

Feature-Opinion Extraction using Formal and Informal Classifier Technique

Jennifer Selvaraj
Mumbai University,
Pillai Institute of Information Technology,
New Panvel

J.W. Bakal, Ph.D
Mumbai University,
Shivajirao S. Jondhale College of Engineering
Dombivli, Thane

ABSTRACT

Merchants selling products on the Web often ask their customers to share their opinions and hands-on experiences on products they have purchased. This is not only true for organizations but also true for individuals. Our beliefs and perceptions of reality, and the choices we make, are, to a considerable degree, conditioned upon how others see and evaluate the world. For this reason, when we need to make a decision we often seek out the opinions of others. Unfortunately, reading through all customer reviews is difficult, especially for popular items, the number of reviews can be up to hundreds or even thousands. This makes it difficult for a potential customer to read them and make an informed decision. Thus a compressed and precise opinion list is what a user would generally desire. The contents available on the Web are not in the standard format. The idea is to classify these contents as formal and informal. The type of algorithm used is linguistic. It incorporates grammatical and other knowledge of the language which helps in understanding the text, thus trying to improve the mining approach to mine product features and their opinions from Web opinion sources for formal as well as for informal text.

General Terms

Opinion mining, NLP, Data mining

Keywords

Online, informal, formal, tagging, parse

1. INTRODUCTION

Among the 2 main types of textual information - facts and opinions, a major portion of current information processes methods such as web search and text mining work with the former. Opinion Mining refers to the broad area of natural language processing, computational linguistics and text mining involving the computational study of opinions, sentiments and emotions expressed in text. A thought, view, or attitude based on emotion instead of reason is often referred to as a sentiment. Hence, an alternate term for Opinion Mining, namely Sentiment Analysis. This field finds critical use in areas where organizations or individuals wish to know the general sentiment associated to a particular entity - be it a product, person, public policy, movie or even an institution. Opinion mining has many application domains including science and technology, entertainment, education, politics, marketing, accounting, law, research and development.

The purpose of the analysis is to extract, organize, and classify the information contained in the required documents. The main highlights are Stanford parser and POS tagging where different tags are used to identify the different parts of speech in the opinion sentences that are parsed. Here the document is applied to formal and informal classifiers. The primary phase of the project is to categorize the document into formal and informal text. Now all the previous related works follow a different meaning for the informal part whereas the

idea implemented here is quite different. Here the concept of informal style is related to the sms (short message service) style of writing. To overcome this problem, a dictionary of common and related short form words has been created. Thus this dictionary will be used to further refer to what is the original actual word and then further processing takes place. The large set of review data is cleaned and divided into small chunks of data. This is followed by the application of the POS tags and finally the identification of features and their respective opinions. The output is a list of features and their opinions. Although the concept is somewhat different from all the existing systems, however, other kinds of preprocessing and feature extraction models may be tested for a better recognition rate in the future research in opinion mining system. The POS tagging method which is incorporated in this work could be improved to handle large variety of words that occur often in the review documents. There is a huge scope for designing the complete model for feature extraction of informal reviews.

2. RELATED WORK

The work is partly based on and closely related to opinion mining and sentence sentiment classification. Extensive research has been done on sentiment analysis of review text and subjectivity analysis (determining whether a sentence is subjective or objective). Another related area is feature/topic-based sentiment analysis, in which opinions on particular attributes of a product are determined. Most of this work concentrates on finding the sentiment associated with a sentence (and in some cases, the entire review). There has also been some research on automatically extracting product features from review text. Though there has been some work in review summarization, and assigning summary scores to products based on customer reviews, there has been relatively little work on ranking products using customer reviews.

The following are some of the related work proposed by the researchers in the field of opinion mining and sentiment analysis.

Yin-Fu Huang and Heng Lin have proposed a product ranking system where users can specify product features to get back the ranking results of all matched products [9]. Three issues have been considered while calculating product scores: 1) product reviews, 2) product popularity, and 3) product release month. The experimental results show that the system is practical and the ranking results are interesting. Especially, the system can be used to find the release year distribution of top-ten products with specified features. The results reveal that new products are not always more favorable than old products.

Francis Heylinghen and Jean Marc Dewaele have given the definition, measurement and behavior of formality of languages [2]. They have classified and categorized formal as surface formal and deep formal. A method to determine the

degree of formality for any text using a special formula is proposed. This formula is the F-score measurement which is based on the frequencies of different word classes (noun, verbs, adverbs, etc.) in the corpus. The texts with high F-score are considered formal, while the ones with low F-score are considered informal. In our work, we want to build a model based on main characteristics of the two styles, rather than based on the frequency of word classes.

