ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 05, pp.31701-31705, May, 2016

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

SENTIMENT ANALYSIS OF COMPARATIVE SENTENCES IN TEXT DOCUMENT USING OSA AND PMI TECHNIQUES

[1,*]Aman Singla and [2]Gurwinder Singh

[1]Thapar Polytechnic College, Patiala
[2]BBSBEC, Fatehgarh Sahib, India

ARTICLE INFO

Article History:
Received 23rd February, 2016
Received in revised form
25th March, 2016
Accepted 14th April, 2016
Published online 31st May, 2016

Key words:

Sentiment, Opinion, Positive opinions, Negative opinions, Opinion Mining, Sentiment analysis.

ABSTRACT

Textual information in the world can be broadly categorized into two main types: facts and opinions. Facts are objective expressions about entities, events and their properties. Opinions are usually subjective expressions that describe people's sentiments, appraisals or feelings toward entities, events and their properties. With the growing availability of online resources on web and popularity of fast and rich resources of opinion sharing such as online review sites and personal blogs, Opinion Mining has become an interesting area of research. Identifying sentiments from an opinion is a challenging problem. For a popular product, the number of reviews can be in hundreds or even more. This makes it difficult for a customer to read them to make an informed decision on whether to purchase the product. It also makes it difficult for the manufacturer of the product to keep track and to manage customer opinions. For the manufacturer, there are additional difficulties because many merchant sites may sell the same product. In this research, we aim to mine and to summarize all the customer reviews of a product and summarize whether the opinions are positive or negative.

Citation: Aman Singla and Gurwinder Singh, 2016. "Sentiment Analysis of Comparative Sentences in Text Document using OSA and PMI Techniques", *International Journal of Current Research*, 8, (05), 31701-31705.

INTRODUCTION

Data mining sometimes called data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. This makes Data Mining techniques a useful necessary tool in many applications, especially in systems where huge amount of data has to be analyzed, something infeasible for a human being to perform manually. One of the most important ways of evaluating an entity or event is to directly compare it with a similar entity or event. The objective of this work is to extract and to analyze comparative sentences in evaluative texts on the Web, e.g., customer reviews, forum discussions, and blogs. This task has many important applications. For example, after a new product is launched, the manufacturer of the product wants to know consumer opinions on how the product compares with those of its competitors. Extracting such information can help businesses in its marketing and product benchmarking efforts.

Mining an Opinion: Opinion Mining is a field of Web Mining that aims to find valuable information out of user's opinions. As the usage of e-commerce is increasing year after year many people had changed the habit of going to a shop for the comfortable virtual shopping. Mining opinions on the web is a fairly new subject, and its importance has grown significantly mainly due to the fast growth of e-commerce, blogs and forums. A major problem however, is finding the desired information on the products. It is not difficult to find web sites with thousands of reviews for a single product, and thus finding any useful information among them can be a very difficult task, so it is important to differentiate between types of opinion holders. An expert opinion is usually far superior in quality, richer in technical details, and goes through all the most relevant aspects of a product. Users and customers usually give opinions with less commitment. Basic components of an opinion:

Object: on which an opinion is expressed.

Corresponding author: Aman Singla,
Thapar Polytechnic College, Patiala

Opinion holder: The person or organization that holds a specific opinion on a particular object.

Opinion: is a personal belief or judgment of a subject.

Opinion Orientation: The orientation of an opinion on a feature indicates whether the opinion is positive, negative or neutral.

Sentiment Analysis: Textual information includes two types of information in it: facts information and opinion information. Facts information is objective statement about objects, and opinion information is subjective statement that expresses person's opinion about objects. The rise of World Wide Web brings us many user generated information (e.g. forum post, blog, review), which contains a large number of opinion information. When one wants to see how well one product is, he or she wants to buy it, it is not necessary to ask other friends if we can fetch opinion information about the product on Web. All these reasons push the development of research on opinion mining and sentiment analysis.

