iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** VIII **Month of publication:** August 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue VIII, August 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Sentiment Analysis of Web Scraped Product Reviews using Hadoop

Mahek Merchant[1], Ricky Parmar[2], Nishil Shah[3], P.Boominathan[4]
[1,3,4]SCOPE, VIT University, Vellore, Tamilnadu, India

Abstract*— As in the world there is large volume of data, individuals are proposed to develop a framework that can distinguish and classify data based on opinions. Sentiment analysis (opinion mining) can be considered as the one of the major application of Natural Language Processing (NLP). We are looking forward to sentiment product reviews using sentence level categorization in Hadoop. We have used web scraping method to retrieve product reviews from different websites which is then used as dataset in this experiment. Finally, we compare the sentiment reviews between both the websites, so that decision making will be lot easier for the user regarding purchasing of the product.

Keywords*— Sentiment Analysis, Machine Learning, Web Scraping, Product Reviews, Natural Language Processing

I. INTRODUCTION

To track the mood of a people regarding particular product or topic, we used sentiment analysis which is a part of natural language processing. Sentiment analysis (Opinion mining) includes a framework which collects and look for opinions about the product made in blogs, reviews or tweets. There are many applications in which opinion mining can be used. For instance, in advertising it helps in judging the achievement of an advertisement battle or new product dispatch, figure out which adaptations of a product or administration are well known and even distinguish which demographics like or dislike specific elements.

From a scientist's point of view, several social media websites provide their application programming interfaces (APIs), provoking information gathering and investigation by scientists and engineers. For example, Twitter as of now has three distinct variants of APIs accessible [9], in particular the REST API, the Search API, and the Streaming API. With the REST API, engineers can assemble status information and client data; the Search API permits engineers to question particular Twitter content, while the Streaming API can gather Twitter content in real-time. Besides, engineers can blend those APIs to make their own applications.

Nevertheless, those sorts of online information have a few defects that conceivably thwart the procedure of opinion mining. The principal defect is that since individuals can unreservedly post their own content, the nature of their sentiments can't be ensured. For instance, rather than imparting theme related insights, spams are posted by online spammers on collection. Some spam are good for nothing by any means, while others have insignificant opinions otherwise called fake sentiments [7-9s]. Another defect is that ground truth of such information provided by online is generally inaccessible. A ground truth is like a label, showing whether the opinion of sentiment is positive, negative, or neutral.

Web scraping is used to collect online product reviews from different websites which are used as dataset in this paper. There are two faults which can be tackled in following ways: First, before posting every product gets assessments. Second, rating is specified on each product which can be used as the prior truth. There is a star-scaled framework for rating, where 5 stars indicates most noteworthy and 1 star indicates least worthy.

The following sections of this paper are as follows. In section 2, is about literature review. In section 3, we discuss regarding web scraping of data. In section 4, we discuss our technique and methods. In section 5, we show our experimental results and discussion. Finally, we give conclusion in section 6.

II. RELATED WORKS

Fang [1] concentrates on two critical assignments in sentiment analysis, i.e., supposition dictionary extension and target extraction. They propose a proliferation way to deal with concentrate keywords and targets iteratively given just a seed supposition dictionary of little size. The extraction is performed utilizing recognized relations between keywords and targets, furthermore opinion words/targets them. The relations are depicted grammatically taking into account the relevant language structure. The author additionally proposes novel techniques for new conclusion of words for target pruning.

Kim [2] present an unsupervised framework for removing perspectives and deciding opinion in review content. The strategy is basic and adaptable with respect to area and dialect, and considers the impact of perspective on opinion mining. They present a local

Volume 4 Issue VIII, August 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

theme model, which looks at the sentence and keywords of subjects that naturally surmise the angles.

A way to deal with keywords and to distinguish the sentiments connected with these components from reviews through syntactic data in view of relevant examination is depicted in [3]. In [4] Tao and Yi is introduced a novel way to deal with gain from lexical information as space autonomous assumption Laden terms, in conjunction with some labeled records. This model depends on an obliged non-negative tri factorization of the term-record framework which can be executed by utilizing basic upgrade rules.

The procedure of evaluating the support for the review remarks by surveying the qualities, add quality to the review inspection process [5]. The review remarks could originate from chat rooms or online examination discussions. In numerous situations, it is prudent to utilize a computerized customer survey operators for gathering and making review models [6]. Different analysts use distinctive machine-learning procedures for performing classification.

