Analysis of Feature Classification as Formal or Informal of Online Reviews Samples

Jennifer Selvaraj, Dr. J. W. Bakal, Department of Computer Engineering, Mumbai University, Mumbai, India.

Abstract— **The Internet has become the hub of information for almost all the people. Hence with the rapid development of web, most of the customers express their opinions on various kinds of entities, such as products and services on the web. These reviews provide useful information to customers for reference. These reviews are also valuable for merchants to get constructive feedback from customers and improve the qualities of their products or services. However, the contents available are not in the standard format. We are classifying these contents as formal and informal. The extracted feature-opinion pairs and sentence-level review source documents can be modelled using a graph structure.**

Keywords— **mining, online, informal, formal.**

I. Introduction

As the exponential explosion of various contents generated on the Web has been increasing enormously, recommendation techniques have become increasingly indispensable in recent times. Innumerable different kinds of recommendations are made on the Web every day, including music, images, books recommendations, query suggestions, etc. Thus there is a need to extract, classify, and understand these opinions expressed in various online sources. Here opinion mining refers to computational techniques for analysing the opinions that are extracted from various sources. Current opinion research focuses on business and e-commerce such as product reviews and movie ratings.

In the real world, businesses and organizations always want to find consumer or public opinions about their products and services. Individual consumers also want to know the opinions of existing users of a product before purchasing it, and others' opinions about political candidates before making a voting decision in a political election. In the past, when an individual needed opinions, he/she asked friends and family. When an organization or a business needed public or consumer opinions, it conducted surveys, opinion polls, and focus groups. Acquiring public and consumer opinions has long been a huge business itself for marketing, public relations, and political campaign companies. Opinions and its related concepts such as sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining. The inception and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, micro blogs, Twitter, and social networks, because for the first time in human history, we have a huge volume of opinionated data recorded in digital forms. Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing. It is also widely studied in data mining, Web mining, and text mining. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole. In recent years, industrial activities surrounding sentiment analysis have also thrived. Numerous start-ups have emerged. Many large corporations have built their own in-house capabilities. Sentiment analysis systems have found their applications in almost every business and social domain.

Due to all these advances, industrial activities have flourished in recent years. Applications related to sentiment analysis have spread to almost every possible domain, from consumer products, services, healthcare, and financial services to social events and political elections. Many big corporations have also built their own in-house capabilities. These practical applications and industrial interests have provided strong motivations for research in sentiment analysis.

II. Related Work

Considerable amount of work has already been done in this field where different approaches like the rule based approach, the stochastic approach and the transformation based learning approach along with modifications have been tried and implemented. However, if we look at the same scenario for South-Asian languages such as Bangla and Hindi, we find out that not much work has been done, as much as it is in English. The main reason for this is the unavailability of a considerable amount of annotated corpora of sound quality.

Francis Heylinghen and Jean Marc Dewaele have given the definition, measurement and behaviour 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. In [16], Turney proposed an algorithm to classify a review as positive or negative, which applies POS analysis to identify opinion phrases in review documents and uses PMI-IR algorithm [15] to identify their semantic orientations. 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.

In [13], the authors have proposed a supervised pattern mining method, which identifies product features from pros and cons sections of the review documents in an automatic way. 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. In [13] the author proposed a feature extraction method that uses ontology for opinion mining. Although this method worked well semantically, the main problem is the maintenance of the ontology to address the constant expansion of the review data. In this system, the ontology is manually constructed and when new features are added it must be updated. In addition, a concept that is defined in the ontology is only able to be classified. Thus, it is necessary to construct an automatic system to avoid continued intervention.

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 Senti-WordNet; (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.

Reference [10] proposed rule based system for feature extraction method. This method extracts a relatively large number of features compared with the amount of review data. For example, it generates 189 features from 50 reviews for digital cameras. The main reason for the extraction of so many features is that terms that have the same or similar meanings are not considered as the same features but they are considered as different words. Consequently, this system could not provide proper summary information for the product. This problem is solve in FEROM in that the number of features are reduced by merging words that have similar meanings using the semantic similarity between features and then providing reliable summary information for the product based on the merged features.

