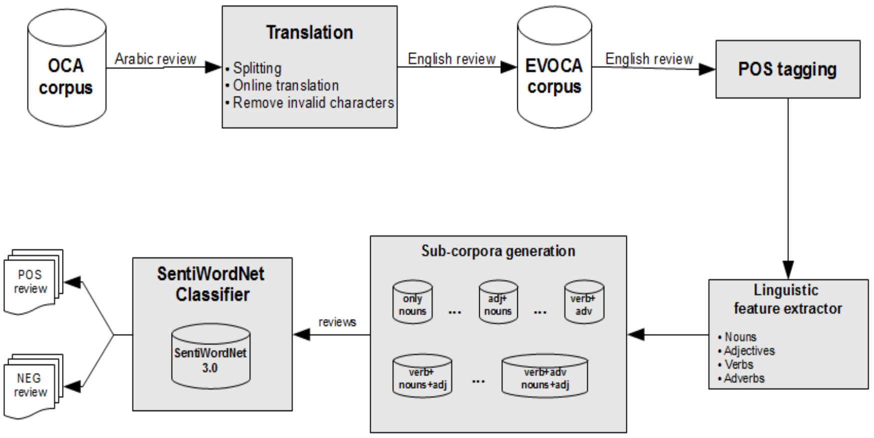


Figure 3. Overview of the semantic orientation approach

5. Results and discussion

In this section we show the results obtained with the EVOCA corpus by applying both approaches proposed in this work: machine learning and semantic orientation using SWN 3.0.

As we have already pointed out, 10-fold cross validation has been used for ML experiments in order to evaluate the polarity of the EVOCA corpus, making use of two learning algorithms, SVM and NB, to classify each review as positive or negative. For SVM, several kernels have been tested configuring the remaining default parameters. Moreover, we have studied the behavior of different heuristics such as filtering the stop words, the use of stemming, filtering those tokens with less than four characters and the use of unigrams, bigrams or trigrams. In addition, we have tested two configurations to generate the learning vectors: TF and TF·IDF. The combination of all these features produces a wide range of experiments (240 for SVM). However, the results for the different kernels are very similar in all of them and there are not many significant differences. For this reason, we have decided to show only the experiments of the Linear Kernel because it is the kernel which has obtained the best results. Table 2 shows the results for the 48 experiments with SVM using the linear kernel.

On the other hand, only 48 experiments were necessary for NB since it does not have particular parameters to configure. Table 3 shows the results obtained using the NB machine learning.

