Sentiment Analysis Using a Novel Human Computation Game

Claudiu-Cristian Musat THISONE **Alireza Ghasemi** **Boi Faltings**

Artificial Intelligence Laboratory (LIA)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
IN-Ecublens, 1015 Lausanne, Switzerland
firstname.lastname@epfl.ch

Abstract

In this paper, we propose a novel human computation game for sentiment analysis. Our game aims at annotating sentiments of a collection of text documents and simultaneously constructing a highly discriminative lexicon of positive and negative phrases.

Human computation games have been widely used in recent years to acquire human knowledge and use it to solve problems which are infeasible to solve by machine intelligence. We package the problems of lexicon construction and sentiment detection as a single human computation game. We compare the results obtained by the game with that of other well-known sentiment detection approaches. Obtained results are promising and show improvements over traditional approaches.

1 Introduction

We propose a novel solution for the analysis of sentiment expressed in text media. Novel corpus based and lexicon based sentiment analysis methods are created each year. The continual emergence of conceptually similar methods for this known problem shows that a satisfactory solution has still not been found. We believe that the lack of suitable labelled data that could be used in machine learning techniques to train sentiment classifiers is one of the major reasons the field of sentiment analysis is not advancing more rapidly.

Recognizing that knowledge for understanding sentiment is common sense and does not require experts, we plan to take a new approach where labelled data is obtained from people using human computation platforms and games. We also prove that the method can provide not only labelled texts, but people also help by selecting sentiment-expressing features that can generalize well.

Human computation is a newly emerging paradigm. It tries to solve large-scale problems by utilizing human knowledge and has proven useful in solving various problems (Von Ahn and Dabbish, 2004; Von Ahn, 2006; Von Ahn et al., 2006a).

To obtain high quality solution from human computation, people should be motivated to make their best effort. One way to incentivize people for submitting high-quality results is to package the problem at hand as a game and request people to play it. This process is called gamification. The game design should be such that the solution to the main problems can be formed by appropriately aggregating results of played games.

In this work, we propose a cooperative human computation game for sentiment analysis called Guesstiment. It aims at annotating sentiment of a collection of text documents, and simultaneously constructing a lexicon of highly polarized (positive and negative) words which can further be used for sentiment detection tasks. By playing a collaborative game, people rate hotel reviews as positive and negative and select words and phrases within the reviews that best express the chosen polarity.

We compare these annotations with those obtained during a former crowd-sourcing survey and prove that packaging the problem as a game can improve the quality of the responses. We also compare our approach with the state-of-the-art machine

Proceedings of the 3rd Workshop on the People's Web Meets NLP, ACL 2012, pages 1–9,
Jeju, Republic of Korea, 8-14 July 2012.

learning techniques and prove the superiority of human cognition for this task. In a third experiment we use the same annotations in a multi faceted opinion classification problem and find that results are superior to those obtained using known linguistic resources.

In (section 2) we review the literature related to our work. We then outline the game and its rules (section 3). We compare the Guesstiment results to the state-of-the-art machine learning, standard crowd-sourcing methods and sentiment dictionaries(section 4) and conclude the paper with ideas for future work (section 5).

2 Related Work

In this section we review the important literature related and similar to our work. Sine we propose a human computation approach for sentiment analysis, we start by reviewing the literature on human computation and the closely related field of crowd-sourcing. Then we move on by having a brief look on the human computation and knowledge acquisition games proposed so far by the researchers. Finally, we briefly review major sentiment analysis methods utilized by the researchers.

2.1 Human Computation and Crowd-Sourcing

The literature on human computation is highly overlapping with that of crowd-sourcing, as they are closely connected. The two terms are sometimes used interchangeably although they are slightly different. Crowd-sourcing in its broadest form, "is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call"(Quinn and Bederson, 2011; Howe, 2006). Since the first use of the word crowd-sourcing by J. Howe (Howe, 2006), there has been a lot of interest in this field due to the wide accessibility of anonymous crowd workers across the web.

The work described in (Rumshisky, 2011) uses crowd-sourcing to perform word sense disambiguation on a corpus. In (Vondrick et al., 2010), crowd-sourcing is used for video annotation. Moreover, (Christophe et al., 2010) has used crowd-sourcing for satellite image analysis.