III. Proposed System

The purpose of the analysis is to extract, organize, and classify the information contained in the required documents. The proposed method is based on object-oriented approach to software development.

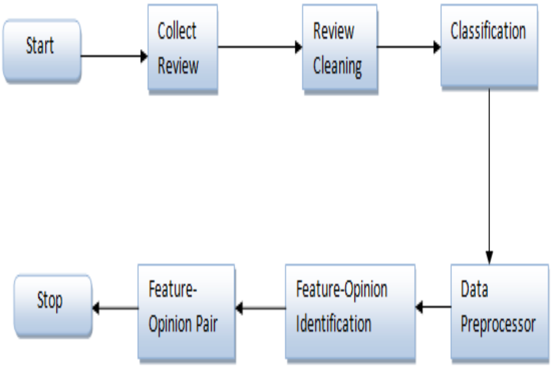

Fig 1. State diagram representing the system

Initially the crawler retrieves reviews document from sources such as web. These sources could range from small blogs to bigger review sites. Then locate and download the reviews. The language we are dealing here with is the English language. After that review document is processed to review cleaning or filtering. In the filtering process, we will filter out or remove noisy review. Noisy data could be related to any unwanted symbols used, double words, improper definitions etc. After removing noisy review classify the remaining data review according to formal and informal style [2].

TABLE I

Sr. No	Formal	Informal
1	Purchase	Buy
2	Finish	End
3	Obtain	Get
4	Are not	Aren’t
5	Good	gud

The above table describes how we classify the review. Filtered review document are divided into manageable record size chunk. Filtered review document are divided into manageable record size chunk. This is assign as input for document preprocessor to Parts of Speech tag (POS) to each word, like Stanford Parser. It converts each sentence into set of dependency relationship between pair of words. This is how generally how every system works, but it will alternate if the desired end result differs. Various methods have been proposed which have different methods. These reviews can be about places, goods, services, people or any idea in which a particular is interested in. Hence the system and its approach will differ according to the intended application.

Vocabulary choice is likely the biggest style marker. In general, longer words and Latin origin verbs are formal, while phrasal verbs and idioms are informal. There are also many formal/informal style equivalents that can be used in writing. The formal style is used in most writing and business situations, and when speaking to people with whom we do not have close relationships. Some characteristics of this style are long words and using the passive voice. Informal style is mainly for casual conversation, like at home between family members, and is used in writing only when there is a personal or closed relationship, such as that of friends and family. Some characteristics of this style are word contractions such as “won’t”, abbreviations like “phone”, and short words.

A. Algorithm

In a dependency relation R, if there exist relationships nn(w1; w2) and nsubj(w3; w1) such that POS (w1) = POS (w2) = NN, POS (w3) = JJ and w1, w2 are not stop-words,

Or

If there exists a relationship nsubj(w3; w4) such that POS (w3) = JJ, POS (w4) = NN and w3, w4 are not stop-words, then either (w1; w2) or w4 is considered as a feature and w3 as an opinion.

IV. RESULTS AND DISCUSSION

Results can be evaluated using standard Information Retrieval (IR) metrics Precision, Recall and F-score. 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.

V. CONCLUSIONS

A great issue has been addressed here regarding the informal reviews. After implementation of the system we are able to recognize formal as well as informal opinions. An attempt is made to process the informal reviews. There are many factors that 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.

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.

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.

[15] Marie-Catherine de Marneffe and D.Manning, *Stanford typed dependencies manual*. Revised for Stanford Parser v.1.6.9 in September.

[16] P. D. Turney. *Mining the web for synonyms: Pmi-ir versus lsa on toe*. In Proceedings of the 12th Euro- pean Conference on Machine Learning, EMCL '01, pages 49-502. Springer-Verlag, 2001.