Sowmya Kamath S, Anusha Bagalkotkar, Ashesh Khandelwal, Shivam Pandey, Kumari Poornima have studied the various existing approaches for performing sentiment analysis and classification [7]. Different approaches have been discussed related to sentiment analysis in e-learning, news videos and twitter messages. Their study has more scope related to analyzing bias in online content.

A. Kennedy and M. Shepherd discussed the task of web page classification by genre, namely how to distinguish home pages from non-home pages as noise, and then classify those home pages as personal home page, corporate home page or organization home page [4]. The corpus they used is rather small: 312 web pages. They tried the hard task of subgenre discrimination. The best accuracy they obtained is (71.4%) on personal home pages with a single classifier, manual feature selection, and without noisy pages. Has a problem in identifying organizational home page and identification and classification of noise is not included.

Ahmad Kamal, Muhammad Abulaish, and Tarique Anwar implemented a rule-based system, which applies linguistic and semantic analysis of texts to mine feature-opinion pairs which are then modeled using a bipartite graph structure [14]. Considering feature-opinion pairs as hubs and source documents as authorities, Hyperlink-Induced Topic Search (HITS) algorithm is applied to generate reliability score for each feature-opinion pair with respect to the underlying corpus. The system is able to identify product features and opinions that are related either directly or indirectly. The rule set needs to be refined in order to identify more dependency relationships to improve the precision and recall values of the proposed system and to identify implicit features. The system does not handle informal texts.

Yu-shan Chang and Yun-Hsuan Sung have deployed the Name Entity Recognition (NER) method which is applied to informal document, such as E-mail message and Newsgroup postings [5]. Maximum Entropy Classifier (Max-Ent) and Conditional Random Field (CRF), approaches are used in this system. They probably need more global features and their system has relatively lower recall than precision.

V.K. Singh, R. Piryani, A. Uddin and P. Waila have worked on the sentiment analysis of movie reviews where the textual reviews of a movie are analyzed and a sentiment label is assigned on each aspect [12]. Senti-WordNet based scheme is used to compute the document-level sentiment for each movie reviewed and then the results are compared with the results obtained using Alchemy API. The results obtained are more accurate and focused. It explores the use of 'Adverb + Verb' combine with 'Adverb + Adjective' combine for document-level sentiment classification. It proposes a new feature-based heuristic scheme for aspect-level sentiment classification of a movie.

Pasi Tapanainen and Timo Jarvinen suggested a practical parser for unrestricted dependencies [6]. The parser creates links between words and names the links according to their syntactic functions. The framework describes non-projective

dependency grammars and applies them efficiently. The output contains more information which is more accurate and explicit. The ambiguity rate is reduced to a quarter without any compromise in correctness.

Luole Qi and Li Chen have used the Conditional Random Fields (CRFs) model to perform the opinion mining tasks [8]. The algorithm's ability in mining intensifiers, phrases and infrequent entities, optimized training and decoding process is highlighted. Although the system is recommendable more experiments need to be conducted to compare CRFs-based approach with other non-model approaches.

Fadi Abu Sheikha and Diana Inkpen [1] have classified opinions as formal and informal. They have used Decision trees, Naive Bayes and SVM methods to do the same. More accurate classifiers required for the system. They have made use of WEKA, a collection of machine learning algorithms for the data mining tasks. The list for formal and informal words is very limited.

Weishu Hu, Zhiguo Gong, and Jingzhi Guo have proposed a Senti-WordNet based algorithm [10]. There are three steps to perform the task: (1) identifying opinion sentences in each review which is positive or negative via SentiWordNet; (2) mining product features that have been commented on by customers from opinion sentences; (3) pruning feature to remove those incorrect features. According to the algorithm, only features of the product in opinion sentences are mined. Compared to the previous work, the experimental result achieves higher precision and recall, but the algorithm has not been implemented practically.

Lizhen Liu, Zhixin Lv and Hanshi Wang have proposed a novel method to deal with the feature-level opinion mining problems [11]. The proposed method considers the explicit features and the implicit features of the opinions. The opinion words are divided into two categories, vague opinion words and clear opinion words, which are then used to identify the implicit features and cluster the features. The feature clustering depends on three aspects: the corresponding opinion words, the similarity of the features and the structures of the features. Small scale corpus cannot perform well with this method. The structure of the vague opinions dictionary and part-of-speech dictionary increases the cost of the method too.