(Settles, 2011a) is another approach which aims at combining active learning with crowd-sourcing for text classification. The principal contribution of their work is that as well as document annotation, they use human computation also to perform feature selection. (Law et al.,) is another recent work which proposes a game for acquisition of attribute-value pairs from images.

2.2 Human Computation Games

Luis Von Ahn, the pioneer of the field of human computation, designed a game to encourage players to semantically annotate a large corpus of images (Von Ahn and Dabbish, 2004). It was the first human computation game.

Following Von Ahn's work, more researchers were encouraged to package computational problems as joyful games and have a group of non-expert users play them (Von Ahn, 2006). Verbosity (Von Ahn et al., 2006a) was designed with the goal of gathering common sense knowledge about words. KissKissBan (Ho et al., 2009) was another game for image annotation.

Peekaboom(Von Ahn et al., 2006b) aimed at image segmentation and "Phrase Detectives" (Chamberlain et al., 2008) was used to help constructing an anamorphic corpus for NLP tasks. Another human computation game is described in (Riek et al., 2011) whose purpose is semantic annotation of video data.

2.3 Sentiment Analysis

The field of sentiment analysis and classification currently mostly deals with extracting sentiment from text data. Various methods (Turney, 2002; Esuli and Sebastiani, 2006; Taboada et al., 2011; Pang et al., 2002) have been proposed for effective and efficient sentiment extraction of large collections of text documents.

Sentiment classification methods are usually divided into two main categories: Lexicon based techniques and methods based on machine learning. In lexicon-based methods, a rich lexicon of polarized words is used to find key sentences and phrases in text documents which can be used to describe sentiment of the whole text (Taboada et al., 2011). Machine learning methods, on the other hand, treat the sentiment detection problem as a text classification task (Pang et al., 2002).

Most of the research has been oriented towards finding the overall polarity of whole documents. The problem was broken down even more by using of the faceted opinion concept (Liu, 2010). The goal of this attempt was to determine precisely what aspects of the concepts the expressed opinions should be linked to. We will use this distinction to assess our method's viability in both overall and multi faceted opinion analysis.

The work in (Brew et al., 2010) is an attempt to use crowd-sourcing for sentiment analysis. The authors use a crowd of volunteers for the analysis of sentiments of economical news items. Users provide annotations which are then used to learn a classifier to discriminate positive articles from negatives. It uses active learning to select a diverse set of articles for annotations so that a generalizable, precise classifier can be learned from annotated data. The work in (Zhou et al., 2010) is another approach to use active learning to improve sentiment classification. It uses a deep network to learn a sentiment classifier in a semi-supervised manner. Moreover, this method uses active learning to learn from unlabeled data which are the most informative samples that need to be labeled.

Two more recent works that have focused on sentiment classification by designing human computation games are (Weichselbraun et al., 2011) and (Al-Subaihin et al., 2011). In (Weichselbraun et al., 2011) the game ”Sentiment Quiz” has been proposed that aims at finding the degree of polarity of words in a lexicon. In each round of the game, the player is asked to vote about polarity of a given words from most negative to most positive. The player is score based on the agreement between his vote and the votes of previous players. ”Sentiment Quiz” demands annotation in the word level and therefore can only be used to construct a sentiment lexicon.

Another work which aims at sentiment classification is (Al-Subaihin et al., 2011). In this work, a multi-player game is proposed which aims at finding the sentiment of individual sentences. The game is played by three groups of two players each. Each team is shown a sentence and its members are asked to highlight the sentiment carrying terms of the sentence separately and quickly. The first team whose players' votes match wins and the current game round finishes. The game continues by introducing different sentences to the teams and hence gathers information about polarity of terms and their corresponding context.

3 The Proposed Game

In this section we propose a novel human computation game called Guesstiment. We use the information provided while playing this game to obtain a reliable dataset of sentiment annotated data as well as a lexicon of highly polarized positive and negative words.

Having two by-products as the result of playing instead of merely trying to obtain document annotations is the most important contribution of Guesstiment. The idea of using crowd-sourcing for feature extraction has already been used in (Settles, 2011b), but not as a human computation game. In the rest of the following section, we will discuss the game play and rules of Guesstiment.