Weighting scheme	Stop words	Stemmer	Length > 3	n-grams	P	R	Acc	F1
TF·IDF	No	No	No	1	0.8798	0.9240	0.8980	0.9006
	No	No	Yes	1	0.8900	0.9200	0.9020	0.9039
	No	Yes	No	1	0.8878	0.9040	0.8940	0.8948
	No	Yes	Yes	1	0.8820	0.8840	0.8820	0.8823
	Yes	No	No	1	0.8801	0.9000	0.8880	0.8882
	Yes	No	Yes	1	0.8739	0.9040	0.8860	0.8875
	Yes	Yes	No	1	0.8831	0.8680	0.8740	0.8731
	Yes	Yes	Yes	1	0.8689	0.8840	0.8740	0.8753
	No	No	No	2	0.8810	0.9400	**0.9060**	**0.9087**
	No	No	Yes	2	0.8827	0.9080	0.8920	0.8940
	No	Yes	No	2	0.8904	0.9200	0.9020	0.9039
	No	Yes	Yes	2	0.8951	0.8960	0.8940	0.8944
	Yes	No	No	2	0.8794	0.8920	0.8840	0.8840
	Yes	No	Yes	2	0.8780	0.9040	0.8880	0.8893
	Yes	Yes	No	2	0.8846	0.8840	0.8820	0.8826
	Yes	Yes	Yes	2	0.8812	0.8920	0.8840	0.8850
	No	No	No	3	0.8562	0.9440	0.8900	0.8959
	No	No	Yes	3	0.8856	0.9200	0.8980	0.9001
	No	Yes	No	3	0.8821	0.9320	0.9000	0.9035
	No	Yes	Yes	3	0.8874	0.9080	0.8940	0.8957
	Yes	No	No	3	0.8822	0.9000	0.8880	0.8891
	Yes	No	Yes	3	0.8848	0.9160	0.8960	0.8983
	Yes	Yes	No	3	0.8856	0.8840	0.8820	0.8825
	Yes	Yes	Yes	3	0.8932	0.8960	0.8920	0.8928
TF	No	No	No	1	0.7001	0.9280	0.7620	0.7967
	No	No	Yes	1	0.8577	0.9000	0.8740	0.8769
	No	Yes	No	1	0.7068	0.9280	0.7680	0.8011
	No	Yes	Yes	1	0.8480	0.8560	0.8500	0.8504
	Yes	No	No	1	0.8692	0.8840	0.8740	0.8732
	Yes	No	Yes	1	0.8695	0.8800	0.8720	0.8718
	Yes	Yes	No	1	0.8813	0.8800	0.8780	0.8776
	Yes	Yes	Yes	1	0.8822	0.8760	0.8780	0.8764
	No	No	No	2	0.7050	0.9200	0.7660	0.7975
	No	No	Yes	2	0.8465	0.8880	0.8620	0.8658
	No	Yes	No	2	0.7035	0.9240	0.7660	0.7982
	No	Yes	Yes	2	0.8581	0.8600	0.8580	0.8569
	Yes	No	No	2	0.8777	0.8880	0.8800	0.8789
	Yes	No	Yes	2	0.8823	0.8960	0.8860	0.8855
	Yes	Yes	No	2	0.8829	0.8760	0.8780	0.8769
	Yes	Yes	Yes	2	0.8977	0.8800	**0.8880**	**0.8857**
	No	No	No	3	0.7180	0.9240	0.7740	0.8056
	No	No	Yes	3	0.8490	0.9000	0.8680	0.8711
	No	Yes	No	3	0.7245	0.9280	0.7800	0.8110
	No	Yes	Yes	3	0.8668	0.8560	0.8600	0.8595
	Yes	No	No	3	0.8869	0.8840	0.8840	0.8845
	Yes	No	Yes	3	0.8808	0.8840	0.8800	0.8813
	Yes	Yes	No	3	0.8878	0.8680	0.8780	0.8762
	Yes	Yes	Yes	3	0.8823	0.8680	0.8740	0.8728

