Utilizing Support Vector Machines in Mining Online Customer Reviews

Taysir Hassan A. Soliman[1], Mostafa A. Elmasry[2]
Information Systems Dept.
Faculty of Computers & Information, Assiut University[1], Fayoum University[2]
Assiut[1], Fayoum[2], Egypt
taysirhs@yahoo.com[1], mostafa_elmasry2006@yahoo.com[2]

Abdel Rahman Hedar[3], M. M. Doss[4]
Computer Sciences Dept.[3], Electrical Engineering Dept.[4]
Faculty of Computers & Information[3], Faculty of Engineering[4], Assiut University,
Assiut, Egypt
hedar@aun.edu.com[3], magdy@aun.edu.eg[4]

***Abstract*—as e-commerce is increasingly becoming popular, the number of customer reviews that a product receives grows rapidly. However, for popular products, many online product reviews exist but for other reviews product reviews are very few. These online discussions about particular products may help other online users to make a decision in buying/ not buying those products, like in amazon.com[1] and ebay.com[2]. Since an enormous number of unstructured and ungrammatical reviews on a product exist, opinion mining is getting a crucial research area for better decision making of buying products. In this paper, we apply an opinion mining approach to summarize the unstructured and ungrammatical users' reviews, based on Support Vector Machine (SVM). Two levels of classification is applied: 1) Features classification and 2) Polarity classification for every feature class. Our approach has been tested on Amazon data with dataset of 535 sentences, where a summary is obtained and analysis of precision (93.15%) and recall (92.41%) illustrate the accuracy of the proposed system.**

Keywords: - Opinion mining, E-commerce, sentiment analysis, support vector machines, reviews classification, opinion visual summary.

I. INTRODUCTION

In recent years, the use of Internet commerce has grown drastically, where the internet became an enormous market for the products. Companies are using electronic commerce to enter new markets that would have otherwise been excluded from due to geographical locations, cost, or other reasons. Companies look to electronic commerce to extend their products to new sets of customers and new parts of the globe. The Web enables a company to introduce a new product into the market, get immediate customer reaction to it, refine and perfect it, all without incurring enormous investment in a physical distribution infrastructure or buying a shelf space at a retailer of distributor [1]. As customer buys products without seeing them physically, he only sees an image or a video of the required product.

The only way for a customer to get persuaded with the product is to see the opinions of other customers who have bought the same product and used it before him. For example, customers can give their feedback or opinion (review) about a specific Nokia phone as follows: "This is a great nokia phone; it cheap; the touch screen is responsive and smooth." However, the new customer must read many reviews to take a decision of the buying operation. As a result, the process of getting feedback (Opinions) from the customers supports the e-commerce operations. Opinion Mining is the process of extracting the judgment or evaluating of a user's review of feedback about a specific topic by using text analysis. However, there is a large number of opinion reviews existing on the web. In many cases, opinions are hidden in blogs and forums, where it is hard for a human reader to extract sentences that are useful for him. In addition, most of the sentences are written in an unstructured and ungrammatical format.

In this work, we propose an opining mining methodology to help new customers to make a decision of buying/not buying a product by summarizing the reviews. The proposed approach is based on Support Vector Machine (SVM) for reviews' classification to conclude the summary of the customers' opinions, visualizing the results as well. The paper is organized as follows: Section two illustrates required definitions for opinion mining problem. Section three illustrates previous work. Section four shows the proposed system architecture, a visual summary of the results, and a comparison of our results with other approaches; section five concludes our work and introduces future work in this track.

II. PROBLEM DEFINITION

The main problem of opinion mining is: Given a set of evaluative text documents D that contain opinions (or sentiments) about an object. Opinion mining aims to extract attributes and components of the object that have been commented on in each document d Є D and to determine whether the comments are positive, negative or neutral [2]. In general, opinions can be expressed on anything, e.g., a product, a service, a topic, an individual, an organization, or an event. The general term object is used to denote the entity that has been commented on. An object has a set of components (or parts) and a set of attributes. Each component may also have its sub-components and its set of attributes. Thus, the object can be hierarchically decomposed based on the part-of relationship [2]. In this section, there are three main definitions that will be clarified: an object, model of feature-based opinion mining, and

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

opinion summary, which will be used later in the following sections.