3.1 Rules of Guesstiment

Guesstiment is a two-player asynchronous game. It aims at annotating a large corpus of text documents, similar to the goal of the ESP game in (Von Ahn and Dabbish, 2004) for images. However, Guesstiment does this in a different way because of its rules and asynchronous approach. The differences allow Guesstiment to obtain more useful information from played game rounds than ESP does, since each player contributes in providing a different type of information.

The two players of the game are called ”Suggester” and ”Guesser”. These roles are initialized randomly and interchanged between the two players after each round of the game.

The Suggester, who starts each round will be given the whole text of a review document and he/she is supposed to:

1. Decide whether the whole text is positive or negative, i.e. the author is praising about a subject or criticising it.

2. Select a single word (or a sequence of words, as short as possible) which best describes the polarity (positive or negative) he has selected in part (1). For example, when the negative polarity is chosen, the word ”terrible” would be

a good choice for the representative word (provided that it is present in the text).

The Guesser, on the other hand, will be given only the word (or word sequence) suggested by the Suggester (he won't see the whole text) and he has to guess polarity of the whole text just based on that single word. If the polarities suggested by the two players agree, they are both given some positive score (based on factors described below) otherwise 0. Then the roles are interchanged and the game continues with a new document. The guesser can also refuse to make a guess about polarity of the text (when for example the suggested word is ambiguous or not very discriminative) in which case the suggester has two more opportunities to suggest another word from the text.

Guesstiment is a cooperative game. It means that the two players are not opponent and they both receive equal score after each round (Not high score for one player and low score for the other). Therefore, the Suggester should make his best efforts to select the most polarized word from the test which best describes the selected sentiment or polarity. The UI screens for Suggester and Guesser are depicted in figures 1a and 1b respectively.

3.1.1 Scoring

The score of each suggested word (or word sequence) depends on a variety of factors, including the length of the sequence and its novelty, i.e. how many times it has already been selected by other players. Suppose that the word sequence w is present in the current text document and also it has been present in text documents of n_w of previously played game rounds. Assuming w has been selected k_w time before current game round, the potential score of w, PS_w is defined as:

$$PS_w = \left[\frac{1}{length(w) \times \frac{k_w}{n_w}} \right] \quad (1)$$

In (1), $length(w)$ is the length (number of words) of phrase w. Using this scoring strategy, players are encouraged to select as shortest phrases as possible. Single words that are not already selected by other players will yield the highest score.

Moreover, some words are not allowed as suggestions and will yield zero score regardless of the agreement in polarity judgments. They are selected by putting a threshold on the potential score of words and placing those with a score lower than the threshold on the forbidden list. These words are colored red in the text and are separately displayed in the forbidden list.

The cooperation between the Suggester and the Guesser requires an agreement between them. This allows the game to collect precise annotations and simultaneously build a good quality lexicon of words which are most important in detecting polarity of the text.

The total score of each player is displayed on the scoreboard at the bottom of the Suggest/Guess graphical user interface. The potential score of a word is also displayed while typing which allows users to avoid selecting words with low score.

4 Experiments

4.1 Implementation Details

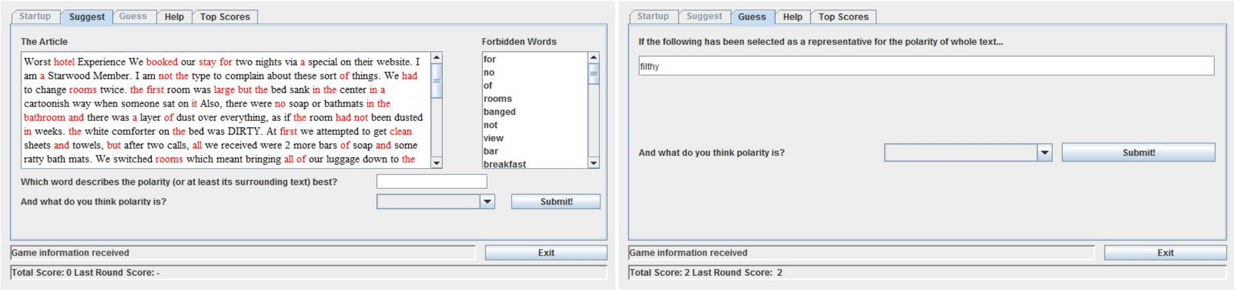
The game was implemented as a traditional three-tier web application. For data storage, we used the H2 embedded database which has proven to be fast enough for scientific information retrieval applications. For the server side of the application we used the Play! framework, which is a lightweight easy to use framework for Java MVC web application framework.

