

Data Mining Case Study for Analysing Opinion Mining from Emotions in Social Network Sites in Theni Dt., Tamilnadu

B. Nasreen Fathima¹, S. Jothi², R. Aswini³

¹PG Student, Department of Computer Science,
Jayaraj Annapackiam College for Women, Periyakulam

²Assistant Professor, Department of Computer Science,
Jayaraj Annapackiam College for Women, Periyakulam

³PG Student, Department of Computer Science,
Jayaraj Annapackiam College for Women, Periyakulam

Abstract: In today's scenario Social Networking Sites (SNS) are a part and parcel of our lives that are firm to be avoided by anyone. These SNS are contributing in bringing people jointly from various communities of like interest at a single collaborated destination. Also this SNS helps people in distribution their views, experiences, likes and dislikes about a common topic of interest. Moreover most of the active and energetic users have a great impact on influencing others views and opinion. Such an impact can be unenthusiastic as well as positive depending upon the conveyer and receptor's capability to think and understand and these factors emotion on how a person react to a particular state of affairs and there by influence others. Teenagers are among the most prolific users of SNS. Emerging studies find that youth spend a considerable portion of their daily life interacting all the way through social media. Subsequently, questions and controversies come out about the Emotion in SNS have on adolescent development. This case study is an exploratory research conducted to the students of colleges and schools in Theni and Periyakulam. This research strands articulate to high-profile concerns and controversies that enclose youth participation in these online communities, and offer ripe areas for future research.

Key Words: Data Mining, Opinion Mining, Apriori Algorithm, R Tool

1. INTRODUCTION

1.1 Introduction:

In today's scenario social networking sites are a part and parcel of our lives that are hard to be avoided by anyone. These SNS are contributing in bringing people together from various communities of similar interest at a single collaborated destination. This helps people in sharing their views, experiences, likes and dislikes regarding a common topic of interest. Moreover most of the active and vibrant users have a great impact on influencing others views and thoughts. Such an impact can be negative as well as positive depending upon the conveyer and receptor's ability to think and understand and these factors effect on how a person react to a particular situation and there by influence others.

The computer-aided detection, analysis and application of emotion, particularly in text, have been a growth area in

recent years. Almost all of this research has focused on detecting opinions in large bodies of text. From a wider social perspective, emotion is important to human communication and life and so it seems that the time is ripe to exploit advances and intuitions from opinion mining in order to detect emotion in a wider variety of contexts and for primarily social rather than commercial goals.

1.2 Statement of the Problem:

Social media, like Twitter, YouTube, Facebook and Flickr have changed the way people interact with each other during past decades. Because people is human, their interactions online always show social activities, forming different groups according to their preferences, hobbies, education levels.

We try to use machine learning method to extract groups from network and predict people into different groups related to their preference, hobbies, and education levels.

2. ALGORITHM USED

Data mining is the application of efficient algorithms to detect the desired patterns contained within the given data. To analyze the large amount of collected information, the area of Knowledge Discovery in Database (KDD) provides techniques which extract interesting patterns in a reasonable amount of time.

2.1. Apriori Algorithm

Apriori algorithm is a fundamental algorithm mining association rule. It contains two processes:

- Detect all frequent itemsets by scanning db.
- Form strong association rules in the frequent itemsets. Process one needs to scan DB several times, which consumes a lot of time and space. As a result, what needs to be improved is the mining competency of frequent group of things in DB. Apriori is used to detect all frequent itemsets in a provided database db. The keynote of Apriori algorithm is to form multiple passes over the database. It employs repetitive approach called as a breadth-first search (level-wise search).

- **Frequent Itemsets:** The itemsets which has minimum help (denoted by l_i for i^{th} -itemsets), Apriori property: any subgroup of frequent things must be frequent.
- **Join Operation:** to detect l_k , a group of candidate k -group of things is developed by adding l_{k-1} with itself.

2.2. Association Rule:

Association rule of data mining involves picking out the unknown inter-dependence of the data and finding out the rules between those items. A rule is defined as an implication of the form $A \Rightarrow B$, where $A \cap B \neq \emptyset$. The left-hand side of the rule is called as antecedent. The right-hand side of the rule is called as consequent.

Support: $I = \{i_1, i_2, i_3, \dots, i_m\}$ is a collection of items. T be a collection of transactions associated with the items. Every transaction has an identifier TID. Association rule $A \Rightarrow B$ is such that $A \in I, B \in I$.

A is called as Premise and B is called as Conclusion. The support, S , is defined as the proportion of transactions in the data set which contains the itemset. **Support ($X \Rightarrow Y$) = Support (XUY) = $P(XUY)$.**

Confidence: The confidence is defined as a conditional probability $\text{Confidence } (X \Rightarrow Y) = \text{Support } (XUY) / \text{Support}(X) = P(Y/X)$. Lift: is the ratio of the probability that L and R occur together to the multiple of the two individual probabilities for L and R , i.e. $\text{lift} = \text{Pr}(L,R) / \text{Pr}(L) \cdot \text{Pr}(R)$.

Conviction: is similar to lift, but it measures the effect of the right-hand-side not being true. It also inverts the ratio. So, a conviction is measured as:

$$\text{conviction} = \frac{\text{Pr}(L) \cdot \text{Pr}(\text{not } R)}{\text{Pr}(L,R)}$$

Data Mining in Social Networking sites

A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest. **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior. A **social networking service** is a platform to build social networks or social relations among people who share interests, activities, backgrounds or real-life connections. A social network service consists of a representation of each user, his or her social links, and a variety of additional services. Social networking sites allow users to share ideas, pictures, posts, activities, events, interests with people in their network.