Table 2. SVM with *linear* kernel experiments over the EVOCA corpus

Weighting scheme	Stop words	Stemmer	Length > 3	n-grams	P	R	Acc	F1
TF·IDF	No	No	No	1	0.7024	0.7640	0.7180	0.7300
	No	No	Yes	1	0.6844	0.7320	0.6940	0.7059
	No	Yes	No	1	0.7286	0.7640	0.7360	0.7435
	No	Yes	Yes	1	0.7165	0.7560	0.7240	0.7335
	Yes	No	No	1	0.6936	0.7400	0.7040	0.7146
	Yes	No	Yes	1	0.6900	0.7400	0.7000	0.7124
	Yes	Yes	No	1	0.7199	0.7480	0.7260	0.7317
	Yes	Yes	Yes	1	0.7121	0.7520	0.7200	0.7295
	No	No	No	2	0.8202	0.8080	**0.8140**	**0.8129**
	No	No	Yes	2	0.8191	0.6840	0.7640	0.7442
	No	Yes	No	2	0.7817	0.8280	0.7980	0.8035
	No	Yes	Yes	2	0.7920	0.6920	0.7540	0.7361
	Yes	No	No	2	0.8249	0.6720	0.7600	0.7382
	Yes	No	Yes	2	0.8222	0.6720	0.7629	0.7380
	Yes	Yes	No	2	0.8419	0.6960	0.7780	0.7572
	Yes	Yes	Yes	2	0.8170	0.7240	0.7780	0.7649
	No	No	No	3	0.9316	0.3640	0.6680	0.5167
	No	No	Yes	3	0.8815	0.2560	0.6120	0.3878
	No	Yes	No	3	0.9192	0.4080	0.6860	0.5614
	No	Yes	Yes	3	0.9031	0.2640	0.6180	0.4011
	Yes	No	No	3	0.8974	0.3120	0.6360	0.4484
	Yes	No	Yes	3	0.8902	0.3320	0.6440	0.4693
	Yes	Yes	No	3	0.8870	0.3200	0.6400	0.4627
	Yes	Yes	Yes	3	0.8762	0.3200	0.6380	0.4619
TF	No	No	No	1	0.6848	0.8920	0.7360	**0.7727**
	No	No	Yes	1	0.6851	0.7840	0.7060	0.7287
	No	Yes	No	1	0.6833	0.8760	0.7300	0.7659
	No	Yes	Yes	1	0.6939	0.7840	0.7160	0.7348
	Yes	No	No	1	0.7125	0.7640	0.7240	0.7352
	Yes	No	Yes	1	0.7009	0.7640	0.7140	0.7290
	Yes	Yes	No	1	0.7197	0.7880	0.7360	0.7503
	Yes	Yes	Yes	1	0.7064	0.7920	0.7280	0.7451
	No	No	No	2	0.6073	0.9760	0.6680	0.7477
	No	No	Yes	2	0.8106	0.7000	0.7660	0.7487
	No	Yes	No	2	0.6046	0.9720	0.6640	0.7445
	No	Yes	Yes	2	0.7847	0.7400	0.7680	0.7608
	Yes	No	No	2	0.8322	0.7000	0.7760	0.7585
	Yes	No	Yes	2	0.8319	0.7040	**0.7780**	0.7611
	Yes	Yes	No	2	0.8134	0.7160	0.7740	0.7594
	Yes	Yes	Yes	2	0.8141	0.7280	**0.7780**	0.7665
	No	No	No	3	0.5776	0.9920	0.6300	0.7295
	No	No	Yes	3	0.8833	0.4320	0.6840	0.5689
	No	Yes	No	3	0.5695	100	0.6200	0.7253
	No	Yes	Yes	3	0.8537	0.4280	0.6760	0.5647
	Yes	No	No	3	0.8716	0.4800	0.7020	0.6070
	Yes	No	Yes	3	0.8788	0.4760	0.7020	0.6086
	Yes	Yes	No	3	0.8618	0.4440	0.6860	0.5794
	Yes	Yes	Yes	3	0.8684	0.4440	0.6860	0.5823

Table 3. NB experiments over the EVOCA corpus

The rule-based SWN classifier was applied in the second set of experiments. In this case, each generated corpus from EVOCA was tested in order to evaluate the behavior of nouns, adjectives, verbs and adverbs taking into account the procedure explained in the previous section. The main goal in this set of experiments was to check if the use of a lexical resource like SWN could be a valid strategy when we do not have a training corpus for polarity classification using an English opinion corpus. Table 4 summarizes the results obtained using the semantic orientation approach. The second column shows the positive reviews that have been classified as positives. The third column indicates the positive reviews that have been classified as negatives. The fourth and fifth columns present the negative reviews that have been classified as positives and negatives, respectively.

