ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Extracting Comparative Sentences from Twitter Using POS tagging

Ashwini D. Pawar [1], Sachin N. Deshmukh[2]

M. Tech Student, Dept. of Computer Science and Information Technology, Dr. B. A. M. University Aurangabad, Maharashtra, India[1]

Professor, Dept. of Computer Science and Information Technology, Dr B. A. M. University, Aurangabad, Maharashtra, India[2]

ABSTRACT: The importance of Text Mining applications has increased in recent years because of the large number of web-based applications which lead to the creation of such data. Now a days, newer aspects of Text Mining can be apply on emerging platforms such as Social Networks. Opinion Mining and Sentiment Analysis are one of the applications of Text Mining. Opinion Mining refers to the extraction of lines and phases from the social networks that contain some opinion. Sentiment analysis identifies the polarity of opinion being extracted. Daily huge amount of data is generated by these Social networks such as Twitter. Users not only use these social networks but also give their valuable feedback, thus generating additional information. Due large amount of user's opinion, views, feedback and suggestion available through social networks in terms of comparative thoughts represent a way of users express their preferences about two or more entities, it's very much essential to explore, analyze and organize their views for better decision making.

KEYWORDS: Text Mining, opinion mining, comparative sentences.

I. INTRODUCTION

If we want to take a decision, we first prefer to seek others opinion, we evaluate opinions and take decision. Same thing is applied to organizations when they introduce new product or on the way to introduce it; organizations take opinions of its customers in the form of reviews of product on official websites of organization, social media sites such as Facebook, Twitter, Blogs or online shopping sites. Customer also wants to know opinions of existing users before they use service or purchase a product. These reviews help organizations and its Customers to evaluate the response or love among people about product or service.

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.

DOI: 10.15680/IJIRCCE.2016. 0408076

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.SENTIMENT ANALYSIS IN TWITTER

Sentiment analysis is a type of natural language processing for analyzing the mood of the public about a specific product or subject. Opinion Mining and Sentiment Analysis are the branches of Text Mining which refer to the process of extracting nontrivial patterns and interesting information from unstructured script documents. We can say that they are the expansion to data mining and knowledge discovery. Opinion Mining and Sentiment Analysis concentrate on polarity identification and emotion recognition correspondingly. Opinion Mining has higher attractive potential than data mining, as it is the most natural type of storing the information in text format. It is much complex task than data mining because it needs to manage unstructured and fuzzy information.

A. 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.*

1) *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"*

2) *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*"

3) *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

Twitter is a social networking service that lets its users to post real time messages, called tweets. Tweets have many unique characteristics. Twitter, with nearly 600 million users and over 250 million messages per day, has rapidly turned into a gold mine for organizations to monitor their reputation and brands by extracting and analyzing the sentiment of the Tweets posted by the public about their remarks, markets, and other contenders. Performing Sentiment Analysis on

Twitter is complicated than doing it for large reviews. This is because the tweets are very short and mostly contain slangs, emoticons, hash tags and other twitter language.

There is no large public available data set of Twitter tweets with sentiment, so we use Twitter API to collect data. The Twitter API has a parameter that specifies in which language you want to retrieve tweets and we set this parameter to English. We acquire tweets of *iPhone mobile* then we performed preprocessing on tweets

B.Preprocessing

In the tweets, people use acronyms; make the spelling mistake, use emoticons and the other characteristics that express the social meaning. Particularly the other characteristics includes the used of “@” and “#” symbols in tweets, where “@” symbol is act as the target and used to refer the other users on microblog while “#” symbol represents the content of the tweet.

We pre-process the tweets as follows.
All the words are transformed into lower case.Removes the numbers from the tweets.Removes the URL from the tweets.Removes the Punctuation from the tweets.Stop Word Dictionary: Stop word dictionary recognizes a stop words in the tweets.Removes Whitespaces from the tweets.