K.B. Dempsey, P.M. McCarthy, and D.S. McNamara proposed that phrasal verbs can be used as a text genre identifier [3]. Their results indicate that phrasal verbs significantly distinguish between both the spoken/written and the formal/informal dimensions. Their experiments are performed on the frequency of occurrence of phrasal verbs in spoken versus written text and in formal versus informal texts. Their study is beneficial for textbook writers, researchers and for computational tools like Coh-Metrix. Large corpus of data is needed for this system.

Liu Gongshen, Lai Huoyao, Luo Jun, and Lin Jiuchuan [13] have proposed, a new vector representation of texts where total score computing and SMV are used for predicting their semantic orientation. The texts are represented as vectors with nouns as dimensions and their related polarity scores as dimensional values. In this way, topic correlation is taken into account when assigning polarity labels for reviews. Experiment results show that the model is effective for identifying sentiment polarity of reviews. In order to get a better predicting accuracy, some improvements should be introduced in future work. Use of more refined adjective

scoring algorithms can be used to generate elemental sentiment scores. More subjective words such as verbs and adverbs can be considered with effects of valence shifters.

3. SYSTEM ARCHITECTURE

The need to identify and interpret possible difference in the linguistic style of text, such as formal or informal is increasing, as more and more people are using the Internet as their main research resource. In this section, the architecture and functional detail of the proposed opinion mining system to identify feature-opinion pairs is presented.

The proposed system takes care of the informal reviews. The idea of formal and informal words has been revised little in the proposed system. The use of English language can be grammatically incorrect with chances of spelling mistakes and wide use of shortcuts. Technically they can be called as formal opinions and informal opinions, where formal opinions refer to the use of proper English with no grammatical mistakes and informal refers to the improper use of English with grammatical mistakes, involving use of words which are not standard spellings e.g. “4get” instead of “forget”. They say that the UK English is said to be formal English and US English is considered to be informal English. Most of the customers use US English, hence it is very important to concentrate on these opinions too, or else they might be discarded as noise. Hence the system proposed deals with this aspect of informal reviews.

The review documents are the documents that contain the user reviews, opinions or recommendations that are related to any particular product or service. These review documents are the main source of information that will be required for mining. In this module the crawler retrieves review documents from the review site on the Web. Then locate and download the reviews. The sources of data for reviews can contain noise or unwanted data. Its needs to be cleansed as a part of the mining process. Hence after the review document has been acquired, it is processed, through the review cleaning or filtering. Filtering process filters out or removes the noisy reviews. After removing noisy reviews the remaining data reviews are classified according to the formal and informal style. The review classification depends upon the use of English language. As described earlier while commenting on the products and services the customers do not conform to the standard use of the language. Filtered review documents are divided into manageable record size chunk. This is assigned as an input to the document preprocessor and then to Parts of Speech (POS) tagger which tags each word, like the Stanford Parser. It converts each sentence into a set of dependency relationship between pair of words. In Feature and opinion identification module we represent the dependency relations between a pair of words w_1 and w_2 as a relation type (w_1, w_2) , in which w_1 is called the head or governor and w_2 is called dependent or modifier. This may be direct or indirect relation type id. In direct relation type, one word depends on other directly and in indirect, one through other word or both of them depends on third word indirectly. For informal text for example we are in a dependency relation R , if there exists a abbreviation (w_1, w_2) relation such that $POS(w_1) = NN$, $POS(w_2) = JJ$, and w_1 and w_2 are not stop-words then w_2 is assumed to be an opinion and w_1 as an feature. Opinion words and features are generally associated with each other and consequently, there exists inherent as well as semantic relations between them. Therefore, the feature and opinion learner module is implemented as a rule-based system, which analyzes the dependency relations to identify information components from review documents.

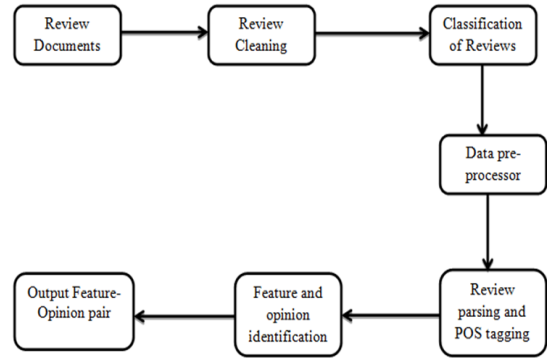


Fig 1: System Architecture

The table below, shows how formal and informal are actually classified.

