|  |
| --- |
| Opinion Mining |
| The Pre-SWOT Analysis Opinion Mining |
|  |
| Dr. Samir AbdelRahman |
| TA. Laila Moustafa |
| **14/2/2010** |

|  |
| --- |
| This document presents an overview of Opinion Mining (OM) technology and its techniques. It discusses the state of art for both Arabic and Latin languages. And then it present the SWOT analysis of applying OM in Arabic language.  |

**Table of Contents**

1. **Brief Overview ..................................................................................................3**
2. **State of the Art (For Latin Languages) ................................................................4**
	1. Technology and Reported Performance**........................................................**4
	2. Future Trends……………………………………………………………………………………………….6
	3. Opinion Mining Applications.............................................................................7
3. **State of the Art (For Arabic Language)**………………………………………………………………**.7**
4. **Dependency Between Technologies**…………………………………………………………………..**8**
5. **Language Resources**…………………………………………………………………………………………**…8**
	1. Available Resources (English, Arabic)……………………………………………………………8
	2. Needed Resources (English, Arabic)……………………………………………………………..8
6. **Strengths, weaknesses, opportunities and threats .………………………………………….8**
7. **Suggestions for Survey Questionnaire…………………….………………………………………..9**
8. **List of people/organizations pioneers ………………..…….………………………………………9**
9. **Key persons in each application area (on technical/LR levels)…….…………………….9**
10. **Suggestions for Language Resources .…….……………………………………..………………….9**

**Opinion Mining**

1. **Brief Overview**

*Opinion mining* (OM) is a recent sub discipline at the crossroads of information retrieval and computational linguistics which is concerned not with the topic a document is about, but with the opinion it expresses. Opinion-driven content management has several important applications, such as

* Determining critics’ opinions about a given product by classifying online product reviews, or
* Tracking the shifting attitudes of the general public toward a political candidate by mining online forums.

This field is very important since “What other people think” has always been an important piece of information for most of us during the *decision-making process*. Long before Web, many of us asked our friends to recommend an auto mechanic or to explain who they were planning to vote for in local elections, or consulted *Consumer* *Reports* to decide what dishwasher to buy. But the Internet has now made it possible to find out about the opinions and experiences of people that are neither our personal acquaintances nor well-known professional critics.

Indeed, according to two surveys of more than 2000 American adults each,

* 81% of Internet users (or 60% of Americans) have done online research on a product at least once;
* among readers of online reviews of restaurants, hotels, and various services (e.g., travel agencies or doctors), between 73% and 87% report that reviews had a significant influence on their purchase;
* 32% have provided a rating on a product, service, or person via an online ratings system, and 30% (including 18% of online senior citizens) have posted an online comment or review regarding a product or service.

However, consumption of goods and services is not the only motivation behind people’s seeking out or expressing opinions online. A need for political information is another important factor.

The user hunger for online advice and recommendations that the data above reveals is one reason behind the interest in new systems that deal directly with ***opinions extraction***. But while a majority of American internet users report positive experiences during online product research, at the same time, 58% also report that online information was missing, impossible to find, confusing, and/or overwhelming. Thus, there is a clear need to aid consumers of products and of information by building better information-access systems than are currently in existence.

However, aside from individuals, an additional audience for systems capable of automatically analyzing consumer sentiments (opinions) in online resources is companies. They need to understand how their products and services are perceived. Companies can respond to the consumer insights they generate through social media monitoring and analysis by modifying their marketing messages, brand positioning, product development, and other activities accordingly.

Research on opinion mining started with

* Identifying *opinion* (or *sentiment*) *bearing words*, e.g., great, amazing, wonderful, bad, and poor. Many researchers have worked on mining such words and identifying their *semantic orientations* (i.e., positive or negative).
* Classification of text (entire documents or sentence) by their contents as expressing a positive or a negative sentiment about an object (e.g., a *movie*, a *camera*, or a *car*).
	+ Classification is useful but it does not find what the reviewer liked and disliked about object (i.e. a negative opinion on an object does not mean the reviewer dislike everything of object). So the solution is to go to the feature level of each object. For example, picture quality or battery life of camera object.
* Extraction of opinion expression from text, eventually including relations with the rest of content, e.g. recognizes an opinion, who is expressing it, who/what is the target of the opinion.
1. **State of the Art (For Latin Languages)**
	1. **Technology and Reported Performance:**
		1. **Sentiment Words/Phrases Identification**

Sentiment words or phrases are those that are primarily used to express the writer’s sentiment.

* As current work on sentiment analysis focus on content words (nouns, verbs, adjectives and adverbs), most of the work use *part-of-speech* (POS) tagging to extract them (Hu and Liu, 2004; Turney, 2002).
* Other natural language processing technique like *stop words removal*, *stemming and fuzzy matching* are also used in the preprocessing stage to extract sentiment words/phrases.
	+ 1. **Sentiment Word/Phrase Orientation Identification**

The main approaches to identify the semantic orientation of sentiment word/phrases are

* Statistical-based or
* Lexicon-based.