III. WEB SCRAPING

With web scrapping you can apply approximate tree pattern matching to web scrap the data. The design of a site, i.e. the presentation of information, is portrayed using Hypertext Markup Language (HTML). A HTML report essentially comprises of four kind of components: Archive structure, piece, inline and intuitive components.

There is a Document Type Definition (DTD) for every variant of HTML which portrays how the components are permitted to be nested. It is organized as a punctuation in amplified Backus Naur structure. There is a script and a transitional form of the DTD for in reverse compatibility. The most well-known conceptual model for HTML report are trees. An illustration of a HTML archive demonstrated as a tree which is appeared in a figure below:

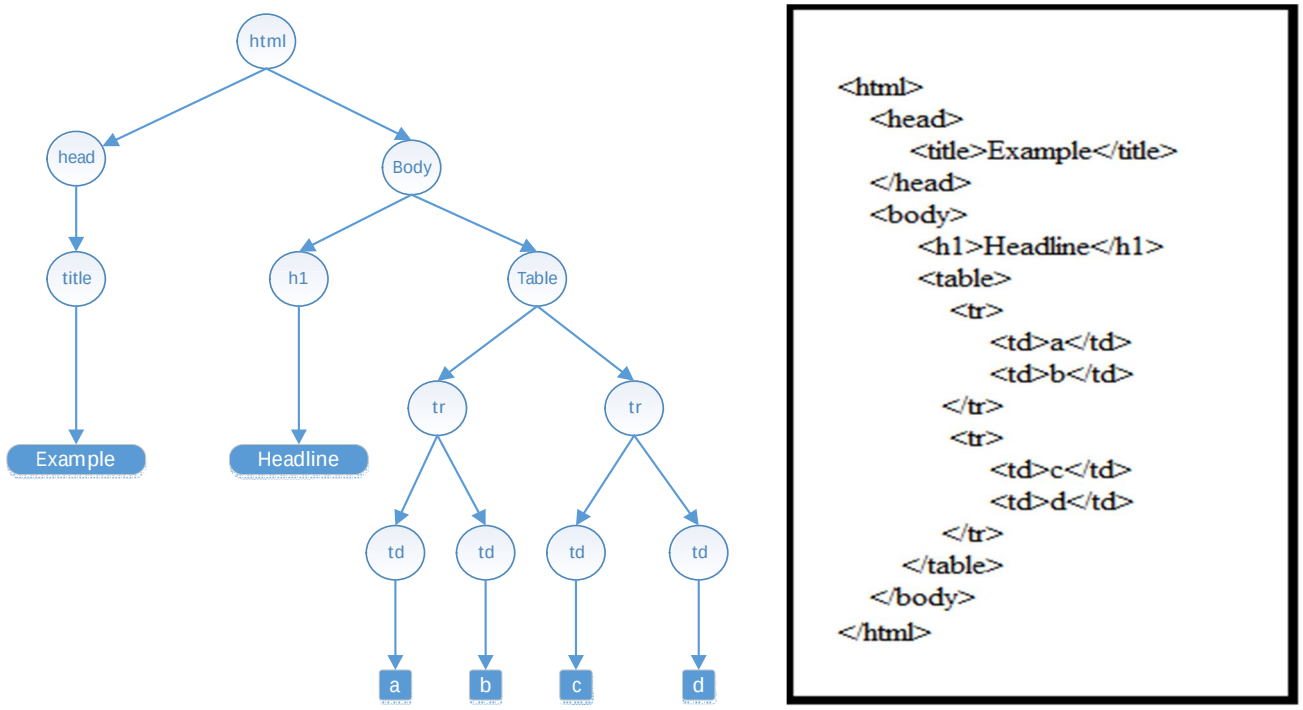


Fig.1 Tree model for HTML document

Change in the HTML archive influences, so a tree can be utilized to distinguish auxiliary changes. Tree design coordinating can be utilized to discover matching of an example in the HTML tree. For this reason, the above tree model is a subset of an HTML report.

IV. METHODOLOGY

The motivation behind this examination is to scrap, arrange and classify the data contained in the dataset. Fig 2 demonstrates our architecture, which consists of various practical parts.