Table 1. List of Formal and Informal words

Formal	Informal
approximately	About
in addition	And
anyone	Anybody
request	ask for
employer	Boss
however	But
purchase	Buy
finish	End
sufficient	Enough
must	have to
are not	aren't
cannot	can't

But as discussed earlier, nowadays, most of the people have started using the sms language. Hence there is a large set of review data which will contain opinions in the short-form/sms language. Let us look at the following table to get an idea.

Table 2. Proposed list of formal and informal words

Formal	Informal
I am	M
Behave	Bhv/behv
Have	Hv
This	Dis
Good	Gud
Awesome	Awsum
Different	Diff
Bright	Brite
Working	Wrkin

Now all the previous systems used to eliminate such words and opinions as noise, which is not the right thing. Since most of the users are used to using this form of language, such opinions should not be ignored. As shown in table 1 above, the informal words are the words which can be used in casual conversation; they are not basically short forms of the words. This concept of informal words has been changed here, where now the sms form of words shall be considered as informal style. Thus such valuable opinions will not be discarded as noise and will be considered for further processing.

4. RESULTS

Results can be evaluated using standard Information Retrieval (IR) metrics i.e. Precision, Recall and F-score that are defined in equations given below, respectively.

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (4.1)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (4.2)$$

Calculating precision and recall is actually quite easy. Imagine there are 100 positive cases among 10,000 cases. You want to predict which ones are positive, and you pick 200 to have a better chance of catching the 100 positive cases. You record the IDs of your predictions, and when you get the actual results you sum up how many times you were right or wrong. There are four ways of being right or wrong:

TN / True Negative: case was negative and predicted negative

TP / True Positive: case was positive and predicted positive

FN / False Negative: case was positive but predicted negative

FP / False Positive: case was negative but predicted positive

The output graphs for both are shown below. In the first one, reviews containing proper English grammar and spellings are selected and processed. The output graph is as follows:

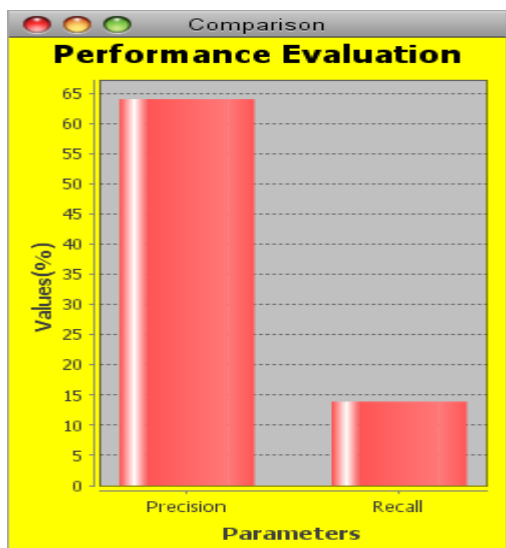


Fig 2: Precision and Recall graph

In the second output graph, the proposed idea takes form, where manually written reviews are selected, which contains sms(short message service) language. These reviews are fed into the system, which yields the following output:

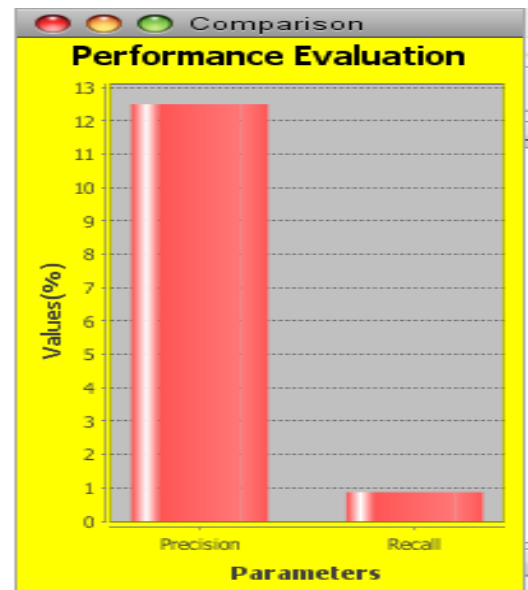


Fig 3. Precision and Recall for sms language reviews

None of the previous systems have deployed an idea where the short messaging language; namely the sms language is taken into consideration. The motivation for the same was acquired from the thought that majority of the people use short cuts while writing or chatting.

