Sentiment Analysis on Citations in Research Paper

[1]P. Suryaja, [2]N. Sai Pranav,[3]B. Rohith,[4]K. Chandra Sekhar,[5]G. V. Gayathri

[1,2,3,4]Student, [5]Asst. Professor, Dept. of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh, India.

[1]suryaja.parvatini@gmail.com, [2]pranav605@gmail.com,

[3]bankarohith1999@gmail.com,[4]chandunayak1314@gmail.com,[5]gayathri.ganivada@gmail.com

***Abstract*: Scientific papers are the papers including citations. Generally scientific papers include references at the end of the paper. Authors of these papers may include related work in their paper. The related work may or may not help the author. We apply sentiment analysis on the related work in a scientific paper and find out whether the references are helpful or not for the author. In this paper, we apply data pre-processing on the related work and find out the polarity for each mentioned reference using Naive-Bayes theorem. We also apply h-index ranking algorithm on the citations of references. Finally, we find out the overall polarity of the references. Three parameters are used to calculate the result. They are related work mentioned in the research paper, number of citations for each reference in the research paper and h-index value for each author in the references of the research paper respectively. First parameter is used to calculate naïve bayes and the rest two are used to calculate h-index. The expected result is to find the polarity of each reference cited by the author.**

***Keywords*: Citations, Naïve-Bayes, Polarity, Sentiment Analysis, Scientific Papers.**

I. INTRODUCTION

In the field of computer science many new research papers are being published every single day. Whenever a researcher starts to develop a project he goes through several research papers related to the field in which they are developing. In the process a researcher needs to go through the whole paper and then analyze the paper to know if the references mentioned by a particular author affected the paper in a fruitful way or not. This will become a tedious task when many papers are to be referred , the algorithm proposed in this paper does that tedious work of finding how a particular reference influenced a paper.

This is achieved with the help of Sentimental Analysis, It is the process of identifying the sentiment of a particular set of text. A sentiment is the tone or the mood of the author[10]. A sentiment is often represented in subtle or complex ways in a text[8]. The data is collected from a research paper and then data pre processing is performed. Basically it is data mining which means to collect useful information from a huge collection of data. Here with the help of data mining and sentimental analysis the polarity is predicted for a particular reference mentioned by the author in a paper.

II. PROPOSED ALGORITHM

There are three modules in our system.

2.1. Citation Text Identification

2.2. Text Pre-processing

2.3. Naïve Bayes

2.4 H-Index

2.1 Citation text Identification

Citation means a reference to or quotation from a paper, author, or book, particularly in a scholarly work. Purpose of citation is to allow users to find and verify sources for papers, show type and degree of support, give attribution, show paper is well researched. In this system, Citation context is defined as the textual statement that contains the citation[5]. we first collect the data i.e., the scientific papers and from the respective paper we take the related work which is described by the author. From the related work we first identify the cited text and after that we implement the algorithms on the resulting data.

2.2 Text Pre-processing

Text pre-processing is one of the module where the text is converted into array of words. After citation text identification, the related data is given as an input to this module. The input is the related work mentioned by the

 DOI : 10.35291/2454-9150.2020.0320

author of the scientific paper. This related data is in the form of text. Once text is obtained, it needs to undergo text normalization. In text normalization, there are several steps. All the characters of the text are to be converted into either upper or lower case. After this, numbers are removed. Also punctuations, accent marks, white spaces, stop words, sparse words are removed.

Tokenization:

This is the process where the text is split into tokens I.e., small pieces. Tokens such as words, punctuation marks, numbers etc.

Stop Words:

The stop words are removed from the set of words. These words does not have any specific meaning. Some of them are 'the', 'a', 'on' etc

Stemming:

This is a process of reducing words to their root or base form.

Pots – Pot

Was – wa

Studies – Studi

Studying - Study

Lemmatization:

This is the process similar to stemming, to reduce inflectional form to base form. After this, we obtain an array of words.

Geese – Goose

Was – Be

Studies – Study

Studying - Study

2.3 Naïve Bayes

Formula to calculate Naïve Bayes[8]

$$P(X \mid Y) = P(X)\,P(Y \mid X) / P(Y)$$

Where:

$P(X)$ = *Probability of X's occurrence.*

$P(Y)$ = *Probability of Y's occurrence.*

$P(X / Y)$ = *Probability of X when given Y .*

$P(Y / X)$ = *Probability of Y when given X .*

Naïve Bayes theorem is basically a classification technique which is used to classify the give set. Based on training data Naïve Bayes theorem is applied to the trail data and is classified.

Here, Naïve Bayes is applied to classify the related work mentioned by the author into three different classes namely, positive polarity words[1], negative polarity words[1] and neutral based on the polarity of the words of the training data. After text classification, the resultant data is given to the Naïve Bayes Theorem. Based on training data, the polarity for each reference is calculated accordingly.

2.4 H-Index

H-Index is one of the algorithms using which we can calculate the impact generated by a researcher[2]. H-index is purely based on citation count. The H-index is defined as follows : "A scientist has index h if g of his or her Np papers have at least h citations each and the other (Np – h) papers have <= h citations each." (Hirsch, 2005) Mathematically, it can be represented as formula:

$$h(f) = max/min(fi,i)$$

where f is the function that corresponds to the number of citations for each publication, sorted in descending order[2] . H-index finds a balance point between publication amount and the citation count of each publication. It is however not able to model the different polarities of citations.

Algorithm:

1 . Start.

2 . Take all the referred papers in the related work of the base paper .

3 . Start a for loop until all papers are finished:

for(i = papers ; i > 0 ; i --)

4 . Take the number of citations for each paper in h[i] , where I is the number referring to the paper.

5 . Now for each paper, find the h-index value for author in a[i] array. (from google scholar)

6 . Now apply h value on the array such that :

a . If h[i] > a[i], then positive.

b . Else if h[i] < a[i], then negative.

c . Else if h[i] = a[i], then neutral.

7 . End.

III. EXISTING SYSTEM

The main goal of the system is to find the polarity of a research paper. In the existing system, the polarity of the whole paper is calculated i.e., the effect of the whole paper. Whether the paper is relatively helpful to the student or not and the effect of the whole paper. Here, the result is whether the paper has positive polarity or negative polarity. There are various methods to find the polarity of the paper.

 DOI : 10.35291/2454-9150.2020.0320

IV. PROPOSED SYSTEM

The proposed system is an extension of the existing system. In this system, a research paper is taken as an input. From the paper, the related work mentioned by the author in the paper is extracted. The related work consists the methods or data used in each reference paper. Author mentions whether any data is useful for them or not. Here, the system finds out the polarity for each paper mentioned in the related work. The result is such that whether the author finds the reference positive or negative. The method used here is Naïve Bayes and H-index. The result is author's point of view. Using the result produced by both the algorithms we determine the way in which a author has affected the paper .

V. SYSTEM ARCHITECTURE

System Architecture means "the overall structure of the system and the ways in which the structure provides conceptual integrity". Architecture is the hierarchical structure of a program components (modules), the process in which these components interact and the structure of data that are used by that components.

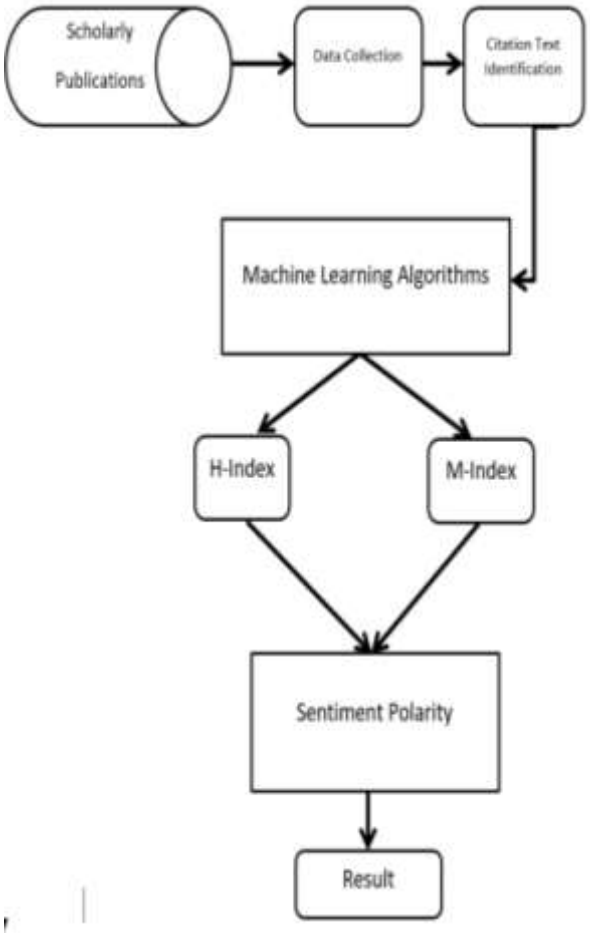

Fig : Architecture

Scholarly publications include several papers written by scholars. The research conducted by the scholar is published in a paper and are submitted in journal. The data base contains several such papers. The papers include research conducted in different departments

In the stage of data collection we pick a paper from several papers present in the scholarly publications database and perform data pre-processing on that paper.

In citation text identification we identify the text related to each cited paper present in the reference section.

After the text is identified we apply the machine learning algorithms and evaluate the final sentiment polarity of every reference cited by the author.

VI. CONCLUSION AND FUTURE WORK

In this paper, the techniques used are text classification and Naïve Bayes theorem. This paper results the point of view of author for each reference they used in their paper. The main objective is where the references are useful to the author or not. To calculate the polarity, the technique is based on sentimental analysis. Sentiment for each word is taken into consideration and is fed as the training data. When the polarity for the related data is to be calculated, it is given as training data and the polarity is calculated. The polarity is chosen based on the highest probability between the three classes (positive, negative, neutral).When once h-index and naïve bayes are applied on the respective data, the resultant data is compared. Comparative analysis is applied on both h-index and naïve bayes results. According to the findings h-index result is not accurate compared to naïve bayes. H-index result is based on the number of citations which may not result in the accurate result for finding the polarity of the paper or related work. Naïve bayes gives the accurate result for the related work. So when both h-index and naïve bayes are compared, naïve bayes is more accurate. For future work, many new algorithms and techniques can be nourished to get more accurate results when given with more training data. In future semantic analysis can be applied on the data where the feeling of the each word, sequence of words can be given as training data. The training data can be given in such a way that as if the feelings and understanding of humans. Furthermore, not only based on previous data but it can also be based on the current data. These techniques can be applied wherever there is a necessity to calculate the polarity.

REFERENCES

[1] Souvick Ghosh, Dipankar Das and Tanmoy Chakraborty. Determining sentiment in citation text and analysing its impact in the proposed ranking index.

[2] Zheng Ma, Jinseok Nam and Karsten Weihe. Improve Sentiment analysis of citations with author modelling.

[3] J. E. Hirsch. An index to quantify an individual's scientific research output. In Proceedings of the National

 DOI : 10.35291/2454-9150.2020.0320

Academy of Sciences, 102(46):16569–16572, November (2005).

[4] Amjad Abu-jbara, Jefferson Ezra. 2013 and Dragomir Radev. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of NAACL*-2006..

[5] Citation Sentiment Analysis in Clinical Trial Papers Jun Xu, Ph.D., Yaoyun Zhang, Ph.D., Yonghui Wu, Ph.D., Jingqi Wang, M.S., Xiao Dong, M.D., Hua Xu, Ph.D. School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA.

[6] Sentiment Analysis of Citations Using Word2vec Haixia Liu School Of Computer Science, University of Nottingham Malaysia Campus, Jalan Broga, 43500 Semenyih, Selangor Darul Ehsan.

[7] Effective Sentimental Analysis and Opinion Mining of Web Reviews Using Rule Based Classifiers Shoiab Ahmed and Ajit Danti

[8] Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. Christos Troussas, Maria Virvou Department of Informatics, University of Piraeus, Piraaeus, Greece. Kurt Junshean Espinosa, Kevin Llaguno, Jaine Caro, Department of computer science, university of Philippines, cebu city, Philippines.

[9] Sentiment Analysis of Political Tweets : Towards an Accurate Classifier Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, RonO'Brien, LamiaTounsi and MarkHughes.

[10] Sentiment Analysis of Movie Reviews A new Feature-based Heuristic for Aspect-level Sentiment Classification. V.K. Singh, R. Piryani, A. Uddin Department of Computer Science ,South Asian University ,New Delhi, India, P. Waila DST Centre for Interdisciplinary Mathematical Sciences Banaras Hindu University Varanasi, India

 DOI : 10.35291/2454-9150.2020.0320