Part of Speech Tagger: Part-of-speech (POS) tagging is the task of determining the correct parts of speech for a sequence of words. POS tagging is useful for a large number of applications: It is the rst analysis step in many syntactic parsers.It is used in information extraction, speech synthesis, lexicographic research, term extraction, and many other applications. A large number of methods have been applied to POS tagging over the years. Among them are Hidden Markov Models (Church, 1988; Cutting et al., 1992; Brants, 2000), transformation-based learning (Brill, 1992), memory-based learning (Daelemans et al., 1996), maximum-entropy modeling (Ratnaparkhi, 1996), support vector machines (Gim´enez and M`arquez, 2004), neural networks (Benello et al., 1989etc. The typical accuracy of POS tagger is between 95 % and 98 % depending on the tagset, the size of the training corpus, the coverage of the lexicon, and the similarity between training and test data. One special application of natural language processing is determining the part of speech of each word in a sentence, known as part-of-speech (POS) tagging. For Example: *The speed of Mozilla Firefox is better than internet explorer*. After applying POS Tagger the following tags are annotated with the different words of the sentence. *The*/DT *speed*/NN *of*/IN *Mozilla*NN *Firefox*/NN *is*/VBZ *better*/JJR *than*/IN *internet*/NN *explorer*/NN.

Comparative Sentences and their classification: A comparative sentence expresses an ordering relation between two sets of entities with respect to some common features. In this work, we focus on comparisons. Clearly, product comparisons are not only useful for product manufacturers, but also to potential customers as they enable customers to make better purchasing decisions.

Types of Comparatives

We group comparatives into four types. The first three of which are gradable comparatives and the fourth one is non-gradable comparative. The gradable types are defined based on the relationships of greater or less than, equal to, and greater or less than all others. These are:

Non-Equal Gradable: Relations of the type greater or less than that express an ordering of some objects with regard to certain features.

Equal Gradable: Relations of the type equal to that state two objects as equal with respect to some features.

Superlative: Relations of the type greater or less than all others that rank one object over all others.

Non-Gradable: Sentences which compare features of two or more objects, but do not grade them.

Related Work

Sentiment analysis has been studied by many researchers recently. Two main directions are sentiment classification at the document and sentence levels, and feature-based opinion mining. Sentiment classification at the document level investigates ways to classify each evaluative document as positive or negative (Pang et al., 2002; Turney 2002). Sentiment classification at the sentence-level has also been studied (e.g., Riloff and Wiebe 2003; Kim and Hovy 2004; Wilson et al., 2004; Gamon et al., 2005; Stoyanov and Cardie 2006). These works are different from ours as we study comparatives. Fiszman et al. (2007) studied the problem of identifying which entity has more of certain features in comparative sentences. It does not find which entity is preferred. Soo-Min Kim and Eduard Hovy[14] presented a system that, given a topic, automatically finds the people who hold opinions about that topic and the sentiment of each opinion.YeongHyeonGu and SeongJoonYoo (2009) proposed a study of comparative online opinions which is about sorting comparative sentences out of given sentences. Their work focused on the documents in Korean, may be the first of its kind in Korea although there have been a few of such studies in English spoken countries. The works in (Hu and Liu 2004; Liu et al 2005; Popescu and Etzioni 2005; Mei et al 2007) perform opinion mining at the feature level. The task involves (1) extracting entity features (e.g., "*picture quality*" and "*battery life*" in a camera review) and (2) finding orientations (positive, negative or neutral) of opinions expressed on the features by reviewers. Again, our work is different because we deal with comparisons. Discovering orientations of context dependent opinion comparative words is related to identifying domain opinion words (Hatzivassiloglou and McKeown 1997; Kanayama and Nasukawa 2006). Both works use *conjunction rules* to find such words from large domain corpora. One conjunction rule states that when two opinion words are linked by "and", their opinions are the same.

Problem Statement

This work studies a text mining problem related to sentiment analysis from the user generated content on the web ,In particular it focuses on mining opinions from comparative sentences i.e., to determine which entities in a comparison are preferred by the user.

Objectives

Given the comparative opinions the work proposes the following objectives:

- To categorize the comparative sentences into different types.
- To extract the comparative relation from a comparative sentence.
- To identify which of the entities used in the sentence has positive orientation.

Proposed Technique

We now present the proposed technique. As discussed above, the primary determining factors of the preferred entity in a comparative sentence are the feature being compared and the comparative word, which we conjecture, form the context for opinions (or preferred entities). Here we are evaluating our results by using two techniques one is OSA i.e One side association and the second is PMI i.e Pointwise Mutual Informtion. We develop our ideas from here.

Identifying Gradable Comparatives

From the large collection of online opinions the gradable comparative sentences are figured out. Some are defined to identify the gradable comparatives from the opinion text.

These rules are:

Standard comparatives: Comparatives and Superlatives having standard words that express gradable comparatives suffixes "-er", "-est". Sentences formed with more, most, less, least, better, best, worse, worst, further/farther, furthest/farthest. The words which are tagged as JJR, JJS, RBR and RBS are commonly observed as standard Comparatives.