A. An object

An object O is an entity which can be a product, topic, person, event, or organization. It is associated with a pair, O: (T, A), where T is a hierarchy or taxonomy of components (or parts) and sub-components of O, and A is a set of attributes of O. Each component has its own set of sub-components and attributes. The word "features" is used to represent both components and attributes. Using features for objects (especially products) is quite common in practice [2].

B. Model of Feature-Based Opinion Mining

An object O is represented with a finite set of features, F = $\{f_1, f_2, \ldots, f_n\}$, which includes the object itself. Each feature f_i Є F can be expressed with a finite set of words or phrases W_i, which are synonyms [3]. That is, there is a set of corresponding synonym sets W = $\{W_1, W_2, \ldots, W_n\}$ for the n features [2].

C. Opinion Summary

There are many ways to analyze and summarize the mining results. One simple way is to produce a feature-based summary of opinions on the object [3].

III. Related work

Many mining online reviews exist but it this area of research still needs more work to effectively mine the evolving online product reviews data. Zhang and Tran [4] propose a mining method to rank and classify the online reviews as Helpful and Not Helpful, or to classify the reviews into qualified and bad claims [5]. However, the results of the two methods are not related with knowledge about the product itself. Das and Sivaji [6] use SVM to make subjective classification for customer reviews but their method produce low precision and recall. Zhang et al. [7] divide the reviews into sentences, where the sentences look as a news corpus in [6]. Gamon [8] apply clustering techniques to find the salient patterns in sentences but results produce low accuracy in free text reviews. Abulaish [9] utilizes a markup Language as a filter to get the sentences from the reviews. Hiremath et al. [10] differentiate the reviews into three formats: format 1 is Pros and Cons, format 2 is Pros, Cons and detailed review and finally format 3 is the free text format. The authors apply k-means algorithm to cluster the reviews and assess the weight for a given reviews by considering cluster weight but the k-means clustering algorithm will produce many outliers because of unstructured reviews.

There are two types of sentences in the review: subjective sentences and comparative sentences, [7,11,12] summarize the orientation of the sentence based on set of positive (POS) and set of Negative (NEG) words. Wang and Ren [11] identify the set of features to be extracted from the reviews, calculating the Point wise Mutual Information(PMI) of the opinion words, but the free text reviews may contain more than one word that describe the feature. Wang and Ren [11] apply Vector of Feature Intensities (VFI) for Binary Polarity (BP) to assign a numeric degree of polarity to every sentence and Probability of Polarity (PP) to assign probability of polarity to every sentence. Although there is a lot of opinion mining research, there is still need more efficient algorithms because of the wide varieties of products available, the increasing number of reviews on the web, unstructured and ungrammatical customers' reviews and web users' needs of visual summarization of product evaluation.

Our work is based on classifying the reviews' sentences of a product according to features by calculating the cosine similarity to classify the reviews. Then we suggest a set of stop words, good words and a set of bad words. Inside every feature class, the sentences are classified to two classes according to sentence polarity, as clarified in the following section.

IV. Proposed Opinion Mining Approach and Methodology

In this section the proposed opinion mining approach based on SVM, will be explained in details in the next section. It is introduced mainly to summarize the customer reviews. It consists of four main phases, as illustrated in Fig. 1: data collection, data preprocessing, data processing and reviews visual summarization. These phases are classified as follows:

Phase 1: Data Collection (Customer Reviews): there are many websites that get feedback from users like amazon.com and ebay.com. The proposed system uses a collection of reviews for different models of Nokia phone's products which applied in [14,15].

Phase 2: Data preprocessed: Removing unneeded words is a basic operation when we mine the unstructured data because the data will be converted to numbers as input for statistical equations. Therefore, collected reviews are preprocessed by removing the stop words. A list of stop words in [7, 14], are used to remove unneeded words to facilitate the data processing. A list of 317 words is used for removing unneeded words (a, the, on, in, etc) [8,11,14]. Then, a stemmer is used to stem the yield words of each review to get the words' root. We stem the words to return its root to easily extract features and identify polarity. Every product has many reviews; these reviews are going to be split into sentences, as the user describes each feature in a sentence.