Current works to identify sentiment words and their sentiment orientation or polarity are mainly focused on *adjectives* and *adverbs*, as they are often considered as the most obvious indication of *subjectivity*

Hu and Liu (2004) apply POS tagging and some natural language processing techniques to the text to extract the adjectives as sentiment words. Then they use WordNet to determine whether the extracted adjective has a positive or negative polarity. They used the semantic orientation of synonyms and antonyms to predict the orientation of the adjectives. They start with a seed list which consists of 30 manually selected common adjectives. Then they use WordNet to predict the orientation of all the adjectives in the extracted sentiment word list by finding out whether its synonyms or antonyms are in the seed list or not. Once the adjective’s orientation is predicted, it is added to the seed list and can be used to determine other adjectives’ orientation.

Hatzivassiloglou and McKeown (1997) used a method to automatically retrieve semantic orientation information using indirect information collected from a *large corpus* as they pointed out those dictionaries such as WordNet do not include semantic orientation information and there lacks direct association between antonyms and synonyms especially when they are domain dependent. They first extract all conjunctions of adjectives from the corpus with relevant morphological relations. Then they use a log-linear regression model and combine the information from the different conjunctions to determine if each two conjoined adjectives are of same or different orientation. The adjectives are represented in a graph with hypothesized same or different orientation links and are then separated into two subsets of different orientation using a clustering algorithm. Lastly they compare the average frequencies in each adjective group and label the higher frequency group as positive.

Turney (2002) used mutual information between two words to classify the adjective or adverb’s orientation in reviews of different domains. Prior to word sentiment classification, they use POS tagging to extract adjectives and adverbs. They assume that terms with similar orientation tend to co-occur in documents. Based on the pointwise mutual information (PMI) approach which is a measure of the strength of semantic association between two words is used. The semantic orientation of a word/phrase x is then calculated as PMI(x, ”excellent”)−PMI(x, ”poor”)–the word/phrase is classified as positive if it is more strongly associated with ”excellent”, and negative otherwise. They chose the words" excellent” and" poor” because these two words are commonly used to express the two ends of sentiments in reviews.

However, the accuracy of previous systems ranging from 78% to 87%

* + 1. **Sentiment Sentence/Document Classification**

Sentiment word/phrase orientation identification is used for sentence/document classification as in (Hu and Liu, 2004), whereas other works (Pang et al., 2002) classifies sentiment sentences/documents without the knowledge of each sentiment words.

Hu and Liu (2004) predict the orientation of the opinion sentence in their study of customer reviews. If the number of positive/negative opinion words exceeds the other one, then the sentence is classified as positive/negative. In case of a draw, the average orientation of closest opinion word for the product feature or the orientation to the previous opinion sentence is used for classification. Their sentence orientation accuracy is 84.2%.

Pang et al. (2002) used supervised machine learning to classify movie reviews. Without classifying individual sentiment words or phrases, they extract different features from the review and use machine learning algorithms Na¨ıve Bayes (NB), Maximum Entropy (ME) and Support Vector Machine (SVM) to classify the reviews. The features include a single item or a combination of the following:

* presence of unigrams,
* frequency of unigrams, bigrams, POS tag,
* adjectives, top 2633 unigrams and
* Position of the word in text.

They achieved accuracies between 78.7% and 82.9%.

* + 1. **Opinion Holder Identification**

This phase is about determining who the owner of this opinion is, Techniques used:

* Classification (Maximum Entropy)
* Conditional random fields
* Hidden Markov Model
* Rule-based approach
	1. **Future Trends:**
* **Use of other types of words**, Most of the work done for sentiment analysis so far has been focused on content words. However, other word types could also have affect sentiment classification. For example, conjunctions as" but” connects two parts of a sentence together but emphasize on the part following" but”. For example, ”The movie is good but difficult to understand” would be classified as neutral if we simply count the number of positive sentiment words (”good”) and negative ones (”difficult”). It may even be classified as positive if we look at the opinion word (”good”) closest to the feature (”the movie”). However, if we make use the conjunction" but” and give a higher weight to the sentence part following" but”, in this case" difficult”, the sentence would be classified correctly as negative.
* **Sentiment lexicon construction**.

* **Dealing with negation expressions**, unlike the operations in mathematics where the negation of positive is negative and vice versa, the adding negations to a word or phrase in real world text does not equal to the effect of putting a minus sign in front of a number. For example," late” is negative; but adding a" not” in front does not make ”not late” to be positive as ”not late” is not equal to ”early” which is the opposite of ”late”.
* **Complexity of sentence/document,** current approaches only attempt to classify sentences with simple structures. Without analyzing the whole sentence structure, the overall sentiment may be classified wrongly. For example, in some movie review; when the writer uses a lot of paragraphs to describe how he/she hates one of the movie actors but uses only a small paragraph to express how he/she still loves the movie after all. Current approaches may very well be fooled to classify the review as negative.
* **Contextual Sentiment,** current works for sentiment orientation identificationof words have not considered much of the contextenvironment.Same words in different contexts can have the different sentiment orientation. For example, the word" poor” in" the system performance is poor” has a negative sentiment orientation; but in" we should help the poor people"," poor” is neutral.
	1. **Opinion Mining Applications:**