Non-standard words that express gradable comparisons like prefer, superior E.g. "In term of battery life, Kodak is superior to Canon" Kodak is preferred.

Analysis of Comparative Relation

A comparative relation captures the essence of a comparative sentence and is represented with the following parameters:

(Comparative word, Feature, Entity1, Entity2, Type of Comparative)

Sentiment detection: Adjectives in a sentence carry the sentiments. Words with POS tags of JJR, RBR, JJS, and RBS are the indicators of comparative words as in the above example the word "better" which is tagged as JJR is a comparative word.

Feature detection: words with POS tags NN, NNP and NNS may be the feature of an entity. All those words that are tagged as NN, NNP and NNS are extracted from the tagged sentence and to identify the feature, these extracted words are compared with the words present in feature set.

Entity detection: The two entities being compared.Entity1 appear to the left of the relation word in a comparative sentence and entity2 appear to the right of the relation word in a comparative sentence. The words tagged as NN, NNP and NNS may act as entities in a comparative sentence. All the words tagged as NN, NNP and NNS are extracted and then comparison algorithm is used to find both the entities which are being compared

Type of Comparison: Type of a sentence may be gradable and non-gradable. Our study is only limited to gradable comparative sentences so every comparative sentence associated with the type "Gradable Comparative. For Example: A comparative sentence "*The speed of Mozilla Firefox is better than Internet Explorer.*"

A comparative relation **<better, speed, Mozilla Firefox, Internet explorer, Gradable>** is extracted from the sentence.

Further analysis also shows that we can group comparatives into two categories according to whether they express increased or decreased values: *Increasing comparatives*: Such a comparative expresses an increased value of a quantity, e.g., "*more*", and "*longer*". *Decreasing comparatives:* Such a comparative expresses a decreased value of a quantity, e.g., "*less*", and "*fewer*". As we will see later, this categorization is very useful in identifying the preferred entity.

Identifying Preferred Entities

To find the preferred entity in a comparative sentence denote comparative word by *cw* and the feature being compared by *f* Different cases are:

1. If Comparative word *cw* is opinionated then we check the sentimental orientation of the comparative word.

If *cw* has positive orientation then Preferred entity = Entity1
Else

Preferred entity = Entity2

2. If Comparative word *cw* is not opinionated but the Feature *f* being compared in the sentence is opinionated then we check the sentimental orientation of the feature *f* and we follow the following steps:

If orientation of *f* = positive and *cw* is increasing comparative word then

Preferred entity = Entity1

Else

Preferred entity = Entity2

3. The orientation of an opinion sentence depends upon the comparative word and the feature used in the comparative sentence. To find the orientation of the sentence following rules are applied:

i) Increasing comparative+ Negative Feature → Negative Opinion

The first rule says that the combination of an increasing comparative word with a negative opinion adjective or adverb implies a negative orientation of the sentence and entity2 is preferred.

ii) Increasing comparative + Positive Feature → Positive Opinion

The Second rule says that the combination of an increasing comparative word with a positive opinion adjective or adverb implies a positive orientation of the sentence and Entity 1 is preferred.

iii) Decreasing comparative+ Negative Feature→ Positive Opinion

The third rule says that the combination of a decreasing comparative word with a feature word of negative orientation gives a positive orientation of the sentence and entity1 is preferred.

iv) Decreasing comparative + Positive Feature → Negative Opinion

The fourth rule says that when a decreasing comparative word combines with a feature word of positive orientation it implies the negative orientation of sentence and entity2 is preferred.

Evaluation

Evaluation Datasets and Results

A system called Finding Positive Entity in Comparative Sentences (FPECS) is implemented based upon the proposed technique. Our comparative sentence dataset consists of two subsets. The first subset is from (www.toptenreviews.com), which are product reviews sentences on Samsung galaxy S7 and iphone 6s. The original dataset also contains many non-gradable comparative sentences, which are not used here as most such sentences do not express any preferences. To make the data more diverse, we collected more product reviews sentences about various phone sets like htc, Lenovo, nexus from http://www.toptenreviews.com and www.epinions.com. Table 1 gives the number of sentences from these two sources.