A. Scraping of Reviews

This section includes online product reviews which are extracted from different websites using web scraping. For experimental

Volume 4 Issue VIII, August 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

purpose, we scraped mobile reviews of different giants like Apple, Samsung, OnePlus etc.

Reviews can be:

1) Not satisfied with the camera and performance.
2) Light weight, great looks and build quality. Average camera and image software.
3) I am not satisfied with this camera.....I have a hard time getting clear shots that aren't hazy. Not inspired by any means.

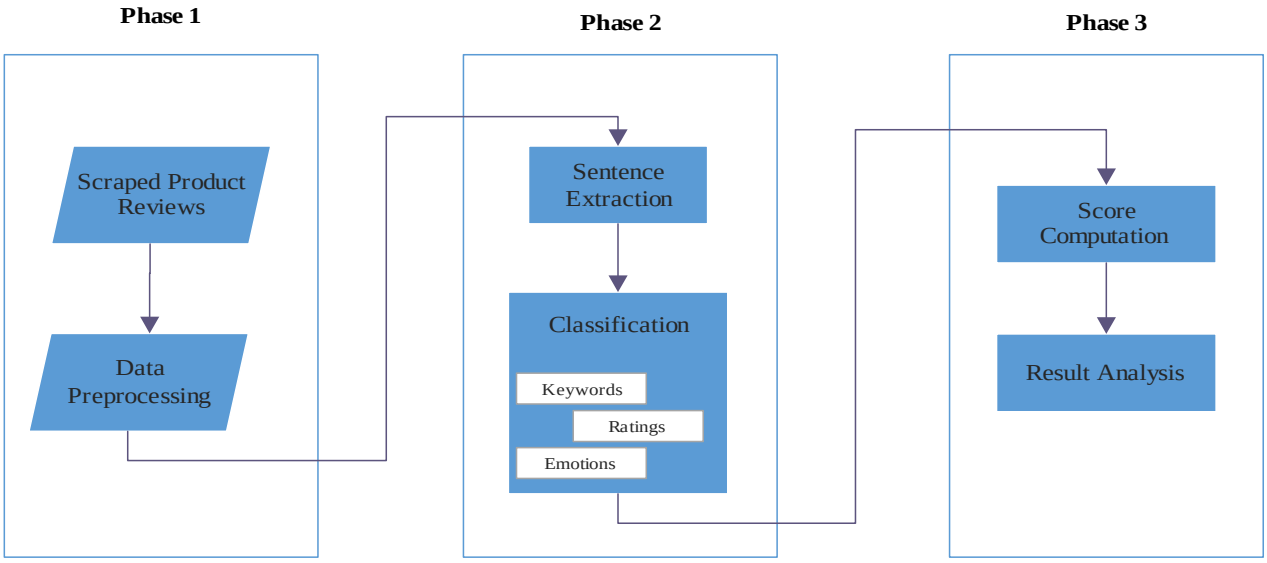


Fig2. Proposed Sentiment Classification

B. Data Preprocessing

Here, scraped product reviews are preprocessed (filtered). Filtering is done step by step:

1) Remove undesirable words (like are, the, was, it). i.e. Stop Words Removal
2) A Part-of-Speech Tagger required in order to identify noun, pronoun, adjective, adverbs etc.
3) Stemming words such as heat-heated, fast-faster should be removed i.e. Stemming Process
4) Keywords extraction.

C. Classification based on Keywords

It utilizes 'pack of words' methodology. Words are space free. Every word has been named positive/negative. We need to give words in right spelling to be arranged. Weight of each word is same. There might be a mix of positive/negative words in a dataset which may result into wrong classification.

Table 1. Positive and Negative keywords.

Positive	Negative
Satisfied	Overheating
Impressed	Bad
Fast	Worst
Smooth	Hated
Excellent	Poor

D. Classification based on Ratings

Ratings are frequently utilized for classification purposes. They are utilized by analysts for positioning things, for example, movies, TV appears, eateries, and hotels. These star evaluations describes summation of ratings which are assembled from various sources, including dealers, customers, article locales and clients.

