International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 3 Issue 10, September 2016

ISSN: 2394-4404

Extracting Comparative Sentences From Text Documents Using POS Tagging

Ashwini D. Pawar

M. Tech Student, Dept. of Computer Science and Information Technology, Dr. B. A. M. University Aurangabad, Maharashtra, India

Sachin N. Deshmukh

Professor, Dept. of Computer Science and Information Technology, Dr B. A. M. University, Aurangabad, Maharashtra, India

Abstract: ***This paper studies the difficulty of distinguishing comparative sentences in text documents. The difficulty is related to but quite different from sentiment/opinion sentence recognition or classification. Sentiment classification studies the difficulty of classifying a document or a sentence based on the subjective opinion of the author. An important application area of sentiment/opinion recognition is business intelligence as a product manufacturer always wants to know consumers' impressions on its products. Comparisons on the other hand can be subjective or objective. Furthermore, a comparison is not concerned with an object in isolation. Instead, it compares the object with others. An example opinion sentence is "the sound quality of CD player P is poor". An example comparative sentence is "the sound quality of CD player P is not as good as that of CD player Q". Clearly, these two sentences give different information. Their language constructs are quite different too.***

Keywords: opinion mining, comparative sentences, genetic algorithm, social media mining.

I. INTRODUCTION

Comparative thoughts represent a way of users express their preferences about two or more entities. Mining comparative sentences from texts can be useful in several applications. For instance, a company might be interested in social media rumors of a new product release among consumers. Or, what are the best and worst features of the new product from consumer's viewpoint? Now days, social medias are great source of this kind of information and mining comparative impressions from them seems to be a very promising direction to unveil valuable Knowledge.

Many researches have been done in the field of regular opinion and sentiment classification. However, comparative impressions represent a different viewpoint of users and an interesting research area. A regular impression about a certain car C is a statement like "car C is ugly". On the other hand, a comparison is like "car C is much better than car D", or "car C is larger than car D". Clearly, these Sentences have rich information from which we can extract cognition with specific mining techniques.

II. RELATED WORK

Related work to ours comes from both computer science and linguistics. Researchers in linguistics focus primarily on defining the syntax and semantics of comparative conceptions. They do not deal with the distinguish of comparative sentences from a text document computationally. Studies the semantics and syntax of comparative sentences, but uses only limited vocabulary. It is not able to do our task of distinguishing comparative sentences. Discusses gradability of comparatives and measure of gradability. The semantic analysis is based on logic, which is not directly applicable to distinguishing comparative sentences. The types of comparatives (such as adjectival, adverbial, nominal, superlatives, etc). The concentration of these researches is on a limited set of comparative conceptions which have gradable keywords like more, less, etc. In summary, although linguists have studied comparatives, their semantic analysis of comparatives based on logic and grammars is more for human intake than for automatic recognition of comparative sentences

by computers. In text and data mining, we have not found any direct work on comparative sentences.

III. PROBLEM DEFINITION

In this section, we state the difficulty that we aim to solve. We first give a linguistic view of *comparatives* (also called *comparative constructions*) and discover some restrictions. We then enhance them by including implicit comparatives, and state the difficulty that we deal with in this paper. Since we need Part-Of-Speech (POS) tags throughout this section and the paper, let us first acquaint ourselves with some tags and their POS categories. We used Brill's Tagger to tag sentences. It follows the Penn Tree Bank POS Tagging Scheme. The POS tags and their categories that are important to this work are: *NN*: Noun, *NNP*: Proper Noun, *VBZ*: Verb, present tense, 3rd person singular, *JJ*: Adjective, *RB*: Adverb, *JJR*: adjective, comparative, *JJS*: adjective, superlative, *RBR*: Adverb, comparative, *RBS*: Adverb, superlative.

IV. COMPARATIVE SENTENCES

An *object* is an entity that can be a person, a product, an action, etc, under comparison in a comparative sentence. Each object has a set of features, which are used to compare objects. A comparison can be among two or more objects, groups of objects, one object and the rest of the objects. It can also be between an object and its previous or future versions.

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*.

- ✓ *NON-EQUAL GRADABLE:* Relations of the type *greater* or *less than* that express an ordering of some objects with regard to certain features. This type includes user preferences, and also those comparatives that do not use *JJR* and *RBR* words. *Ex: "optics of camera A is better than that of camera B"*
- ✓ *EQUATIVE:* Relations of the type *equal to* that state two objects as equal with respect to some features. Ex: "*camera A and camera B both come in 7MP*"
- ✓ *SUPERLATIVE:* Relations of the type *greater* or *less than all others* that rank one object over *all* others. Ex: "*camera A is the cheapest camera available in market*".

V. EXPERIMENTAL WORK

A. DATA DESCRIPTION

A comparative sentence usually expresses an ordering relation between two sets of entities with respect to some shared features (or aspects). For our study, we only used reviews of manufactured products. The Amazon product reviews are about mp3 players and were obtained from www.Amazon.com. Then we performed preprocessing on reviews.

B. PREPROCESSING

We pre-process the reviews as follows.