Table 1. Comparative Sentences from different sources

Data Sources	No of Comparative Sentences
www.toptenreviews.com	85
www.epinions.com	35
Total	120

Accuracy, Recall and Precision

Training Dataset comprises of 120 sentences. Imagine there are 60 positive cases for entity1 i.e. Samsung Galaxy S7 among 120 cases. We want to predict which ones are positive and we pick 80 sentences to have a better chance of catching many of the 60 positive cases. We evaluate the sentences and sum up how many times we were right or wrong. There are four ways of being right or wrong:

True Negative: case was negative and predicted negative (TN)

True Positive: case was positive and predicted positive (TP)

False Negative: case was positive but predicted negative (FN)

False Positive: case was negative but predicted positive (FP).

To find the relation between the comparative word and the feature used in the sentence we have used two approaches, Point wise Mutual Information (PMI) and One Side Association (OSA). The results of these two methods are shown in the following tables.

Entity 1: Samsung Galaxys7

Table 2. Datasheet for Samsung Galaxys7

Entity1	Predicted Negatives	Predicted Positive
Negative Cases	TN = 38	FP= 19
Positive Cases	FN = 22	TP = 41

Entity 2: iphone 6s

Table 3. Datasheet for iphone 6s

Entity 2	Predicted Negatives	Predicted Positives
Negative Cases	TN= 40	FP= 11
Positive Cases	FN = 16	TP= 53

Table 4. Resulting figures of Accuracy, Recall and Precision using Point wise Mutual Information (PMI) method

Product	Accuracy for Sentences	Recall for Sentences	Precision for Sentences
Samsung Galaxy S7	80.4%	81.1%	61.6%
iphone 6s	65.33%	73.6%	66.1%

Table 5. Resulting figures of Accuracy, Recall and Precision using One Side Association (OSA) method

Product	Accuracy for Sentences	Recall for Sentences	Precision for Sentences
Samsung Galaxy S7	81.6%	83.33%	62.5%
iphone 6s	69.6%	80%	66.66%

Table 6. Comparison of PMI and OSA methods

	Entity1 Preferred			Entity2 preferred		
	Acc.	Rec.	Pre	Acc.	Rec.	Pre.
FPECS(PMI)	0.804	0.811	0.616	0.653	0.736	0.661
FPECS(OSA)	0.816	0.833	0.625	0.696	0.800	0.666

The developed system FPECS using One side Association (OSA) method shows the slightly better results in case of Accuracy for Entiy1 and Entity2 than FPECS using point wise Mutual Information (PMI).

Conclusion

This paper studied sentiments expressed in comparative sentences. This paper compares the results of the two

approaches PMI and OSA and the experimental results shows that the OSA method provides the better results in terms of accuracy, recall and precision than PMI.

REFERENCES

"Mining customer opinions from free text". IDA, 2005.

Fiszman, M., Demner-Fushman, D., Lang, F., Goetz, P. and Rindflesch, T. 2007. "Interpreting Comparative Constructions in Biomedical Text." BioNLP.

Ganapathibhotla, M. and Liu, B. 2008. "Mining Opinions in Comparative Sentences"

Hu and Liu, 2007. KDD-04; Liu, "Web Data Mining book"

Hu, M. and Liu, B. 2004. "Mining and summarizing customer reviews", Proceedings of the tenth ACM SIGKDD 04, August 22-25, 2004.

Jindal, N. and Liu, B. 2004. "Mining Comparative Sentences Relations." *AAAI'06*.

Jindal, N. and Liu, B. 2006. "Identifying Comparative Sentences in Text Document" *AAAI'06*.

Kanayama, H. and Nasukawa, T. "Fully automatic lexicon expansion for domain-oriented sentiment analysis." *EMNLP'06*

Kim, S. and Hovy, E. 2004. "Determining the Sentiment of Opinions." *COLING'04*.

Liu, B., Hu, M. and Cheng, J. 2005. "Opinion observer: Analyzing and Comparing Opinions on the Web." *WWW'*05.

Max Bramer, 2007. "Principels of Data Mining".Springer, 2007.

Popescu, A.M. and Etzioni, O. "Extracting Product Features and Opinions from Reviews." EMNLP'05

Simon Corston-Oliver Eric Ringger Michael Gamon, Anthony Aue

Soo-Min Kim and Eduard Hovy, 2004. Determining the Sentiment of Opinions. 2004. In Proceedings of Conference on Computational Linguistics (COLING-04). pp. 1367-1373. Geneva, Switzerland.

Stanford, 1992. "A simple rule-based part of speech tagger." *ANL*, 1992.

www.epinions.com

www.toptenreviews.com

****