Volume 4 Issue VIII, August 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Table 2. Sample Ratings

Ratings	Meaning
★★★★★	Excellent
★★★★	Good
★★★	Average
★★	Poor
★	Bad

E. Classification based on Emotions

This characterization is done on the premise of emoticons. It utilizes R.E (Regular Expression) to identify nearness of emoticons which are then arranged into positive or negative, utilizing a rich arrangement of emoticons which are labeled as positive or negative. It utilizes a collection of positive and negative feelings which are really two text that incorporate positive and negative emotions symbols separately.

Table 3. Emotions

Positive	Negative
:)	:(
:-)	:-(
=)	=(
: o	:'(

F. Review Score

On the basis of extracted keywords, reviews are scored. Each product has its own reviews and ratings. For experimental purpose, we have used the following formula for Score computation:

$$R_p + \sum_{i=1}^{n} \mathbf{Ki}$$

where R_p is the net rating and K_i is the weight for each keywords in the dataset

V. EXPERIMENTAL RESULTS

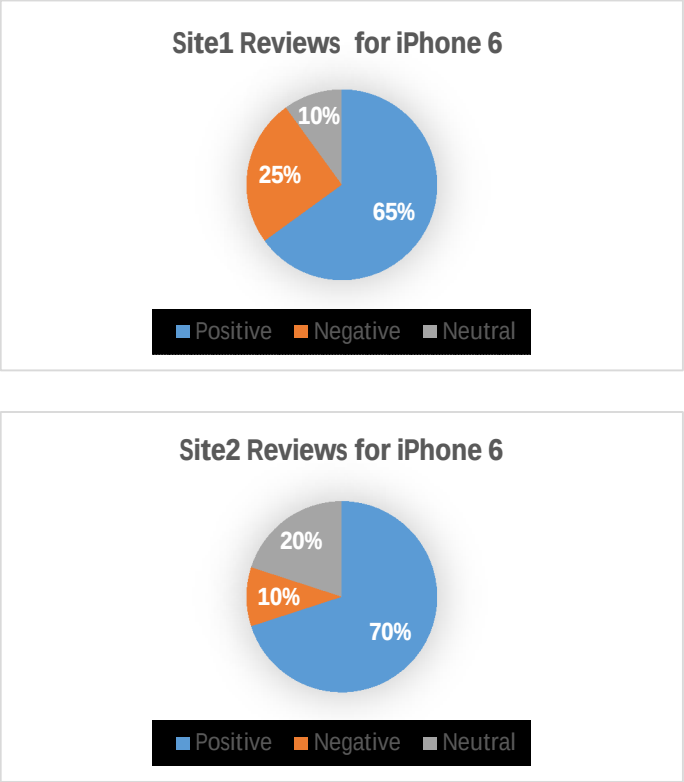
In this section, we show score computation of iPhone 6 for two websites. Experiments were performed using JAVA of version JDK 1.6 in Hadoop Distributed File system(Hadoop 1.21.), which was implemented in Ubuntu 15.04 with an Intel® Core™ i5-2430M Processor (@3.00 GHz) and 4 GB RAM.

Product reviews were extracted using web scraping method which was implemented in Python 2.7 using BeautifulSoup library.

Graph1. Score comparison for iPhone 6

From graph1, Score of Site2 is more compared to Site1,means Site2 has more positive feedback

Volume 4 Issue VIII, August 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)