EVOCA sub-corpus	**Rev POS**		**Rev NEG**		**P**	**R**	**Acc**	**F1**
	Pred POS	**Pred NEG**	**Pred POS**	**Pred NEG**				
only-noun	203	47	189	61	0.5179	0.8120	0.5280	0.6324
only-adj	198	52	152	98	0.5657	0.7920	0.5920	0.6600
only-verb	215	35	193	57	0.5270	0.8600	0.5440	0.6535
only-adv	44	206	32	218	0.5789	0.1760	0.5240	0.2699
adj+noun	212	38	171	79	0.5535	0.8480	0.5820	**0.6698**
adj+verb	209	41	172	78	0.5486	0.8360	0.5740	0.6624
adj+adv	122	128	47	203	0.7219	0.4880	0.6500	0.5823
noun+verb	222	28	199	51	0.5273	0.8880	0.5460	0.6617
noun+adv	96	154	49	201	0.6621	0.3840	0.5940	0.4861
verb+adv	72	178	44	206	0.6207	0.2880	0.5560	0.3934
adj+noun+verb	216	34	188	62	0.5347	0.8640	0.5560	0.6606
adj+noun+adv	149	101	66	184	0.6930	0.5960	**0.6660**	0.6409
noun+verb+adv	136	114	74	176	0.6476	0.5440	0.6240	0.5913
adj+verb+adv	137	113	65	185	0.6782	0.5480	0.6440	0.6062
adj+noun+verb+adv	165	85	87	163	0.6548	0.6600	0.6560	0.6574

Table 4. Experiments with semantic orientation approach applied to the EVOCA corpus

The results vary significantly among the different approaches. Whereas the machine learning approach provides high F1 score in general (0.9087 as best result), the semantic orientation approach achieves 0.6698 as the best F1 score. Nevertheless, it can be considered a good result if we take into account that the semantic orientation approach does not use a learning corpus (the difference among the best F1 results is -35%). Therefore, the use of the rule-based SWN classifier could be an interesting strategy for polarity opinion classification when there is not a training corpus.

A deeper analysis suggests some interesting conclusions. For the machine learning approach, it is clear that the use of SVM as learning algorithm achieves a better performance than NB. Comparing the best F1 scores, SVM is 11.78% better than NB. In addition, when the TF·IDF weighting scheme is applied to generate the learning vectors, it also achieves better results than the use of TF (+2.74% for SVM and +5.2% for NB, taking into account the best F1 scores). Finally, it seems desirable not to filter the stop words nor those tokens whose length is less than four characters, as well as not to apply a stemming process. However, when it comes to the use of n-grams, the best results are obtained by applying bigrams.

For the semantic orientation approach, we can conclude that nouns and adjectives are the lexical features that contain higher semantic importance, since the sub-corpus built with nouns and adjectives achieves the best F1 score (0.6698). It seems clear that verbs also provide important values in the semantic orientation approach, since the sub-corpus generated with adjectives and verbs achieves the second best F1 result (0.6624). In fact, the results obtained with the corpora consisting of only nouns, only adjectives and only verbs are very similar, achieving the best F1 result the *only-adj* corpus (0.66), then the *only-verb* corpus (0.6535) and finally the *only-noun* corpus (0.6324). It is important to pinpoint the poor performance of adverbs since their inclusion in the generated sub-corpora impoverishes the results. The main reason for this behavior is that most adverbs in the EVOCA corpus have assigned the score "1" for the *objective* value in SWN. In addition, the rest of adverbs have a negative polarity and so they are almost always considered as negative in the semantic orientation approach.

On the other hand, if we compare the results obtained with the ML approach based on the EVOCA corpus and those obtained with the same approach for the OCA corpus [8], another conclusion can be drawn. The best results for the OCA corpus were obtained through the TF·IDF weighting scheme, without either stemmer or stopper, and using bigrams. The accuracy for SVM and NB was 0.9060 and 0.8900, respectively. For the EVOCA corpus the best accuracy is also 0.9060 with SVM and 0.8140 with NB. There is no difference between the best accuracy results. However, taking into account the F1 results, the difference in favor of OCA varies between +0.38% and +0.28% for unigram and bigram models, respectively. As a consequence, we can conclude that the lost of precision in the translation process has been minimal, as can be seen in Table 5.