Phase 3: Data processing: this process consists of two main sub phases as following:

Level 1 Classification: This consists of three parts as follows:

- **Subjective sentences:** As a result of the second phase, each review is split into sentences by using comma, full stop and exclamation mark as sentences splitter [8]. Then, sentences, which talk explicitly about at least one feature on the product, are obtained.
- **Feature Extraction:** this process explores the subjective features at every sentence [10,14]. The product features are listed before [9,10], where the famous features for the mobile phone are (*phone, battery, screen, Camera, Price, Wi-Fi*). We suppose a list of feature names for each product (famous features) [9,14]. Then, we extract the opinion word

that describes the feature [6,7,9,14]. A list of 664 good words is used to extract the positive description (Opinion word) of the feature at each sentence and a list of 658 bad words to extract the negative description (Opinion Word) [8,11]. Table I shows some opinion words used in the system.

TABLE I. EXAMPLES OF POSITIVE AND NEGATIVE WORDS

Positive Words	Negative Words
Good, Great, Nice, Well, Beauty, Amazing, Easy, Charm, Best, Fine, Fun, Kind, Love, Magic, Safe, pretty … etc	Bad, Weak, Difficult, Complex, Shy, Worst, angry, annoy, shame, bore, fear, hate, hard, hurt, poor, sad … etc

After sentences are grouped from the reviews of a product, the system classifies the sentences according to the described feature and another classification into the feature class according to the polarity. There is an opinion word that describes the feature of the product like: "*the camera is good*". In this example, the feature is a noun (*camera*), and so the system will classify the sentence according to the feature mentioned in it as a first classification. The system will extract the opinion word which is (*good*) in the example before so, the system will classify each sentence into a positive class or a negative class and this is the second classification. The results will be two classes for each feature: the first one contains the positive sentences and the second contains the negative sentences. The precision and recall will be calculated to evaluate our proposed approach against previous results, which will be presented in section 4.

- **Features Classification**: the system classifies the sentences into classes which are the number of features. Classification is based on the feature that is mentioned in the sentence. However, if two features are described in the sentence (*the battery and the camera of this mobile are nice*), the tool will classify the sentence in the two classes (*Battery* and *Camera* in the example). Each feature of a product may be mentioned in more than one subject, the user can describe the feature in many ways, he can say: "*the camera is good*", and he can say: "*the resolution is good*", which has the same feature. So, we suggest synonyms for every feature as used in [7,8] to classify the sentences into feature classes. Table II shows a partial view of features synonyms.

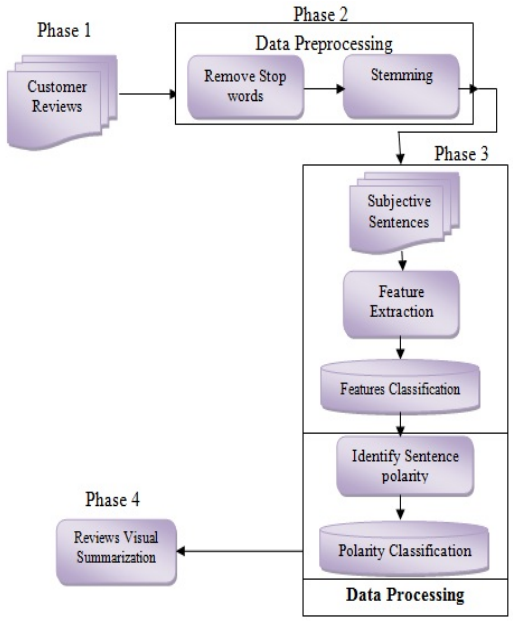


Figure 1. Proposed System Architecture

TABLE II. PARTIAL VIEW OF FEATURES' SYNONYMS

Feature	Synonyms
Camera	Photo, Picture, Resolution, Video, resolution, zoom, pixel, cam, camera, pic... etc
Product itself	Phone, Device, <Name>, Product, nokia, version, model, mobile, <Model No.> … etc

As classification requires labeled classes, we suppose the synonyms of every feature as a class vector V_F=<S_1, S_2 .... S_n>, and the system build a vector for every sentence V_S = <W_1, W_2 ... W_n>. The system uses the Cosine similarity between the synonyms vectors V_{Fi} and all sentences' vectors V_S,

$$\text{Sim}(V_F, V_S) = \frac{\sum_{i=0}^{n}(S_i \cdot W_i)}{\sqrt{\sum(S_i)^2 \cdot \sum(W_i)^2}} \qquad (1)$$

Where n is the vector length. SVM is applied to classify the sentences and define the margin between classes. The similarity will be calculated, as shown in Fig. 2, between the features' synonyms vectors (V_{F1}, V_{F2} … V_{FN}), where N is the number of features and the vector of every sentence of the product V_{Si}.