The client side is a Java Applet. We used a service oriented approach to define interactions between the client and the server so that the game play is defined as a sequence of HTTP requests between client and server. Using this approach,, client and server are maximally separate and therefore various client applications can be written, for instance to run on smart phones.

4.2 Experimentation Environment

A total of 80000 review texts were extracted along with their corresponding ratings from the TripAdvisor website [1]. Among these, 1000 articles were randomly selected and inserted in the game database to be used in game rounds. More than 20 players played the game over the course of one month and 697 annotations were collected during this period,

[1] http://www.trip-advisor.com



The Article

Worst hotel Experience We booked our stay for two nights via a special on their website. I am a Starwood Member. I am not the type to complain about these sort of things. We had to change rooms twice. the first room was large but the bed sank in the center in a cartoonish way when someone sat on it Also, there were no soap or bathmats in the bathroom and there was a layer of dust over everything, as if the room had not been dusted in weeks. the white comforter on the bed was DIRTY. At first we attempted to get clean sheets and towels, but after two calls, all we received were 2 more bars of soap and some ratty bath mats. We switched rooms which meant bringing all of our luggage down to the

Forbidden Words

for
no
of
rooms
banged
not
view
bar
breakfast

Which word describes the polarity (or at least its surrounding text) best?

And what do you think polarity is?

Game information received

Total Score: 0 Last Round Score: -

(a) Suggester Role

If the following has been selected as a representative for the polarity of whole text...

filthy

And what do you think polarity is?

Game information received

Total Score: 2 Last Round Score: 2

(b) Guesser Role

Figure 1: The Suggest/Guess UI.

of which 312 are distinct. The players were volunteer students who played the game after an initial advertisement in Facebook. Almost all of them were non-native English speakers, both graduate and undergraduate.

For selecting articles for each round, a combination of strategies were used. From the set of documents which have not already been labelled by any of the players, we select the article with the least difference between number of positive and negative (as collected in the lexicon constructed so far) words so that we get the most information from annotations. If all document have been annotated at least two times, we make the selection among documents for which the two annotations disagree, so that we solve disagreements by majority vote.

4.3 Quality of Annotations

For each review text in the TripAdvisor website, there is a corresponding rating score given by the very writer of the text. These ratings, score the quality of services of the mentioned hotel from the point of view of the review writer.They give a score of 1 (most negative) to 5 (most positive) to the described hotel which is presumably correlated with the inherent sentiment of the review text written by the same author for the same hotel.

We used these review ratings as the ground truth to assess the quality of player annotations. We considered review ratings higher than 3 as having positive sentiment and those with rating lower than 3 as having negative sentiment. Review with rating equal to 3 were considered neutral and excluded from further experiments. Let $Rate_i$ be the numerical rating of document i, according to the above criteria, we accept document i if and only if:

$$Rate_i \in \{1, 2, 4, 5\} \tag{2}$$

As well as annotations provided by players of the game, we also compared the ground truth to the results of state-of-the-art machine learning techniques adapted for sentiment analysis. We considered sentiment analysis as a typical binary classification problem and used a simple bag of words approach for feature extraction.

For the learning algorithm, we used Support Vector Machines (SVM) and the Naïve Bayes methods which are two well-known learning algorithms for text classification problems (Brew et al., 2010). In the SVM method, document feature vectors are transformed to high-dimensional kernel space and then a maximum margin separating hyperplane is sought in the new kernel space. Training of SVM is quadratic in the amount of training data and therefore it is hardly feasible for large-scale problems.

The Naïve Bayes approach is another learning algorithm which is simpler and faster to run than SVM. In this statistical method, feature words are considered independent random variables and Bayes rule is used to derive posterior probabilities of having positive and negatives sentiment for each document. The sentiment with maximum posterior probability is the predicted sentiment for the given document.

Results of comparison between the ground truth and various annotation approaches are depicted in table 1. For the game results, we aggregated different annotations for individual documents by majority voting. Moreover, for the machine learning algorithms we used cross-validation to adjust the parameters. Moreover, the results were computed by averaging 10-fold cross-validation results over all folds.

Accuracy of each method is defined as:

$$Accuracy = \frac{N_{correct}}{N_{total}} \tag{3}$$

.