Corpus	n-gram model	Precision	Recall	Accuracy	F1
EVOCA	unigram	0.8900	0.9200	0.9020	0.9039
	bigram	0.8810	0.9400	0.9060	0.9087
OCA	unigram	0.8699	0.9480	0.9020	**0.9073**
	bigram	0.8738	0.9520	0.9060	**0.9112**

Table 5. Comparison of OCA and EVOCA best results for machine learning experiments using SVM, TF·IDF and without stemming

6. Conclusion and further work

In this paper we present the EVOCA corpus generated through the automatic translation into English of the OCA corpus. We have performed two different types of experiments with EVOCA in order to check the feasibility of this new resource. On the one hand, we have tested the EVOCA corpus using a machine learning approach. Two different algorithms have been proved, SVM and NB, showing that SVM yields better results than NB. On the other hand, we have integrated the lexical resource SWN into EVOCA in order to apply a semantic orientation approach. Although the results obtained are worse than those of the ML approach, the experiments prove that using lexical resources is a good alternative for polarity detection when we do not have an available labeled corpus.

The results obtained encourage us to continue working in this line. Thus, in future work we will design a combined approach using the semantic features extracted from SWN and applying a machine learning algorithm in order to improve the final results. We will also try to use OCA and EVOCA as parallel corpora in order to analyze in which cases they assign different labels to each review.

Acknowledgments

This work has been partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

It has been also partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER) through the TEXT-COOL 2.0 project (TIN2009-13391-C04-02) from the Spanish Government, and a grant from the Instituto de Estudios Giennenses through the Geocaching Urbano project (RFC/IEG2010). The fourth author is grateful to Agencia Española de Cooperación Internacional para el Desarrollo MAEC-AECID.

References

[1] B. Pang, L. Lee, Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2 (1-2), 2008, pp. 1-135.

[2] J. Wiebe, Learning subjective adjectives from corpora. Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence (AAAI), 2000, pp.735–740.

[3] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79–86. Association for Computational Linguistics.

[4] G. Somprasertsri, P. Lalitrojwong, Mining Feature-Opinion in Online Custumer Reviews for Opinion Summarization, Journal of Universal Computer Science, vol. 16, 2010, pp. 938-955.

[5] R. Mihalcea, C. Strapparava, Learning to Laugh (automatically): Computational Models for Humor Recognition, Journal of Computational Intelligence, Vol. 22, 2006, pp. 126-142.

[6] C. Strapparava, R. Mihalcea, Learning to identify emotions in text. In Proceedings of the 2008 ACM symposium on Applied computing (SAC '08). ACM, 2008, pp. 1556-1560.

[7] P. D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), 2002, pp. 417–424. ACL. Morristown, NJ, USA.

[8] XXXXX

[9] P.J. Stone, D. C. Dunphy, M. S. Smith, D. M. Ogilvie, The General Inquirer: A Computer Approach to Content Analysis, 1966. MIT Press.

[10] C. Strapparava, A. Valitutti, WordNet-Affect: an affective extension of WordNet. In: Proceedings of LREC 2004, 2004, pp. 1083–1086.

[11] A. Esuli, A., F. Sebastiani, SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of the 5th Conference on Language Resources and Evaluation (LREC), 2006, pp. 417-422.

[12] N. Li, D. D. Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast, Decision Support Systems, Volume 48, Issue 2, 2010, pp. 354-368.

[13] M. Hu, B. Liu, Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177. ACM. New York, USA.

[14] V. Hatzivassiloglou, J. Wiebe, Effects of adjective orientation and gradability on sentence subjectivity. Proceedings of the International Conference on Computational Linguistics (COLING), 2000, pp. 299–305.

[15] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis. Computational Linguistics. 2011.

[16] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp.347-354.

[17] J. Kamps, M. Marx, R. J. Mokken, M. D. Rijke, Using WordNet to measure semantic orientation of adjectives. The fourth international conference on Language Resources and Evaluation (LREC), 2004, pp. 1115–1118.