The following table will help to understand the difference between the existing and the proposed system.

Table 3. Comparison of Results

	Precision	Recall
Existing System	0.985	0.985
Proposed System without sms language	0.64	0.16
Proposed System with sms language	0.13	0.015

Although a direct comparison is not made, both the systems performance can be seen. The existing system has a higher precision and recall because they work with proper English words. Thus a higher precision and recall is bound to occur, and hence the proposed system has a lower precision and recall since there is an inclusion of the sms language words. This makes the system different from the existing system. But in due course of time the system can be upgraded to give better results in future.

5. CONCLUSION

A great issue has been addressed here regarding the informal reviews. After implementation of the system, formal as well as informal opinions can be recognised. An attempt has been made to process the informal reviews. There are many factors which affect the performance of the system with regards to the informal reviews. Accuracy of extraction of features from the informal reviews totally depends on the segmentation and

tagging process. The experiments have illustrated that the POS tagging concept can be applied successfully to solve the informal reviews problem.

Although the concept is somewhat different from all the existing systems, however, other kinds of preprocessing and feature extraction models may be tested for a better recognition rate in the future research in opinion mining system. The POS tagging method which is incorporated in this work could be improved to handle large variety of words that occur often in the review documents. There is a huge scope for designing the complete model for feature extraction of informal reviews. Finally to conclude opinion mining is a very interesting and hot topic in recent years and is finding a lot of recognition. Hoping that in the coming years, the level of opinion mining is advanced and fool proof and finds importance in the human-race.

6. REFERENCES

- [1] Fadi Abu Sheikha and Diana Inkpen, "Automatic Classification of documents by formality", IEEE 2010.
- [2] Francis Heylinghen and Jean-Marc Dewaele, "Formality of language: definition and measurement", Internal Report, Center "Leo Apostel", Free University of Brussels, 1999.
- [3] K.B. Dempsey, P.M. McCarthy, and D.S. McNamara, "Using phrasal verbs as an index to distinguish text genres", In D. Wilson and G. Sutcliffe (Eds.), Proceedings of the twentieth International Florida Artificial Intelligence Research Society Conference (pp. 217-222). Menlo Park, California: The AAAI Press, Feb. 2007.
- [4] A. Kennedy and M. Shepherd, "Automatic Identification of Home Pages on the Web", Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.
- [5] Yu-shan Chang and Yun-Hsuan Sung, "Applying Name Entity Recognition to Informal Text", Ling 237 Final Projects, 2005.
- [6] P. Tapanainen and Jarvinen Timo, "A nonprojective dependency parser", In Proceedings of the 5th Conference on Applied Natural Language Processing, pages 64-71, Washington D.C. Association for Computational Linguistics, 1997.
- [7] Sowmya Kamath S, Anusha Bagalkotkar, Ashesh Khandelwal, Shivam Pandey, Kumari Poornima, "Sentiment Analysis Based Approaches for Understanding User Context in Web Content", IEEE International Conference on Communication Systems and Network Technologies, pp. 607-611, 2013.
- [8] Luole Qi and Li Chen, "Comparison of Model-Based Learning Methods for Feature-Level Opinion Mining", IEEE International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 265-273, 2011.
- [9] Yin-Fu Huang and Heng Lin, "Web Product Ranking Using Opinion Mining", IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 184-190, 2013.
- [10] Weishu Hu, Zhiguo Gong and Jingzhi Guo, "Mining Product Features from Online Reviews", IEEE International Conference on E-Business Engineering, pp. 24-29, 2010.
- [11] Lizhen Liu, Zhixin Lv and Hanshi Wang, "Opinion Mining Based on Feature-Level", IEEE International Congress on Image and Signal Processing (CISP), pp. 1596-1600, 2012.
- [12] V.K. Singh, R. Piryani, A. Uddin, P. Waila, "Sentiment Analysis of Movie Reviews - A new Feature-based Heuristic for Aspect-level Sentiment Classification", IEEE, pp. 712-717, 2013.
- [13] Liu Gongshen, Lai Huoyao and Luo Jun, Lin Jiuchuan, "Predicting the Semantic Orientation of Movie Reviews", IEEE International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), pp. 2483-2487, 2010.
- [14] Ahmad Kamal, Muhammad Abulaish, Tarique Anwar, "Mining Feature-Opinion Pairs and Their Reliability Scores from Web Opinion Sources". WIMS" 12, June 13-15, Craiova, Romania, 2012.