In equation (3), $N_{correct}$ is the number of document with computed sentiment equal to the ground truth and N_{total} is the total number of documents. It can be seen in table 1 that our method outperforms machine learning.

Table 1: Comparison of Game Annotation Accuracies With that of Automatic Classifiers

Method	**Accuracy**
Game Collected Annotations	**90.4**
Naïve Bayes	80.5
Logistic Regression	83.6
SVM	82.8

Table 2: Comparison between Quality of the Results of Gamification and Crowd-Sourcing

Method	**Accuracy**
Game Collected Annotations	**100**
Aggregated Crowd Votes	82.5

4.4 Comparison with Classical Crowd-Sourcing

We also made a comparison between our approach and simple crowd-sourcing. For this goal, we used the results of a survey conducted in summer 2011. 40 of the review texts were selected randomly from the whole dataset and given to a crowd of 27 student to be annotated based on their sentiment. Individual annotations for each document were aggregated using majority voting. The ground truth was computed in the same way as the previous section.

We re-executed the Guesstiment in a period of one week using only those 40 reviews and compared the quality of the obtained annotations to that of the survey (aggregated using majority voting). Similar to the survey, we aggregated annotations for individual documents by majority voting.

The results, depicted in table 2, are quite promising. Accuracy of the simple crowd-sourcing was 82.5% whereas gamification acquired an accuracy of 100%. We can see that merely packaging the problem as a game significantly improves accuracy of the results.

We can infer from tables 1 and 2 that gamification actually helps in obtaining good quality annotation results. Therefore, annotations derived from players' effort are highly reliable and can be used for further studies, discussed below.

4.5 Comparison with Sentiment Dictionary Performance

The previous experiments proved the viability of human computation for detecting the polarities of whole documents. Manual classification is however expensive, even if it takes the form of a game. We take a step further and use the result of the game, in the form of a sentiment dictionary, in a subsequent automated classification task. We compare the Guesstiment dictionary with an established resource, OpinionFinder (Wilson et al., 2005) in a multi faceted opinion classification problem.

The OpinionFinder dictionary (OF) contains 8220 entries representing tuples of English words, either in the original form or stemmed, and their most likely parts of speech. Each tuple has an associated polarity which can be positive, negative or neutral. There are 5276 words in the original form in the dictionary that have a positive or negative polarity. By contrast, the Guesstiment dictionary GS only contains 312 terms, nearly 17 times less than OpinionFinder. Of these, 175 words are negative and 137 positive. Each of the words within the two dictionaries has an intrinsic polarity $P(w), \forall w \in D = \{OF, GS\}$.

The opinion extraction task is topic oriented. We extract faceted opinions (Liu, 2010) - occurrences of sentiment that can be attached to a given topic or class within a topic model $z_i \in \theta, i \in \{1..k\}$ where k is the number of independent topics. We used two sets of topics: the first is a statistical topic model obtained with Latent Dirichlet Allocation (Blei et al., 2003) with $k = 50$ topics from which we retained the most probable 5 words for each topic and created sets of topic relevant terms $\mathcal{P}\{z_i\}$. The second set of topic terms contains the most common 90 nouns in all available hotel reviews, which were afterwards manually separated into 11 classes.

Many Guesstiment dictionary words, such as ”value” and ”gentleman” bear meaning by themselves (i.e. are nouns) and are not useful in this analysis. However the great majority of the words are adjectives or adverbs. This makes them useful for faceted sentiment analysis. We only consider combinations of topic words and opinion dictionary terms and the allowed combinations are based on grammatical dependency chains:

$$w_1 \xrightarrow{*} w_2, w_1 \in \mathcal{P}\{z_i\}, i = \{1..k\}, w_2 \in D \tag{4}$$

obtained using the Stanford parser (De Marneffe and Manning, 2008).

This binding brings confidence to the model and prevents the accidental misinterpretation of unigrams. Also, the higher granularity of the opinion description allows clustering users based on their preferences.