[18] R. Prabowo, M. Thelwall, Sentiment analysis: A combined approach. J. Informetrics, 3(2), 2009, pp. 143–157.

[19] A. Devitt, K. Ahmad, Sentiment polarity identification in financial news: A cohesion-based approach. In Proceedings of the 45th Annual Meeting of the ACL, 2007, pp. 984–991.

[20] F. R. Chaumartin, UPAR7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 2007, pp. 422-425. New York, NY: ACM.

[21] B. Ohana, B. Tierney, Sentiment Classification of Reviews Using SentiWordNet. IT&T Conference, 2009.

[22] H. Saggion, A. Funk, Interpreting SentiWordNet for Opinion Classification. Proceedings of the International Conference on Language Resources and Evaluation, (LREC 2010), 2010, pp.17-23, Valletta, Malta.

[23] S.M. Kim, E. Hovy, Identifying and analyzing judgment opinions. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL '06), 2006, pp.200-207.

[24] C. Zhang, D. Zeng, J. Li, F. Y. Wang, W. Zuo, Sentiment analysis of Chinese documents: From sentence to document level. Journal of the American Society for Information Science and Technology (JASIST), 60(12), 2009, pp. 2474–2487.

[25] H. Ghorbel, D. Jacot, Further Experiments in Sentiment Analysis of French Movie Reviews., in Elena Mugellini; Piotr S. Szczepaniak; Maria Chiara Pettenati & Maria Sokhn, ed., 'AWIC' , Springer, 2011, pp. 19-28.

[26] E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, Opinion Classification Techniques Applied to a Spanish Corpus. Proceedings of Natural Language Processing and Information Systems, 2011, pp. 169-176.

[27] C. Banea, R. Mihalcea, J. Wiebe, S. Hassan, Multilingual Subjectivity Analysis Using Machine Translation., in 'EMNLP' , ACL, 2008, pp. 127-135.

[28] K. Ahmad, D. Cheng, Y. Almas, Multi-lingual sentiment analysis of financial news streams. Proceedings of Science (GRID2006), 2006.

[29] K. Denecke, Using SentiWordNet for multilingual sentiment analysis, in 'ICDE Workshops' , IEEE Computer Society, 2008, pp. 507-512.

[30] A. Abbasi, H. Chen, A. Salem, Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems, 26*(3), 2008, 12:11-12.34.

[31] E. Boldrini, A. Balahur, P. Martínez-Barco, A. Montoyo, Emotiblog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. In R. Stahlbock, S.F. Crone & S. Lessmann (Eds.), DMIN, 2009, pp. 491-497. CSREA Press.

[32] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, in Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner & Daniel Tapias, ed., 'LREC' , European Language Resources Association, 2010.

[33] F. Sebastiani, Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), 2002, 1.

[34] C. Manning, H. Schutze, Foundations of Statistical Natural Language Processing. 1999, MIT Press. MA, USA.

[35] T. Joachims, Text categorization with support vector machines: Learning with many relevant features. Machine Learning: ECML-98, 1998, pp. 137–142. Springer.

[36] H. Zhang, The Optimality of Naive Bayes, in Valerie Barr, Zdravko Markov (eds.), Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004), 2004, AAAI Press.

[37] T. O'Keefe, I. Koprinska, Feature selection and weighting methods in sentiment analysis. Proceedings of the 14th Australasian Document Computing Symposium. Sydney, Australia, 2009.

[38] A. Esuli, F. Sebastiani, Determining the semantic orientation of terms through gloss classification. In O. Herzog, H. Schek, N. Fuhr, A. Chowdhury, W. Teiken (eds.), Proceedings of the 14th ACM international conference on Information and knowledge management, 2005, pp. 617–624.

[39] H. Schmid, Probabilistic Part-of-Speech Tagging Using Decision Trees, in Proceedings of the International Conference on New Methods in Language Processing, 1994.