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 alsothe proposed techniques. In grammar, part-of-speech of aword is a linguistic category defined by its syntactic ormorphological behavior. Common POS categories are: Noun, verb, adjective, adverb, pronoun, preposition,conjunction and interjection. Then there are manycategories which arise from different forms of thesecategorieseach word is then replaced with its POS tag. We do not usethe actual words. For each keyword, we combine the actual keyword and the POS tag to form a single item. The reasonfor this is that some keywords have multiple POS tagsdepending upon their uses. Their specific usages can beimportant in deciding whether a sentence is a comparativesentence or not. For example, the keyword “*more*” can be acomparative adjective (*more/JJR*) or a comparative adverb(*more/RBR*) in a sentence.

D. Comparative Sentences Mining Techniques

1.N-grams Classification

The technique of document representation through term vector is the most commonin the sentiment analysis field and can be used as our baseline. In thisapproach, each sentence in the corpus is a document, terms are the most relevantwords and we use TF-IDF matrix to represent them. Such matrix is,therefore, submitted to a classifier that builds a model able to discover whethera given sentence is comparative or not.

2. 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 itfrequently appears in a sequence database, and its frequency is no less than auser-specified minimum support threshold minsup.
According to, a class sequential rule (CSR) is a rule with a sequentialpattern on the left and a class label on the right of the rule. Unlike classicsequential pattern mining, which is unsupervised, in thisapproach sequentialrules are mined with fixed classes. This technique is thus supervised.

VI. RESULT

We collected data from Twitter to represent differenttypes of text. Our data consist ofConsumer reviews on *iphone mobile*. Table 1 and Table 2 shows Training data and testing data respectively, which we labeled automatically using POS tagging.

Table1: Training Data

Data sets	**Comp**	**Non-Comp**	**Total**
Tweets	700	550	1250

Table 2: Testing Data

Data sets	**Comp**	**Non-Comp**	**Total**
Tweets	250	50	300

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

Table 3: Result of Naive-Bayes classifiers

	Comp	**Non-Comp**
Comp	**50**	**0**
Non- Comp	**250**	**0**

We obtain 0.83 Accuracy after Naïve bayes Classifier.

VII. CONCLUSION AND FUTURE WORK

In our research, we have presented a method that automatically collects tweets from twitter using Twitter API. Then on these collected tweets, we perform preprocessing for our desired results and after that we performed POS tagging by using Tree-Tagger. We can use some of tweets to train a sentiment classifier. Our classifier is able to determine comparative, and non-comparative from tweets. The classifier is based on Naive Bayes classifier that uses Unigram and POS tags as features. We can observe a poor performance for n-grams approachvarying classifiers algorithms does not impact on results that reach a maximum accuracy of 68.6% for RBF neural network. The neural network RBF-CSR reached the best accuracy of 81.13%. But by using Naïve bayes classifier and POS tags, our system got 83% accuracy.

This work primarily used POS tags. Infuture work, we also plan to explore other languagefeatures (e.g., named entities, dependency relationships ofdifferent 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, and R.: Forecasting with twitter data. ACM Trans. Intell. Syst. Technol. 5(1), 8:1–8:24 (2014)

3) Ceron, A., Curini, L., Iacus, and 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 Researchand 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) Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., Hsu,M.C.: Mining sequential patterns by pattern-growth: The PrefixSpanapproach.IEEE Trans. on Knowl.And Data Eng. 16(11), 1424–1440 (Nov 2004)
10) Sharma, A., Dey, S.: An artificial neural network based approach for sentiment analysis ofopinionated text. In: Proc. of the 2012 ACM Research in Applied Computation Symposium. pp. 37–42 (2012)

BIOGRAPHY

Ashwini Deeliprao Pawar M.Tech Student Received B.E in Computer Science and Engineering from Godavari College of Engineering Jalgaon, North Maharashtra University (NMU), Jalgaon in 2012. Currently pursuing M.Tech in Computer Science and Engineering from Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India.

Dr. Sachin N. Deshmukh Professor completed his Ph.D. and M.E in Computer Science and Engineering. He is currently a Professor in Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India. His research Area is Text mining, Social Web mining and Intension Mining. He is a member of Adhoc Board of Studies in Bio Informatics and Liberal arts of Dr. B. A. M. University Aurangabad and Adhoc Board of Computer Science at Shivaji University Kolhapur.