We define a construct c relevant to a topic z_i within a review r as

$$\begin{gathered} c_{z_i} \in z_i \times D \\ c = (w_1, w_2 | w_1 \in \mathcal{P}\{z_i\}, w_2 \in D, w_1 \xrightarrow{*} w_2) \end{gathered} \tag{5}$$

The polarity of the said relevant construct is given by the orientation of the contained dictionary word:

$$P(c) = P(w_2) \tag{6}$$

The polarity P of the opinion expressed within a review $r \in R$ with respect to a topic z_i is defined as the sum of the polarities of constructs relevant to z_i. This allows us to assess the strength of the opinion expressed with regard to a topic.

$$\begin{gathered} P : R \times \theta \mapsto \mathcal{R} \\ P(r, z_i) = \sum_r P(c_{z_i}), i = \{1..k\} \end{gathered} \tag{7}$$

while the overall polarity of the review is the sum of all topic dependent polarities

$$\begin{gathered} P : R \mapsto \mathcal{R} \\ P(r) = \sum_{i=1}^{k} P(r, z_i) \end{gathered} \tag{8}$$

We test whether the method assigns positive overall polarities to reviews which have high (4 and 5) numeric ratings $nr(r)$ and negative to those with low ones (1 and 2). We compare the precision and recall of the method using both dictionaries and both topic sets. The dataset consists of 2881 reviews regarding the Las Vegas Bellagio hotel. Table 3 summarizes the results. We confine our analysis to a subset of 2594 reviews from the initial 2881 for which the numeric rating is greater or smaller than 3.

We notice that the recall is consistently lower for the frequent noun topics, which was expected because of the significantly smaller number of topic terms. However the recall does not depend on the chosen dictionary. This is relevant because with a much smaller pool of dictionary terms, similar results are obtained. Precision is constant in all four cases, which also shows that results similar to those of OpinionFinder can be obtained with our much smaller dictionary.

The precision and recall values in Table 3 do not reflect the capacity of the higher grained opinion analysis to extract targeted user preferences. The overall variance of the hotel's numeric ratings $Var(nr(r))$ shows how much the reviewers disagree on the quality of the stay. Generally this disagreement comes from the different sets of values the reviewers have. For example some consider cleanliness the most important aspect while others are interested in a busy nightlife.

We cluster users based on the faceted opinions we retrieved, using the k-Means algorithm (MacQueen, 1967). Each reviewer is represented by a feature vector and each feature i within the vector is the cumulative opinion expressed by the reviewer with regard to topic z_i. The reviews within the same cluster j have a similar representation from the mined opinion perspective. If the quality of the opinion mining process is high, the numeric ratings associated to the reviews within a cluster will also be similar, thus their variance $Var_j(nr(r))$ will be lower than the overall variance. We study the difference between the mean intra cluster variance $avgVar_j(nr(r))$ and overall variance $Var(nr(r))$

and the results are shown in table 4 for different numbers of clusters, using both topic models and both dictionaries.

The results show that we succeeded in decreasing the variance by more than 20% using the Guesstiment dictionary and the frequent noun topics. A 17% decrease is obtained by using the same topic set and the OpinionFinder dictionary, while the decreases for the LDA topics with dictionaries are three times lower. This proves that the dictionary resulted from playing the Guesstiment game is better suited for faceted opinion analysis than an established resource like OpinionFinder.

5 Conclusion

In this paper we introduced Guesstiment, a human computation game for simultaneous feature extraction and sentiment annotation. By conducting various experiments, we showed that quality of the annotations obtained using our approach outperforms those obtained by classic crowd-sourcing. This is an indicator of the fact that packaging a crowd-sourcing problem as a game can improve the quality of the obtained results. It's mostly because that games attract more attention from people than simple questions which are common ways of crowd-sourcing.

We also showed that our approach outperforms state-of-the-art machine learning methods which illustrates that human computation power is still superior to machine intelligence in this problem.

The idea of the game could be further extended by testing other more complicated scoring functions which could better motivate players to submit high quality results. Also other document selection strategies can be created to make a better trade-off between informativeness and interestingness, or exploration and exploitation. Moreover, a computer player could be designed to perform active learning on feature extraction and direct the word suggestion process toward selecting more informative features, hereby obtaining a more discriminative high-quality lexicon.

References

A.A. Al-Subaihin, H.S. Al-Khalifa, and A.M.S. Al-Salman. 2011. A proposed sentiment analysis tool for modern arabic using human-based computing. In *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services*, pages 543–546. ACM.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

Anthony Brew, Derek Greene, and Padraig Cunningham. 2010. Using crowdsourcing and active learning to track sentiment in online media. In *ECAI*, volume 215 of *Frontiers in Artificial Intelligence and Applications*, pages 145–150. IOS Press.