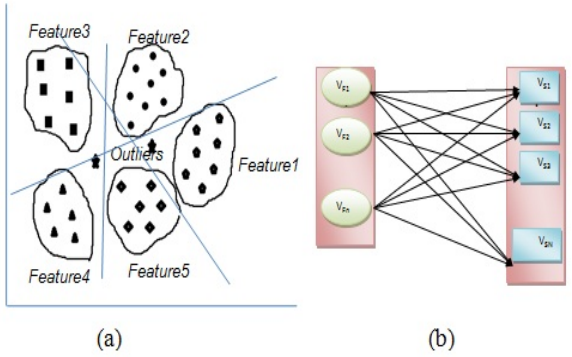


Figure 2. Linear SVM Classification: a. Classes after classification. b. Similarity calculations between features vectors and sentences

The sentence will be classified according to the maximum value of the similarity results between this sentence and all features' classes' vectors. Table III shows the number of sentences classified to every feature's class, where column 2 shows the number of sentences for each product; column 4 shows the number of sentences classified to every feature's

class. Column 5 shows the number of sentences that classified correctly.

TABLE III. NUMBER OF POSITIVE AND NEGATIVE SENTENCES FOR PRODUCTS' FEATURES (POLARITY CLASSIFICATION)

Device	Number of all sentences	Feature's Class	Number of sentences In each Class	Number of classified sentences
Nokia	535	Phone	321	497
		Battery	35	
		Screen	25	
		Camera	65	
		Price	25	
		Wi-Fi	26	

Level 2 Classification: consists of two parts as follows:

- **Identify Sentences Polarity**: inside the class of each feature, the mining system classifies the sentences into two classes (Positive and Negative) according to the polarity of the sentence.
- The classification is based on the opinion word in the sentence, but if two opinion words are in the sentence with a conjunction (and), it seems like word synonyms. Table IV illustrates the results of the second classification process. Because the customers do not follow the language grammatical rules, the results must be less accurate and not all sentences are classified. Recall and Precision, as in equations two and three, respectively, are calculated to show the performance evaluation of the results, as shown in Table V as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

Where, TP = $\sum_{i=0}^{NF}(F^{i}_{Pos} + F^{i}_{Neg})$, NF= number of features, F^{i}_{Pos} =number of sentences in the positive class of feature (i) and F^{i}_{Neg}=number of sentences the negative class of feature (i), FP=N_s – TP, where, N_s=number of all sentences in features classes for a product, FN=N-N_s, where N=number of sentences of the product's reviews, TN=N-TP.

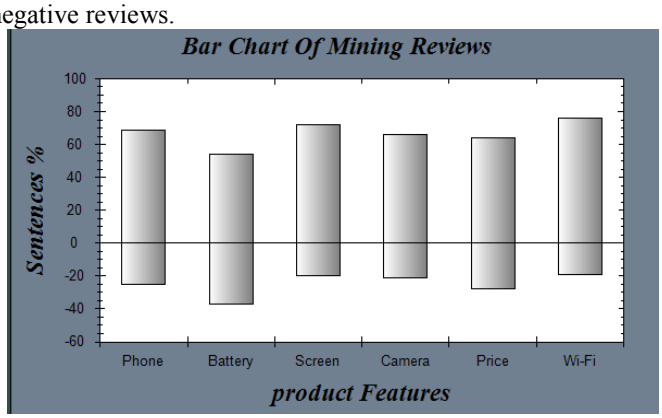
TABLE IV. NUMBER OF POSITIVE AND NEGATIVE SENTENCES FOR PRODUCTS' FEATURES (POLARITY CLASSIFICATION)

Feature	Features Polarity	
	Pos.	Neg.
Phone	222	81
Battery	19	13
Screen	18	5
Camera	43	14
Price	16	7
Wi-Fi	20	5

TABLE V. PERFORMANCE EVALUATION FOR THE RESULTS

TP	FP	FN	TN	Precision%	Recall%
463	34	38	72	93.15	92.41

Phase 4: Reviews Visual Summarization: In order to generate a summary of the customer reviews Nokia product, the system produces a visual summary based on the features of the product and the polarity (Positive and Negative) of the sentences. Fig. 3 shows the visual summary for mining results of Nokia products. It shows the visual summary of the customers' opinions on Nokia product's features, it obvious that price is the battery and price have the most number of negative reviews.