BIBLIOGRAPHY

[ADY06] Khurshid Ahmad, Cheng David, and Almas Yousif. Multi-lingual sentiment analysis of financial news streams. In *First International Conference on Grids in Finance*, volume GRID2006, 2006.

[AMD12] Muhammad Abdul-Mageed and Mona Diab. Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet U?ur Do?an, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[BgNH$^+$08] Sasha Blair-goldensohn, Tyler Neylon, Kerry Hannan, George A. Reis, Ryan Mcdonald, and Jeff Reynar. Building a sentiment summarizer for local service reviews. In *In NLP in the Information Explosion Era*, 2008.

[BMdR06] Krisztian Balog, Gilad Mishne, and Maarten de Rijke. Why are they excited?: identifying and explaining spikes in blog mood levels. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters &#38; Demonstrations*, EACL '06, pages 207–210, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[Che06] Paula Chesley. Using verbs and adjectives to automatically classify blog sentiment. In *In Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches*, pages 27–29, 2006.

[Den08] Kerstin Denecke. Using sentiwordnet for multilingual sentiment analysis. In *ICDE Workshops*, pages 507–512. IEEE Computer Society, 2008.

[EAF12] Mohamed Elarnaoty, Samir AbdelRahman, and Aly Fahmy. A machine learning approach for opinion holder extraction in arabic language. *CoRR*, abs/1206.1011, 2012.

[ES06] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 417–422, 2006.

[Fil12] Elena Filatova. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[GJ11] Hatem Ghorbel and David Jacot. Sentiment analysis of french movie reviews. In *Advances in Distributed Agent-Based Retrieval Tools*, volume 361 of *Studies in Computational Intelligence*, pages 97–108. Springer, 2011.

[HL04] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[HM97] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, EACL '97, pages 174–181, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

[Hu12] Wei Hu. Real-time twitter sentiment toward midterm exams. *Sociology Mind*, Vol.2(2):177–184, 2012.

[Liu12] Bing Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[MC04] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In *In Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2004.

[MI12] A.M. Misbah and I.F. Imam. Mining opinions in arabic text using an improved "semantic orientation using pointwise mutual information" algorithm. In *Informatics and Systems (INFOS), 2012 8th International Conference on*, pages SE–61–SE–69, 2012.

[MVMCPOL13] Maria Teresa Martín-Valdivia, Eugenio Martínez-Cámara, José M. Perea-Ortega, and Luis Alfonso Ureña López. Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Syst. Appl.*, 40(10):3934–3942, 2013.

[MY11] Saif M. Mohammad and Tony (Wenda) Yang. Tracking sentiment in mail: how genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 70–79, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[OBRS10] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

[PL04] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *In Proceedings of the ACL*, pages 271–278, 2004.

[PL08] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, jan 2008.

[PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[PT09] Rudy Prabowo and Mike Thelwall. Sentiment analysis: A combined approach. *J. Informetrics*, 3(2):143–157, 2009.

[Tur02] Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[vALT10] Željko Agič, Nikola Ljubešić, and Marko Tadić. Towards sentiment analysis of financial texts in croatian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[WA07] Guangwei Wang and Kenji Araki. Oms-j: an opinion mining system for japanese weblog reviews using a combination of supervised and unsupervised approaches. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, NAACL-Demonstrations '07, pages 19–20, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[WBO99] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 246–253, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[WWH05] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[YH03] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 129–136, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[YLHA12] Xiaohui Yu, Yang Liu, Xiangji Huang, and Aijun An. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Trans. Knowl. Data Eng.*, 24(4):720–734, 2012.

[ZY07] Wei Zhang and Clement T. Yu. Uic at trec 2007 blog track. In *TREC*, 2007.

[ZZL$^+$09] Changli Zhang, Daniel Zeng, Jiexun Li, Fei-Yue Wang, and Wanli Zuo. Sentiment analysis of chinese documents: From sentence to document level. *J. Am. Soc. Inf. Sci. Technol.*, 60(12):2474–2487, dec 2009.