- ✓ All the words are transformed into lower case.
- ✓ Remove the numbers from the reviews.
- ✓ Remove the URL from the reviews.
- ✓ Remove the Punctuation from the reviews.
- ✓ Stop Word Dictionary: Stop word dictionary recognizes a stop words in the reviews.
- ✓ Remove Whitespaces from the reviews

C. PART-OF-SPEECH (POS) TAGGING

We now give an introduction to part-of-speech (POS) tagging as it is useful to our subsequent discussion and also the proposed techniques. In grammar, part-of-speech of a word is a linguistic category defined by its syntactic or morphological behavior. Common POS categories are: Noun, verb, adjective, adverb, pronoun, preposition, conjunction and interjection. Then there are many categories which arise from different forms of these categories.

In this work, we use Brill's Tagger (Brill 1992).Important POS tags to this work and their categories are:

JJR: Comparative Adjective,
JJS: Superlative Adjective,
RBR: Comparative Adverb,
RBS: Superlative Adverb.

Each word is then replaced with its POS tag. We do not use the actual words. For each keyword, we combine the actual keyword and the POS tag to form a single item. The reason for this is that some keywords have multiple POS tags depending upon their uses.

D. COMPARATIVE SENTENCES MINING TECHNIQUES

a. N-GRAMS CLASSIFICATION

The technique of document representation through term vector is the most common in the sentiment analysis field and can be used as our baseline. In this approach, each sentence in the corpus is a document, terms are the most relevant words and we use TF-IDF matrix to represent them.

b. SEQUENTIAL PATTERNS CLASSIFICATION

Sequential patterns classification for comparative sentences mining had been proposed. Sequential pattern mining (SPM) is an important data mining task. A sub-sequence is called sequential pattern or frequent sequence if it frequently appears in a sequence database, and its frequency is no less than auser-specified minimum support threshold minsup.

However a sentence cannot be handling simply from raw words, as we did on n-grams classification approach. To find sequential POS tags patterns in sentences and, then, build an

input dataset of sentences to be classified (supervised learning) as comparative or non-comparative.

VI. RESULTS

We collected data from disparate resources to represent different types of text. Our data consist of Consumer reviews on such products as *digital cameras*, *DVD players*, *MP3 players* and *cellular phones*. This data set is which studies impressions in reviews. The reviews were downloaded from Amazon.com.

Table 1 and Table 2 shows Training data and testing data respectively, which we labeled automatically using POS tagging. Because labeling the reviews is very time consuming job, the amount of data might not be very much currently

Data sets	Comp	Non-Comp	Total
Reviews	55	70	125

Table1: Training Data

Data sets	Comp	Non-Comp	Total
Reviews	5	25	30

Table 2: Testing Data

From the different machine learning algorithms; we used Naive-bayes Classifier to determine the sentiment of the reviews on Testing Data. Table 3 shows the result of Testing Data using Naive-Bayes Classifier.

	Comparative	Non- Comparative
Comparative	**1**	**3**
Non- Comparative	**2**	**24**

Table 3: Result of Naive-Bayes classifiers

We obtain 0.83 Accuracy after Naïve bayes Classifier.

VII. CONCLUSION AND FUTURE WORK

This paper proposed the study of distinguishing comparative sentences. Such sentences are useful in many applications, e.g., marketing intelligence, product benchmarking, and e-commerce. We first analyzed different types of comparative sentences from both the linguistic point of view and the practical usage point of view, and showed that existing linguistic studies have some restrictions. We proposed a POS (Part Of Speech) and machine learning approach to distinguishing comparative sentences.

This work primarily used POS tags. In our future work, we also plan to explore other language features (e.g., named entities, dependency relationships of different conceptions, etc) to improve the accuracy.

REFERENCES

[1] Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering. pp. 3–14. ICDE ’95 (1995)
[2] Arias, M., Arratia, A., Xuriguera, R.: Forecasting with twitter data. ACM Trans. Intell. Syst. Technol. 5(1), 8:1–8:24 (2014)
[3] Ceron, A., Curini, L., Iacus, S. M.: Using sentiment analysis to monitor electoral campaigns: Technique matters-evidence from the United States and Italy. Soc. Sci. Comput. Rev. 33(1), 3–20 (2015)
[4] Fournier-Viger, P.,Wu, C.W., Tseng, V.: Mining maximal sequential patterns without Candidate maintenance. In: Advanced Data Mining and Applications, vol. 8346,pp. 169–180 (2013)
[5] Jindal, N., Liu, B.: Distinguishing comparative sentences in text documents. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 244–251. SIGIR ’06 (2006)
[6] Liu, B.: Sentiment Analysis and Opinion Mining. Morgan Claypool Pub. (2012)
[7] McAuley, J., Leskovec, J.: Hidden factors and hidden topics: Understanding rating dimensions with review text. In: Proceedings of the 7th ACM Conference on Recommender Systems. pp. 165–172. RecSys ’13 (2013)
[8] Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trendsin Information Retrieval 2, 1–135 (2008).
[9] Sharma, A., Dey, S.: An artificial neural network based approach for sentiment analysis of opinionated text. In: Proc. of the 2012 ACM Research in Applied Computation Symposium. pp. 37–42 (2012).

IJIRAS