J. Chamberlain, M. Poesio, and U. Kruschwitz. 2008. Phrase detectives: A web-based collaborative annotation game. *Proceedings of I-Semantics, Graz.*

Emmanuel Christophe, Jordi Inglada, and Jerome Maudlin. 2010. Crowd-sourcing satellite image analysis. In *IGARSS*, pages 1430–1433. IEEE.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. *Coling 2008 Proceedings of the workshop on CrossFramework and CrossDomain Parser Evaluation CrossParser 08*, 1(ii):1–8.

Andrea Esuli and Fabrizio Sebastiani, 2006. *SentiWordNet: A publicly available lexical resource for opinion mining*, volume 6, page 417422. Citeseer.

C.J. Ho, T.H. Chang, J.C. Lee, J.Y. Hsu, and K.T. Chen. 2009. Kisskissban: a competitive human computation game for image annotation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 11–14. ACM.

J. Howe. 2006. The rise of crowdsourcing. *Wired magazine*, 14(14):1–5.

E. Law, B. Settles, A. Snook, H. Surana, L. von Ahn, and T. Mitchell. Human computation for attribute and attribute value acquisition.

Bing Liu. 2010. Sentiment analysis : A multi-faceted problem. *Science*, 25(1):76–80.

J. B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

A.J. Quinn and B.B. Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1403–1412. ACM.

Table 3: Precision and Recall of Overall Review Polarity Detection

		Total	**TP**	**FP**	**FN**	**Precision**	**Recall**
LDA	**GS**	2594	1399	456	739	0.75	0.65
	OF	2594	1434	461	699	0.75	0.67
Top Frequency	**GS**	2594	1275	405	914	0.75	0.58
	OF	2594	1362	371	861	0.78	0.61

Table 4: Weighted Average of Intra Cluster Variances

k			**1**	**10**	**11**	**12**	**13**	**14**	**15**	**% Decrease**
Variance	**LDA**	GS	1.03	1.01	0.99	0.99	0.98	0.96	1.01	6.79
		OF	1.03	0.99	0.95	0.99	0.99	1.01	0.96	7.761
	Top Frequency	GS	0.62	0.52	0.52	0.51	0.49	0.52	0.5	20.964
		OF	0.68	0.57	0.59	0.56	0.57	0.56	0.56	17.64

Laurel D. Riek, Maria F. O'Connor, and Peter Robinson. 2011. Guess what? a game for affective annotation of video using crowd sourcing. In Sidney K. D'Mello, Arthur C. Graesser, Björn Schuller, and Jean-Claude Martin, editors, *ACII (1)*, volume 6974 of *Lecture Notes in Computer Science*, pages 277–285. Springer.

Anna Rumshisky. 2011. Crowdsourcing word sense definition. In *Linguistic Annotation Workshop*, pages 74–81. The Association for Computer Linguistics.

B. Settles. 2011a. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Conference on Empirical Methods in Natural Language Processing (EMNLP), Edinburgh, Scotland, July*.

B. Settles. 2011b. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. *Conference on Empirical Methods in Natural Language Processing (EMNLP), Edinburgh, Scotland, July*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Peter D. Turney. 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, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

L. Von Ahn and L. Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM.

L. Von Ahn, M. Kedia, and M. Blum. 2006a. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78. ACM.

L. Von Ahn, R. Liu, and M. Blum. 2006b. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 55–64. ACM.

L. Von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Carl Vondrick, Deva Ramanan, and Donald Patterson. 2010. Efficiently scaling up video annotation with crowdsourced marketplaces. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *ECCV (4)*, volume 6314 of *Lecture Notes in Computer Science*, pages 610–623. Springer.

Albert Weichselbraun, Stefan Gindl, and Arno Scharl. 2011. Using games with a purpose and bootstrapping to create domain-specific sentiment lexicons. In Craig Macdonald, Iadh Ounis, and Ian Ruthven, editors, *CIKM*, pages 1053–1060. ACM.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder : A system for subjectivity analysis. *Learning*, (October):34–35.

Shusen Zhou, Qingcai Chen, and Xiaolong Wang. 2010. Active deep networks for semi-supervised sentiment classification. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING (Posters)*, pages 1515–1523. Chinese Information Processing Society of China.