Opinion mining research filed has a lot of applications, therefore there are a good number of companies, large and small, that have opinion mining and sentiment analysis as part of their mission. This section lists some of these applications:

* **Applications to review related websites.**
* Review oriented search engine.
* Summarization user reviews.
* **Applications as a sub component technology,**
* *Recommendation systems*, these systems should not recommenditems that receive a lot of negative feedback.
* In online systems that display advertisement as sidebars, it is helpful to detectWeb pages that contain sensitive content inappropriate for advertisement placement. For example, remove the advertisement when relevant negative statements are discovered.
* *Question answering* is another area where sentiment analysis can prove useful.
* **Applications in business and government intelligence.** Consider, for instance, the following scenario. A major computer manufacturer, disappointed with unexpectedly low sales, finds itself confronted with the question: “Why aren’t consumers buying our laptop?” While concrete data such as the laptop’s weight or the price of a competitor’s model are obviously relevant, answering this question requires focusing more on people’s personal views of such objective characteristics. Moreover, subjective judgments regarding intangible qualities — e.g., “the design is tacky” or “customer service was condescending” — or even misperceptions —e.g., “updated device drivers are not available” when such device drivers do in fact exist — must be taken into account as well.
1. **State of the Art (For Arabic Language)**

To the best of our knowledge, there is only some research to build basic linguistic resources for OM. This research is made in the center of excellence, Faculty of Computers and Information, Cairo University.

Linguistic resources include

* A ***sentiment Arabic lexical semantics database*** where each word has its subjectivity, objectivity and polarity values. This database is similar to SentiWordNet.
* ***Clues database*** which include subjective and objective terms with their orientation and strength.

**However, Arabic opinion mining still in the early stage (i.e. they are some attempts to build basic linguistic resources needed in OM). Therefore, research is open for different topics in OM.**

1. **Dependency Between Technologies**
* **Opinion Mining Technology can serve as sub component technology,**
* **Recommendation systems,** these systems should not recommenditems that receive a lot of negative feedback.
* **Advertisement online Systems,**online systems that display advertisement as sidebars, it is helpful to detectWeb pages that contain sensitive content inappropriate for advertisement placement. For example, remove the advertisement when relevant negative statements are discovered.
* **Question answering**, is another area where sentiment analysis can prove useful.
1. **Language Resources**
	1. **Available Resources (English, Arabic)**

To the best of our knowledge, there is no resource for Arabic OM. But for English there are some

* SentiWordNet. This resource is for public use. It is like WordNet but having emphasis on sentiment orientation of the words. They associate each synset s in WordNet to three numerical scores Obj(s), Pos(s) and Neg(s) to describe how objective, positive, and negative the terms contained in the synset are.
* Annotated corpus for opinions. It contains 535 news articles (11,114 sentences).
	1. **Needed Resources (English, Arabic)**
* Sentiment lexicon. This lexicon is similar to the ordinary one but having emphasis on sentiment orientation of the words.
* Clues database. This database identifies opinion terms and their orientation and strength. The term can be multi-word expression like (not entirely satisfactory)
* Basic NLP resources like POS tagger, morphological analyzer, shallow or deep parser.
* For supervised techniques, we need large annotated corpus.
1. **Strengths, weaknesses, opportunities and threats**

Apply of the Opinion Mining on Arabic Language.

|  |  |
| --- | --- |
| **Strength** |  **Weakness** |
| * The importance of this field for the *decision-making process*.
* Availability of people’s opinions and experiences on the Internet and the Web.
* The Need of aiding consumers of products and of information by building better information-access systems.
* The Market need to understand how their products and services are perceived.
 | * Few researches to build basic linguistic resources for OM.
* Lack of Arabic Opinion Mining Linguistic resources.
* A few limited researches have focused on Opinion Mining for Arabic text due to the complexity of Arabic language and rareness of the linguistic Resources.
* There are no available standardized Opinion Mining Arabic Linguistic resources.
 |
| **Opportunities** | **Threats** |
| * Increasing International and National market need of Arabic Opinion Mining.
* Available local/regional/international funds.
 | * International and regional competition.
* Lack of focused funds.
 |

1. **Suggestions for Survey Questionnaire**
* How can the Arabic Opinion Mining benefit the International and National Market?
* What are the challenges of applying Opinion Mining on Arabic Language?
* Is there any shortage on the Arabic Linguistic Resources that is needed by the Opinion Mining techniques?
1. **List of people/organizations pioneers in each application area to be targeted by the Survey**
* There are no organizations which are considered as pioneer on this area.
1. **Key persons in each application area (on technical/LR levels)**
* Jan Wiebe.
* Bing Liu.
* Rada Mihalcea.
1. **Suggestions for Language Resources (specific to the application area) if ALTEC would like to start collection immediately.**
* Basic NLP resources like POS tagger, morphological analyzer, shallow or deep parser.