Influence Quantification of Online Review Comments

Debanjan Banerjee[1], BikromadittyaMondal[2]
[1]*Dept. of MS-SDE, Manipal Universal Learning,*
Kolkata - 700091, India,
[2]*Dept. of Computer Science & Engineering,*
B P Poddar Institute of Management & Technology
Kolkata - 700052, India,

ABSTRACT

The exponential growth of online expressions produce the need for automated tools for analysing potentially beneficial patterns and trends for the society at large. In this paper we propose a framework to analyse and predict possible influence in the sentence-level online opinions by combining the fundamental influential factors. We have identified these fundamental factors as the opinion maker, the domain of the opinion, the polarity of the expressed opinion and the manner of representation. The proposed framework comes with a mathematical formulation for quantifying the potential influence by combining the effects of these factors in the review comments with respect to receiving "likes" & "dislikes" as well as feedback comments.

Keywords—**Mindset analysis, Sentiment analysis, Favourability analysis, Influence quantification, Natural language processing.**

1. INTRODUCTION

The rise of the World Wide Web and the subsequent rise of the social networks, online user forums, blogs and tweets have brought about an unprecedented way in which humans communicate, explore, inspire and influence each other. The study of sentiment analysis attempts to understand these patterns in day to day communications. One of the most important practices of social network is to influence others with their own ideas, judgments and opinions. Understanding influence through the expression of online communications is an important problem in the realm of sentiment analysis. It is important to analyse potential factors which play an important role behind how any potential idea would gain influence in a social sphere. The study of influence involves chiefly three major areas of interest. These areas can be defined as the potential for any prospective influencer to influence to a certain degree, the potential by the target audience to be influenced by the influencer and the various ways in which the influenced would take action after being influenced. In our work we propose a framework whose primary objective is to quantify the factors involved with the influence associated with the review comments. This paper proposes a method which attempts to combine four distinct factors and predict a potential influence score. These four factors are the person of the opinion maker, the topic about which the review comment has been expressed, the polarity or favourability of the review comment with regards to the topic and the manner of representation of the opinion. Our proposed predictive method generates influence scores with the specific purpose of analysing the trends with regards to influence in a social forum.

2. RELATED WORK

The research community has been working on studying different aspects of social media for the past few years. Smith et al [1] worked on the general purpose analysis of social networks and Tang et al [2] worked on the structured analysis of social networks. Research has been made analyzing the twitter by Java et al [3]. Hui and Gregory [4] worked with a framework for quantifying the sentiment and influence in weblogs. Our work differs with these past works in the sense since we quantify the influence amongst review comments. The topic or subject about which the opinion maker decides to make a comment is a very important aspect in terms of winning over influence. The work of dealing with the subjective properties in any review comment is a significant challenge. Farrar and Moran [5] presented a software framework which deals with the subject of extraction of subjective data formats, Fiscus and Doddington[6] offered a topic detection and tracking program for developing technologies that search, organize and structure multilingual, news oriented textual materials from a variety of broadcast news media, Kim and Myaeng [7] presented a method whereby finite state automata and lexicon for improving topic detection abilities, Makkonen et al[8] proposed a technique which incorporates simple semantics into TDT by splitting the term space into groups of terms, Yang et al[9] made an analytical comparison of many different techniques for topic extraction in subjective documents. Our objective lies in terms of using topic as a source of quantification for the objective of the quantification of influence whereas these works primarily deals with extracting the topic from the subjective texts. Extracting and analyzing the sentiment in the opinionated expressions has also been a significant problem area for sentiment analysis. Choi et al [10] presented a framework which exploited sentiment topic information for generating context-driven features, Lin and He[11] proposed a method for extracting sentiment and topic simultaneously from an opinionated expression, Thet et al[13] proposed a clause level sentiment analysis of opinionated texts. The work of this paper is different from these works since the objective of these works lie in extracting the sentiment from the opinionated expressions whereas our objective lies in the quantification of sentiment as a factor for understanding its contribution behind influence of any opinionated expression.

3. PROPOSED WORK

In this work we identify the potential top influencers in a review web forum. At first the framework collects the reviews of all the reviewers in a particular forum. After collecting the necessary information the next step is to apply necessary natural language processing functions to extract a) the reviewer name b) the exact topic(s) on which the review has been provided c) the rating provided by the reviewer d) the count of the total "likes" which are associated with the review and e) the count of the total "feedback" comments against the particular review. The two step process flow of the analysis is shown in Figure1. Then the proposed framework derives an influence factor based on the information obtained by the natural language processing functions. The main purpose behind the derivation of the influence factor is to quantify all potential sources of subjective information which are relevant in the realm of an online review forum.

Figure1. Building blocks of the influence analysis framework

3.1 SUBJECTIVE PROPERTY EXTRACTION PROCESS

Here we assume that any review comment say *RC* consists of one or more topic(s) $T_n$where n ≥ 1, a reviewer *Re*, a rating *S* which implies sentiment polarity and corresponding set of likes $L_n$where $n \geq 0$ and feedback comments $C_n$where $n \geq 0$. We apply separate natural language process functions to extract the necessary subjective properties which are relevant topic $T_n$and sentiment *S*. The framework uses the functions *TOP* (*RC*,T_n, *Len*), *SENT (RC, S)* and *REPRESENT* (*RC*, *REP*) for deriving subjective properties which are topic and sentiment with respect to the review comment *RC*.The function *TOP* (*RC*,T_n) considers the review comment *RC* and a pre-defined list of topics as input and then from the list of topics $\sum_{i=1}^{n} T_i$this function picks out the particular topic T_i and identifies the same topic in the review comment *RC* and then returns the a quantified value which is between 0 and 1 for the said topic(s). The reason for selecting the topic as an important subjective property in terms of quantification is that if a person comments on a topic which is a very popular one in a particular period of time then there is a considerable amount of chance of the review comment being more influential than the topic on a subject which is not very frequently commented.So the topic *T*is expressed as:

$T = \sum_{k=1}^{n} tm_k * \sum_{i=0}^{a} Ref\ (1)$

The function *SENT (RC, S)* takes a pre-defined set of sentiment ratings [1,5] as well as the *RC* as input andthen it produces the sentiment polarity value for the review comment *RC*. The expression of the function *SENT (RC, S)*is:

$SENT\ (RC, S) = \{S \in RC \mid S \in [0,1]\}$ (2)

The function *REPRESENT* (*RC*, *REP*) provides us with the sentiment value which indicates the quantified value of how the user represents his/her viewpoint in the prospective review comment *RC*. It is observed that there are two ways in which a person represents his/her viewpoint. We define these two varieties of a person representing his/her viewpoint as intelligent representation and emotional representation respectively. We define a set of words [because, since, so, thereby] and the usage of these particular term in any subjective expression is termed as intelligent representation. We categorize absence of any of these words in an opinionated expression as emotional representation. The utmost apparently visible difference between the emotional and intelligent representations lies in the fact that in case of intelligent representation of the opinionated expression, the person provides a cause for his/her feelings, view or judgment which is observable from the use of the aforementioned set of words whereas in case of the emotional representation we certainly observe an absence of any apparent cause. The function *REPRESENT* (*RC*, *REP*) takes a review comment *RC*as input and then assigns a numeric value in the range of [0,1] to each sentence in the subjective expression*RC*. Here the logic is whether the particular sentence in question contains the one/more words belonging to the set. This value is termed as the representation evaluation value or *Re*. Then the function calculates the sentiment representation value *REP* which is the mean value for all representation evaluation values. The *REP* value ranges between [0,1] and if this value is greater than 0.5 and thereabout we term the overall review comment *RC* as intelligently represented and if the less than 0.5 then we categorize the overall review comment *RC* as emotionally represented. So REP is expressed as:

$REP=(\sum_{m=0}^{c} Re_m)/c$ (3)

3.2 QUANTIFICATIONPROCESS

For each review comment RC on a topic $t \in \sum_{i=1}^{n} T_i$ we select the opinion maker OM. The purpose behind the selection of *OM* as a subjective property which impacts influence is the observation that certain opinion makers possess a wide range of followers and friends thus there is a greater percentage of their comments being able to obtain more "like"-s and feedback comments than the opinion makers with a relatively less amount of friends and followers.

We derive *OM* as proportionate to the total number of followers and friends connected with the particular user. Therefore *OM*is expressed as:

$OM = b * (\sum_{i=0}^{e} fr_i)$(4)

Where *fr* represents the total number of friends and b is a coefficient. The value of *OM*lies in the range between 0 to 1.

The opinionated expression itself contains at least one sentiment *S* which could be either positive or negative and the expressed sentiment has been represented in *REP*. The review comment *RC* receive a number of likes $L = \sum_{i=0}^{n} L_i$and feedback comments $C = \sum_{i=0}^{n} C_i$. So we express the potential influence in a review comment in terms of:

$IFF = a*(L+C)$(5)

Where*a*is a coefficient whose value is ½ and *L* can be expressed as:

$L = 1/4(T+REP+OM+S)$(6)

And*C* can be expressed as:

$C = 1/4(T+REP+OM+S)$ (7)

4. RESULTS

We have implemented the proposed methods and tested on data set obtained from goodreads.com book review website for the book "Coming apart" by Charles Murray. Summary of experimental result is shown in Figure 2. Our results show that among the total reviewers having any influence, the highest percentage belongs to those who provide review rating 1 and the largest share amongst the non-influential reviewers belongs to those reviewers who provide the review rating of 4.This particular group of reviewers who provide the review rating of 4 is the reviewer group which has contributed the largest percentage of review comments compared to any other group of reviewers.

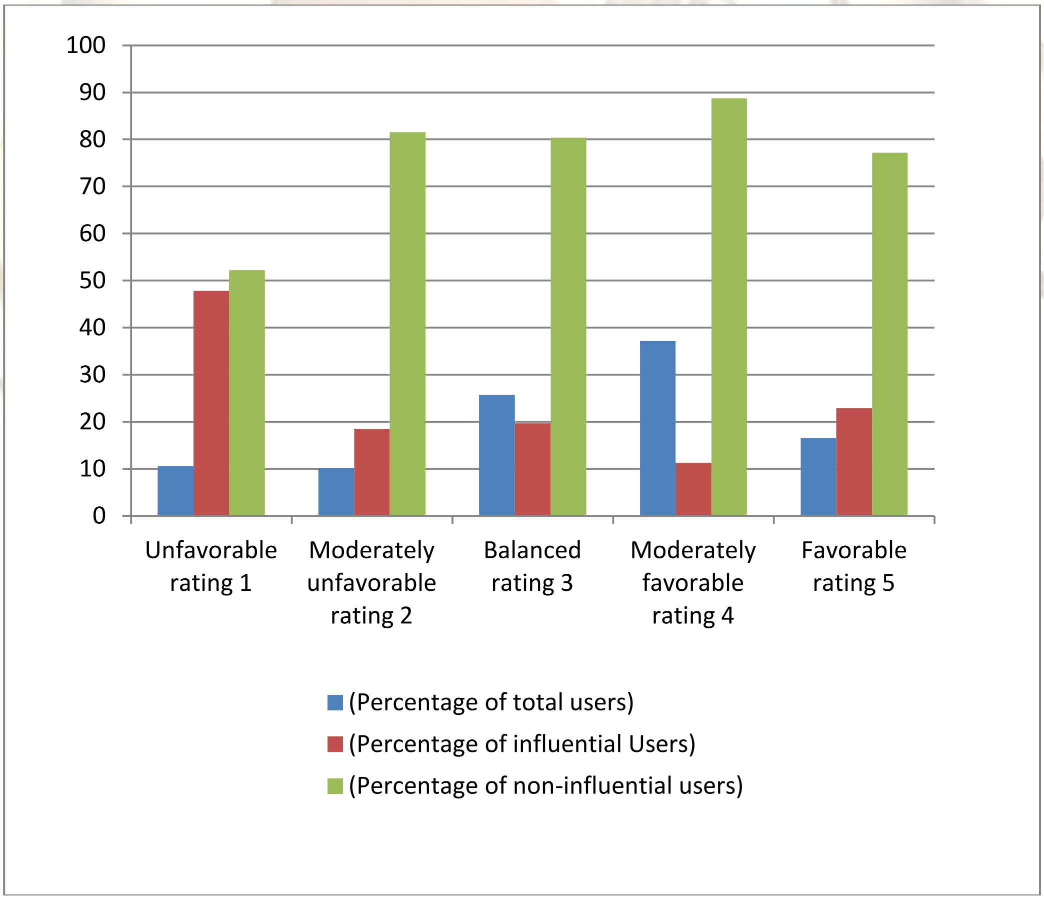

Figure 2: The total percentage of influential users with respect to each sentiment level.

5. CONCLUSIONS

In this work we present a framework for understanding and analysing the influence which is inherent in the opinionated expressions expressed at a popular review forum. This approach may be applied for reviews in cross domains with more versatile data. We also observed that human opinions on socio-political domain depends on the personal socio-political understanding of the subject in question and the existing socio-political belief system under which the person works. Therefore future work may be done consideration these factors in a combinational framework alongside the prevailing combination of the favorability and mindset classes and then observing the potential influence of the overall opinion sentiment on the target people to whom the opinions have been addressed.

REFERENCES

[1] Jonathan G. Fiscus, George R. Doddington, Topic detection and tracking evaluation overview, *The Information Retrieval Series*, *12*, 2002, 17-31

[2] Pyung Kim, Sung HyonMyaeng, Usefulness of temporal information automatically extracted from news articlesfor topic tracking, *ACM transactions on asian language information processing (TALIP), 03(04)*, pp. 227-242,2004.

[3] Juha Makkonen,Helena Myka-Ahonen, MarkoSalmenkivi, Simple semantics in topic detection and tracking, *Information retrieval, 07(03-04)*, 2004, 347-368.

[4] Yiming Yang, Jaime Carbonell, Ralf Brown, John Laffetry, Thomas Pierce, Thomas Ault, Multi-strategy learning for topic detection and tracking: a joint report of cmu approaches to multilingual , *The information Retrieval Series* , (*12*), 2002, 85-114.

[5] Bo Pang, Lillian Lee, Opinion mining andsentiment analysis,*Foundations andTrends in Information. Retrieval, 02(01-02)*, 2008, 1-135.

[6] Mark A. Smith, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, Eric Gleave,Analyzing (social media) networks with noxdel, *Proceedings of the Fourth International conference on communities and technologies* , ASM, USA, 2009, 255-264.

[7] Jintao Tang, Ting Wang, Ji Wang, Dengping Wei,Efficient social network approximate analysis on blogosphere based on network structure characteristics , *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, ACM , France, 2009, 1-8.

[8] Akshay Java, Xiaodan Song, Belle Tseng,Why we twitter: understanding Microblogging usage and communities, *Proceedings of the 9th WebKDD and 1st SNA KDD 2007 workshop on Web mining and social network analysis*, USA , 2007 , 55-65.

[9] Peter Hui, Michelle Gregory, Quantifying Sentiment and Influence in Blogspaces, *1st Workshop on Social Media Analytics*, USA, 2010, 53-61.

[10] Scott Farrar, Steve Moran, Thee-linguisticstoolkit, *Workshop in the 4th IEEE international conference e-science*, IEEE, USA, 2008

[11] YoonjungChoi,YounghoKim,HyonSungMyaeng, Domain-specific sentiment analysis using contextual feature generation, *Proceeding of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion*, ASM, China, 2009, 37-44.

[12] Chenghua Lin, Yulan He,Joint sentiment/topic model for sentiment analysis, *Proceeding of the 18th ACM conference on Information and knowledge management*, ACM, USA ,2009, 375–384.

[13] TunThuraThet, Jin-Cheon Na, Christopher S.G.Khoo, SubbarajShakthikumar , Sentiment analysiof movie reviews on discussion boards using a linguistic approach, *Proceeding of the 1st international CIKM workshop on topic sentiment analysis for mass opinion* , ACM , USA , 2009, 81-84.