Figur 3. Polarity visual summary for Nokia products' features

V. RESULTS AND DISCUSSION

To evaluate the performance of our results, we have applied precision and recall. From manual evaluation for some of customer reviews, we explore some writings' errors in customer reviews: syntax errors, word misuse and missing sentence end. The results are summarized as in Fig. 3. Because the customers' reviews are unstructured and ungrammatical (free text), the average precision is 93.15and the average recall is 92.41. The evaluation of our proposed system is compared to the evaluation of M. Hu and B. Liu [14] and Ding and Liu [15] systems are shown in Table VI. The precision and recall of our proposed system are higher than in Ding and Liu [15] and Hu and B. Liu [14] because of the accurate classification within the 2 levels of classification. The new methodology is built on 2 levels of classification; the first one is feature's classification and the second one is polarity classification. The web users not follow the grammatical rules while writing the reviews which affect the classification results.

TABLE VI. EVALUATION COMPARISON FOR SAME DATASET

System	Precision%	Recall%
M. Hu and B. Liu [14]	80	72
Ding and Liu, WSDM-2008 [15] (Features Based Summary)FBS	92	74
Proposed system	93.15	92.41

VI. CONCLUSIONS AND FUTURE WORK

In the current work, an opinion mining approach was proposed to mine unstructured and ungrammatical customers' reviews. It was based on splitting the product's reviews into a collection of sentences. Two classification techniques were accomplished for the sentences: feature classification and polarity classification. The experiments indicate classification results and a visual summary. In a future study, we will concentrate on the new methodologies to improve the results of customer reviews' summarization. In addition, Arabic data will be used for online mining since Arabic requires a lot of preprocessing.

REFERENCES

[1] Liu, B. Opinion Mining. WWWC 2008

[2] Free Research Papers on E-commerce, http://anyfreepapers.com/free-research-papers/e-commerce-research-paper.html, (Accessed April 8, 2012).

[3] Liu,B. Web Data Mining, Springer-Verlag Berlin Heidelberg 2007

[4] R. Zhang and T. Tran, "Helping E-Commerce Consumers Make Good Purchase Decisions: A User Reviews-Based Approach," MCETECH 2009. Springer-Verlag Berlin Heidelberg 2009.

[5] Shilpa A., Mahesh J. and C. P. Rose, "Identifying Types of Claims in Online Customer Reviews," Proceedings of NAACL HLT 2009. Association for Computational Linguistics.

[6] Das and Sivaji. B, "Opinion-Polarity Identification in Bengali". ICCPOL 2010.

[7] K. Zhang, Ramanathan N. and A. Choudhary "Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking," 3rd Workshop on Online Social Networks Boston, MA, June 2010.

[8] M. Gamon, A. Aue, Simon C. and E. Ringger "Pulse: Mining Customer Opinions from Free Text," IDA 2005.

[9] M. Abulaish, M. N. Doja and T. Ahmad, "Feature and Opinion Mining for Customer Review Summarization," Springer-Verlag Berlin Heidelberg 2009.

[10] P.S Hiremath, Siddu P. Algur and S. Shivashankar "Quality Assessment of Customer Reviews Extracted from Web Pages: A Review Clustering Approach," (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010, 600-606.

[11] J. Wang and H. Ren, "Feature-based Customer Review Mining," System p. 1-9 (2002).

[12] Martin P and S. Becker, "Opinion Summarization of Web Comments," Proceedings of the 32nd European Conference on Information Retrieval, ECIR 2010.

[13] Jorge, C. and Laura, P. A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating. In: ECIR 2011, LNCS 6611, pp. 55–66, 2011

[14] M. Hu and B. Lui. "Mining and Summarizing Customer Reviews," KDD'04, August 22–25, 2004, Seattle, Washington, USA.

[15] X, Ding. Liu, B**.** and Philip S. Yu. A Holistic Lexicon-based Approach to Opinion Mining. In: Proceedings of the International Conference on Web Search and Web Data Mining (WSDM 2008).