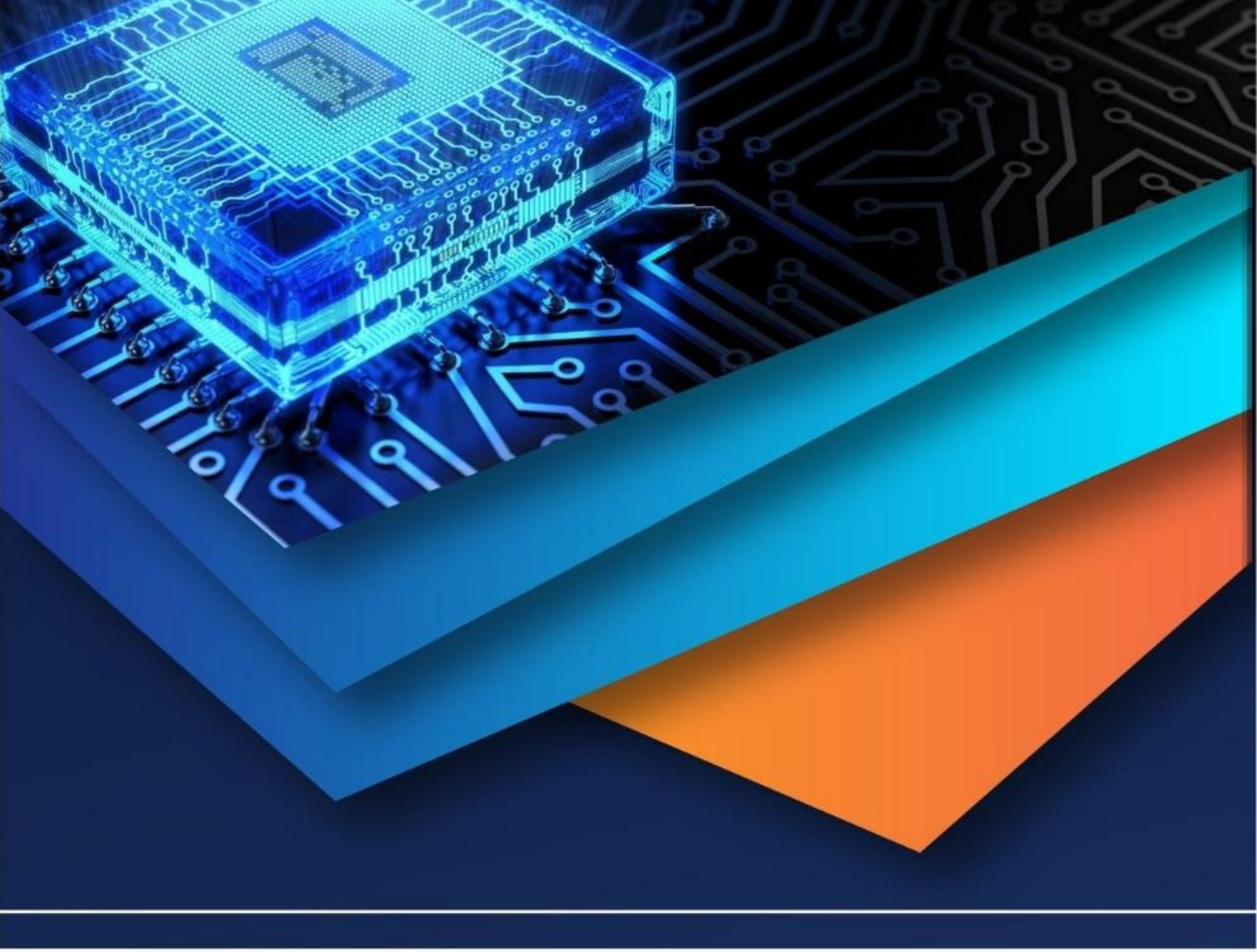
Graph2. Opinions from Site1 and Site2

VI. CONCLUSION

In this experiment, we have proposed a system which makes classification based on keywords, ratings and emotions. In this paper, we compare product reviews from different websites, so that decision making will be a lot easier for the user regarding purchasing of the product. For future work, we can implement this project on Google's TensorFlow, which is an open source framework for machine learning.

REFERENCES

[1] Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, 2(1). Doi:10.1186/s40537- 015-0015-2
[2] Kim S-M, Hovy E (2004) Determining the sentiment of opinions. In: Proceedings of the 20th international conference on Computational Linguistics, page 1367. Association for Computational Linguistics, Stroudsburg, PA, USA
[3] Liu B (2010) Sentiment analysis and subjectivity. In: Handbook of Natural Language\ Processing, Second Edition. Taylor And Francis Group, Boca
[4] Liu B, Hu M, Cheng J (2005) Opinion observer: Analyzing and comparing opinions on the web. In: Proceedings of the 14th International Conference on World Wide Web, WWW '05. ACM, New York, NY, USA. pp 342–351
[5] Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation. European Languages Resources Association, Valletta, Malta
[6] Pang B, Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL '04. Association for Computational Linguistics, Stroudsburg, PA, USA
[7] Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(12):1–135
[8] Turney PD (2002) Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02. Association for Computational Linguistics, Stroudsburg, PA, USA. pp 417–424
[9] Whitelaw C, Garg N, Argamon S (2005) Using appraisal groups for sentiment analysis. In: Proceedings of the 14th ACM International Conference on Information and Knowledge



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY