Maximizing Influential Spread In Social Network

Implementation using Cascade varying Activation Probability

[1]Prabhat Mishra, [2]Kaushal Kishore, [3]Shivam Gupta, [4]Monika Poddar, [5]Asha G R

[1,2,3,4]Students(B.E), [5]Asst. Professor
Computer science and Engineering,
BMS College of Engineering, Bangalore, India

***Abstract*— The amount of data being generated daily is enormous and its analysis and management is becoming difficult and tedious. Based on the ongoing research and development in sentimental analysis and other analysis it has been proved to be difficult lately. Our project using different combinations of techniques like linguistic, NLP, popular visualization techniques we generate graphs which can be understood easily. The whole analysis is done using different algorithms and methods like detailed sentimental analysis of text; opinion mining which mainly focuses on polarity detection (negative, positive, neutral).**

***Index Terms*— Polarity, Sentimental Analysis, NLP, FEM.**

I. INTRODUCTION

World Wide Web can be viewed as a storage area of opinions from users spread across various websites and networks and today's netizens look up Tweets and opinions to judge commodities, visit forums to debate about events and policies. With this extreme volume of data and reliance on user Tweets and opinions, manufacturers and retailers have to face the challenge of automating the analysis of such big volume of data (user Tweets, opinions, sentiments). Equipped with these results, sellers can improve their product and tailor experience for the customer. Similarly, policy makers can review these posts to get quick and comprehensive feedback. Or use it for new ideas that democratize the policy making process. Our paper is the outcome of our research in gathering opinion and Tweets data from popular portals, e-commerce websites, various forums or social networking sites and processing the data using different rules of natural language and grammar so that to find out what was exactly being talked about in the user's Tweets and the sentiments that people are expressing. Our method diligently scans line by line of data, and generates a cogent summary of every Tweet (categorized by aspects) along with graphical visualizations. A well analyzed application of this approach will help out product manufacturers or the government in gauging response.

II. RELATED WORK

Recently summarizing the sentiments by extracting and aggregating sentiment over ratable aspect has become popular area in research. Previously, the user adopted multi-stage methodology in which cutting edge techniques are combined into multi document summarization, polarity classification and sentiment analysis .IT automatically synthesizes, filters, and summarizes user product Tweets into a short list of positive and negative opinions as expressed in the Tweets.

The disadvantage of previous approach is that it just finds Tweets of product manually by collecting the data in the form of a excel file and will produce and classify the Tweets as positive and negative but does not consider any attributes to get the best product.

The activation Probability used in [1] is used to create cascade in the social network which is then used to propagate the information through this cascade. The cascade formation is based on similarities in this probability.

In [2] a data based perspective was proposed by Amit Goyal et,al which completely diverts the traditional method of influential maximization in which the probability between edges is simulated using Monte-Carlo.In this probabilities are learned which are not based on assumptions but using the real world data and credit based model these are learned and hence are less to error prone and accurate probabilities are computed and hence the need for Monte-Carlo simulations are eliminated.

III. IMPLEMENTATION

The implementation of the this problem is done using java based framework which retrieves all the valid tweets on the basis of valid hash tag related to product submitted. All the tweets retrieved are then analyzed. This Analysis done using the methods described in [3] where Xiaowen Ding et,al have used the concept of opinion mining where each sentence are analyzed. In this it has given methods to know if a product is talked about in the context or not. If the sentence has the product names, they need to be identified. If the product name is not mentioned then we infer about the opinion of person about the similar product used. We call this problem entity assignment. Each tweet corresponding to user identifies the sentiment of the user which is used to calculate the polarity of the user with respect to product. The same approach has been used in [4]-[5].This polarity can be positive negative or neutral. Polarity calculated is the used to calculate the Activation probability for each user.

The product we have taken for our implementation is mobile. In [6] Wei Jin et,al have introduced the concept of "Opinion Miner" which we have used in our implementation. This uses a Machine Learning approach to automate the entire opinion into positive

or negative All the tweets related to a mobile practically will consist of hash tag for mobile and sentiment related to common features like Battery, Screen, Touch, Memory, Camera, and Sound. Analyzing a tweet related to a product will give view of a particular user associated with the Tweet that which feature attracts him most in the product. Considering such feature vector will also help in calculating the polarity of user with respect to each feature of the mobile.

The Methodology of the algorithm can be described as follows

1. Collection of Tweets for the given Product using Twitter
2. Break the Tweets into a sequence of statements.
3. Feature Extraction Matrix is generated and product recommendations are generated based on the user searched feature.
4. Compute the Polarity Per product per Feature by using Battery, Memory , Camera, Sound, Touch

The FEM (feature extraction matrix) will help further in counting the number of positive and negative Tweets for each feature
The FEM (feature extraction matrix) will help further in counting the number of positive and negative Tweets for each feature

Input and Output

Tweets Collection

Offline Phase

Input- Tweets Description, Product Name
Output – Tweets Stored at the Data source

Online Phase

Input – URL of Tweets, Product Name, Site-Amazon or Flip kart
Output – Collected Tweets from Amazon or Flip kart and stored in the format of Tweets in the data store

Data Cleaning

Input – Collected Tweets and Stopwords
Output – Cleaned Tweets which does not contain any not meaningful words.

Frequency Computation

Input – Cleaned Tweets
Output – Tokens Formation and Frequency Computation per Tweets.

Feature based Frequency Computation

Input- List of Tokens, Tweets IDs, Product IDs
Output – Computation of Frequency across all products and all Tweets per feature

POS based Frequency Computation

Input – Tweets
Output- Perform POS tagging and compute positive frequency

Feature Extraction Matrix (FEM) generation algorithm

Input – POS frequency, Text Frequency and Product IDs
Output – FEM matrix which has one row per product and one column per feature

Ranking Algorithm

Input – FEM matrix and Search Query
Output – List of Products based on highest value of FEM for searched features

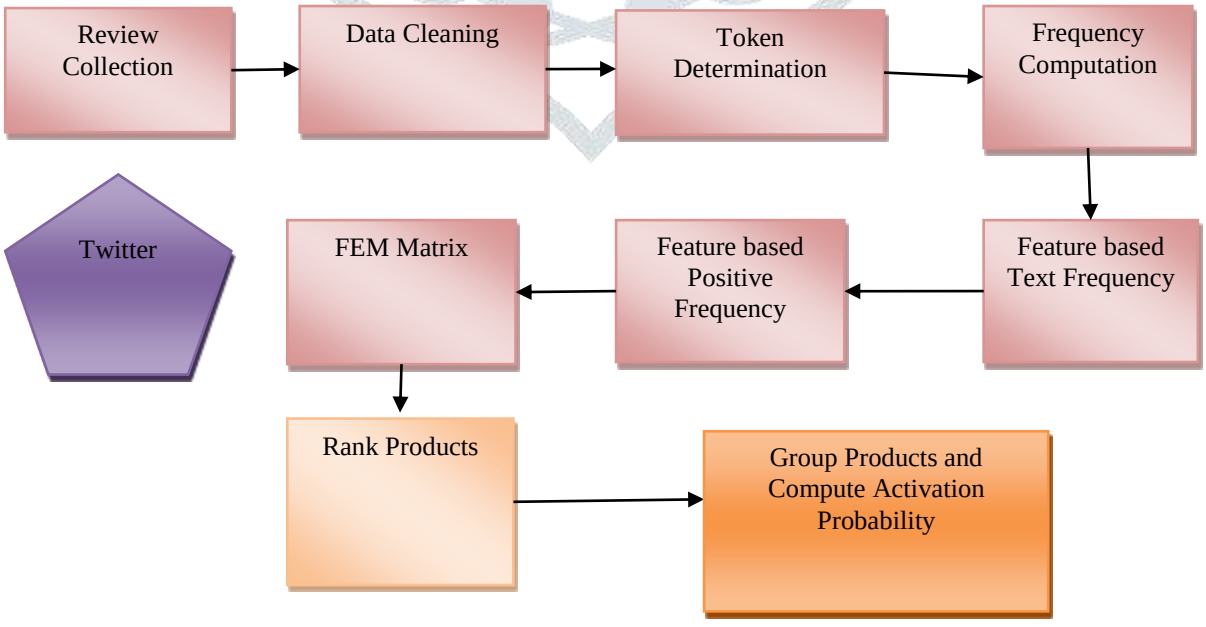

Fig. 1 : Data flow of Implementation using CVAP algorithm

These are three major steps:

Data Collection using Twitter

The data collection in twitter is done using Twitter API's for which you need to have twitter account using your credentials like API key, Access token, API secret you can access the Twitter API. Using this we can collect data related to particular hashtag using REST API. The data returned by this method is in JSON format which is then parsed according to the following flowchart.

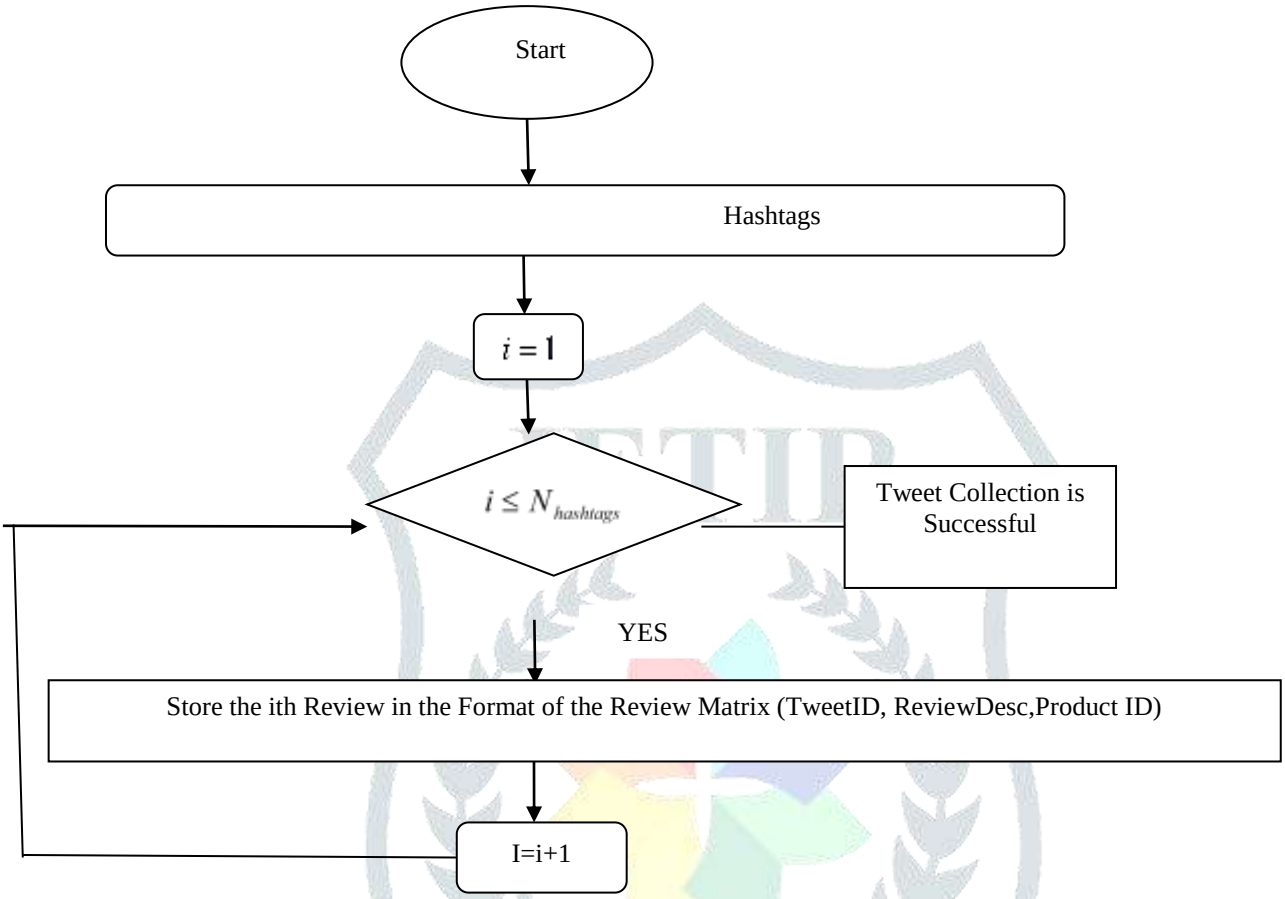

Fig 2*:* Data Collection using Web Crawler

The **Figure 2** shows the Data collection using Tweets ID, Tweets Desc and Product ID acts as an input. All the tweets for the specific Hash Tag are collected and then they are stored in the format of JSON. After collection of data the data needs to be cleaned where stopwords are removed out of them and other traditional ways of data cleaning.

FEM Algorithm

This Algorithm computes the FEM(frequency extraction Matrix) which is nothing frequency of each feature per Product per tweet id. This frequency is evaluated by analyzing each tweet and the feature that tweet describes for each tweet if it contains the feature then the frequency corresponding to that feature is increased. For example a tweet describing a battery feature will have it's frequency increased for a tweet if the tweet contains both the product name and the feature name, this value is then stored in the database as a row column manner which is also called feature extraction matrix.

The algorithm computes feature based nature for tweets which is the used for the polarity calculation in the next step of algorithm. The feature based extraction is an important step as suppose a person does not want to evaluate product on the basis of certain feature the he can skip polarity computation of that feature by just taking only those tweets that describes his required feature thus saving a great ton of computations

The other great Advantage of calculating this feature extraction matrix is that it gives which feature is most talked about in a tweet which is a pre-picture of the importance of feature and also builds the base for polarity calculation. The below flowchart shows how the algorithm proceeds as follow. The **Figure 3** shows the flow of algorithm.

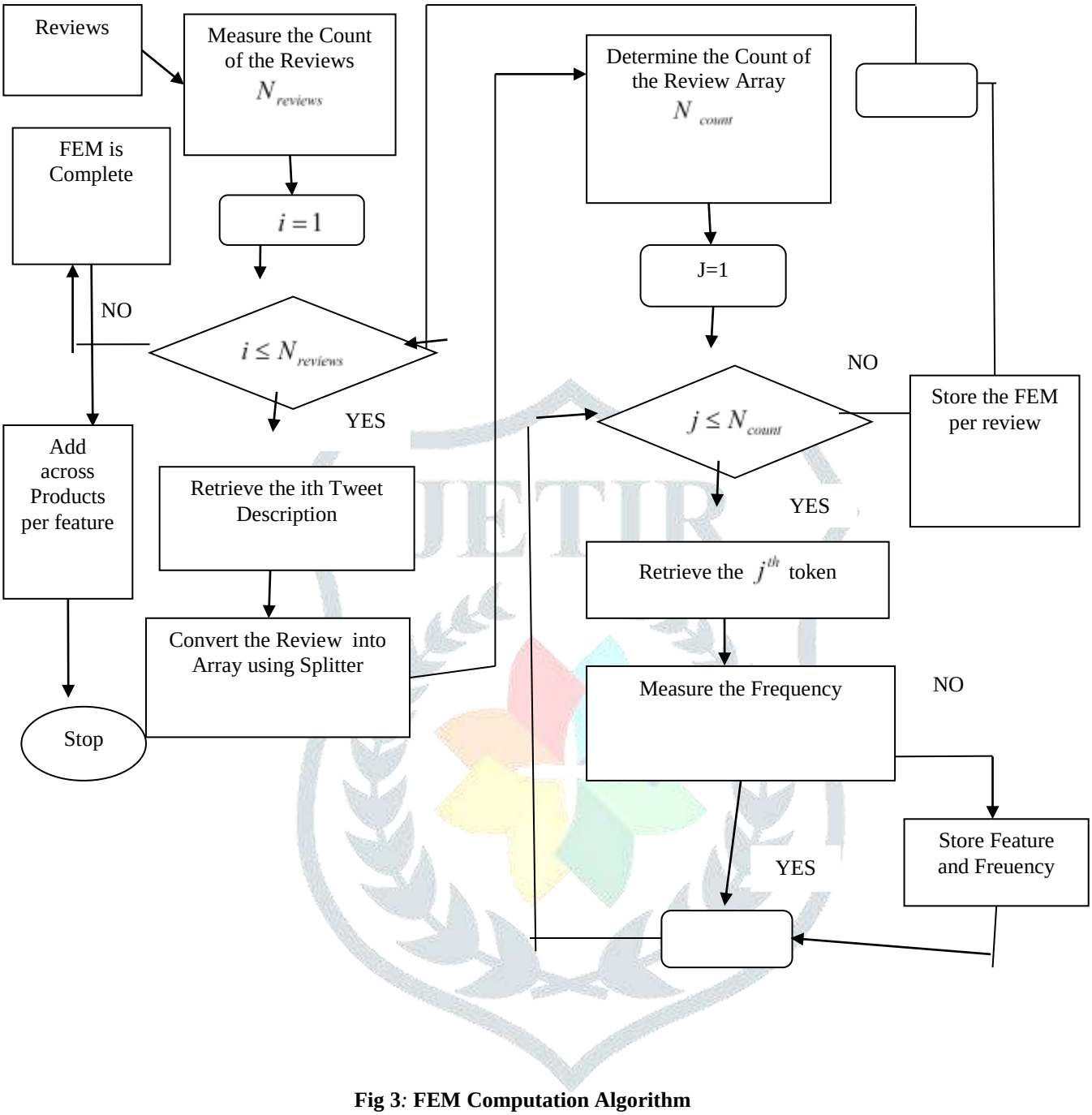

Fig 3: FEM Computation Algorithm

Polarity Computation Algorithm

This Algorithm computes the Polarity of the user on the basis of positive and negative keywords maintained in the database. This is the requirement for the calculation of Activation probability to calculate the sentiment of the person whether it is positive,negative or neutral towards the product.

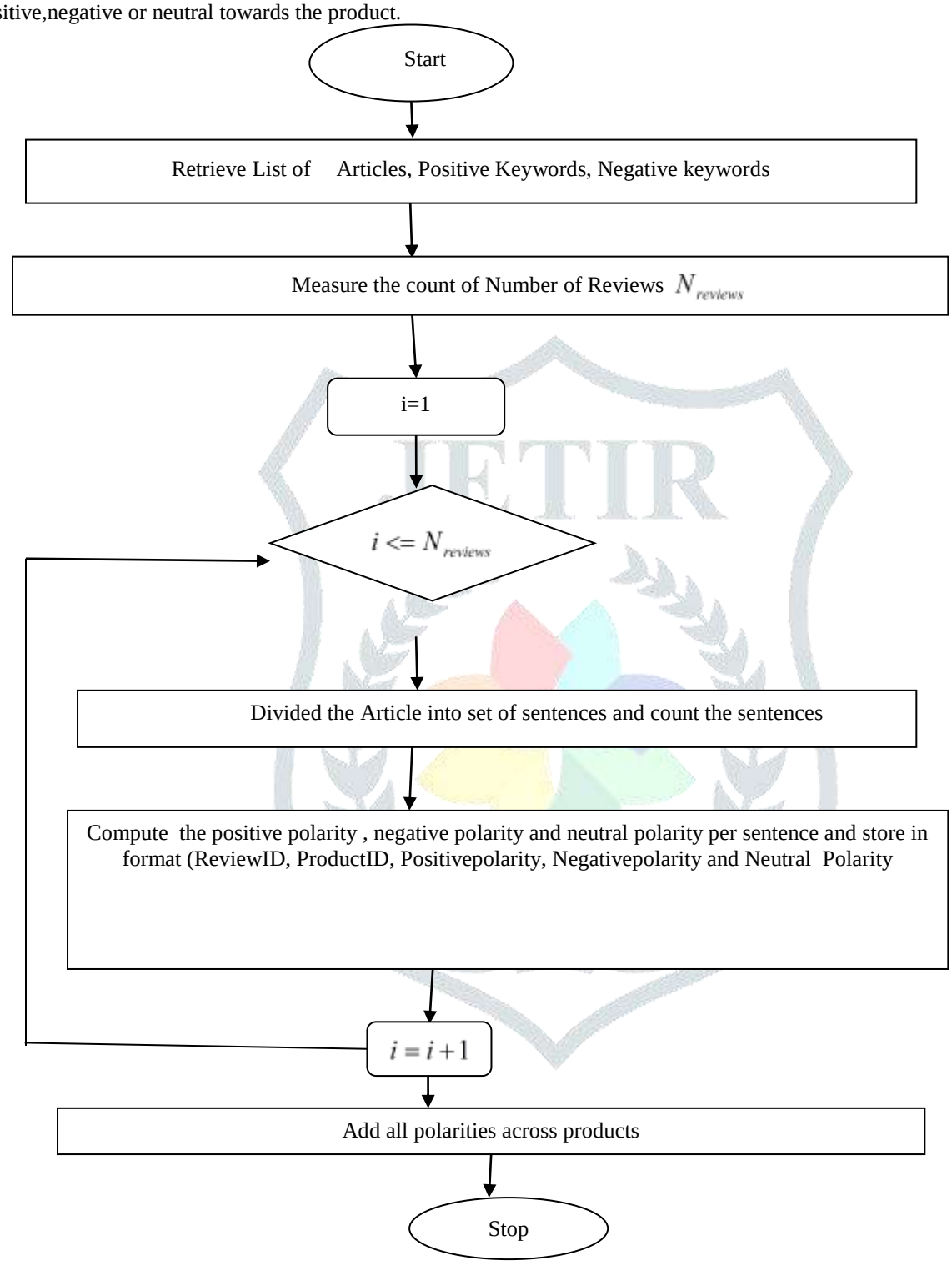

Fig 4: Polarity Computation Algorithm

Activation Probability

This is the Probability which is the probability that user will be influenced by the feature of the Product .The activation probability can be computed as follows

1) Measure for each of the tweets per feature the positive ratio and the other category ratio

$$PositiveRatio_i = Positive\ Polarity_i + Neutral\ Polarity_i + Feature\ \ Based\ \ Frequency_i$$

$$OtherCatRatio_i = Negative\ Polarity_i$$

Where, (1)

$$f1 \le i \le f6$$

$$f = feature$$

2) After computing each of the ratio's then find out the group to which the tweet belongs to and user belongs based on order by Positive Ratio maximum and Negative Ratio Minimum.

3) After Performing the grouping count the number of tweets under each category

4) Measure the Probabilistic measure by using

$$Activation\ \Pr obability_{\ group1|feature1} = \frac{N_{tweetsgroup1}}{N_{tweetsgroup5} + N_{tweetsgroup2} + N_{tweetsgroup3} + N_{tweetsgroup4+} N_{tweetsgroup5}} \quad (2)$$

IV. ACKNOWLEDGMENT

The work, reported in this paper, is supported by the college through the Technical Education Quality Improvement Program [TEQIP-II] of the MHRD, Government of India

REFERENCES

[1] http://www.ijarcsms.com/docs/paper/volume4/issue2/V4I2-0020.pdf

[2] Amit Goyal, Francesco Bronchi, Laks V. S. Lakshmanan, A data based approach to Social Influence Maximization

[3] X. Ding, B. Liu, and L. Zhang, "Entity discovery and assignment for opinion mining applications," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09),* 2009.

[4] S. M. Kim and E. Hovy, "Determining the sentiment of opinions," in Proceedings of Interntional Conference on Computational Linguistics (COLING'04), 2004.

[5] N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'06). 2006

[6] W. Jin, H. H. Ho, and R. K. Srihari, "OpinionMiner: a novel machine learning system for web opinion mining and extraction," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discoveryand Data Mining (KDD'09),* 2009b.