When mining social network data it should be a combination of web structure mining and web content mining. Analyzing the structure of the Social network is known as Social Network Analysis.

2.3 Graph mining algorithms

Web can be considered as a directed graph, where nodes will be the hypertext documents and edges will be hyperlinks. The Algorithm is based on frequent pattern mining in transactional and graph databases with periodic pattern mining in unidimensional and multidimensional sequences.

2.4. Functions of Data Mining

While large-scale information technology has been evolving separate transaction and analytical systems, data

mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of **four types of relationships** are sought:

- **Classes:** Stored data is used to locate data in predetermined groups.
- **Clusters:** Data items are grouped according to logical relationships or preferences..
- **Associations:** Data can be mined to identify associations.
- **Sequential patterns:** Data is mined to anticipate behavior patterns and trends.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- **Artificial neural networks:** Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms:** Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Decision trees:** Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- **Nearest neighbor method:** A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where $k \geq 1$). Sometimes called the k -nearest neighbor technique.
- **Rule induction:** The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization:** The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

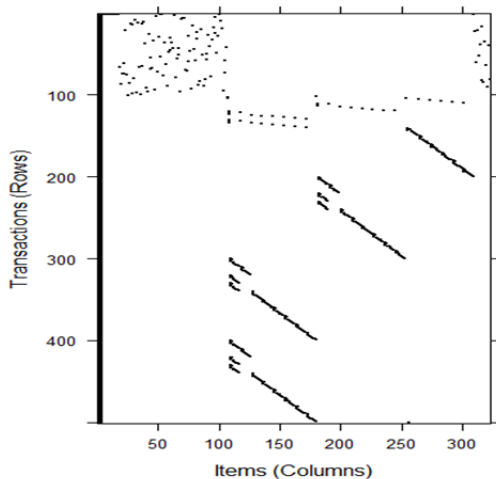
3. ANALYSIS OF DATA

The researchers collected more than 400 responses from samples all over Theni District. After collecting the information, all the details are fed into the software and checked for outlier. The cleaned data was analyzed using single attribute and multiple attributes. Age, Gender, qualification, religion and Community are considered as single attributes of the study.

```
>groupfsets=eclat(trans, parameter =
list(support=0.05),control=list(verbose = FALSE));
>singleItems=fsets[size(items(fsets))=1];
>singleSupport=quality(singleItems) $support;
>names(singleSupport)=unlist(LIST(items(singleItems),
decode = FALSE));
> head(singleSupport, n = 5);
576 955 739 317 185 0.122 0.084 0.074 0.062 0.063
> itemsetList = LIST(items(fsets), decode = FALSE);
> allConfidence = quality(fsets)$support /
sapply(itemsetList, function(x)+
max(singleSupport[as.character(x)]));
> quality(fsets) = cbind(quality(fsets), allConfidence);
> summary(fsets);
```

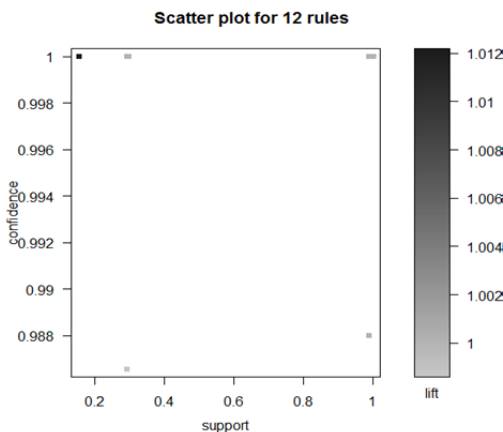
3.1. Conducting the experiment:

```
> image(tr)
```

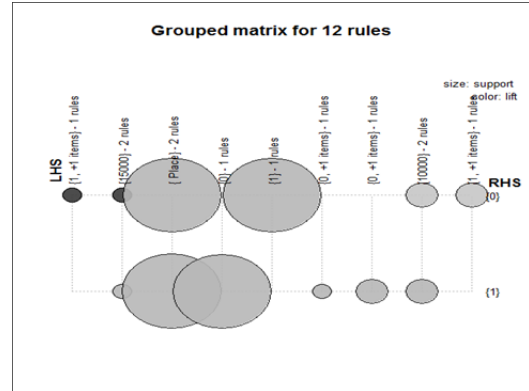


Parameter specification:

```
> plot(rules)
```

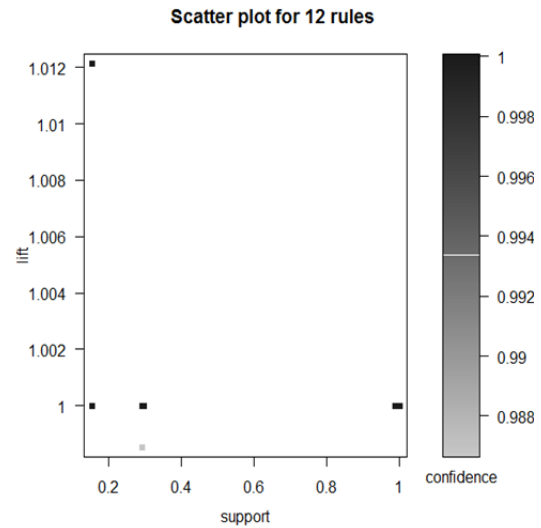


```
> plot(rules,method="grouped")
```

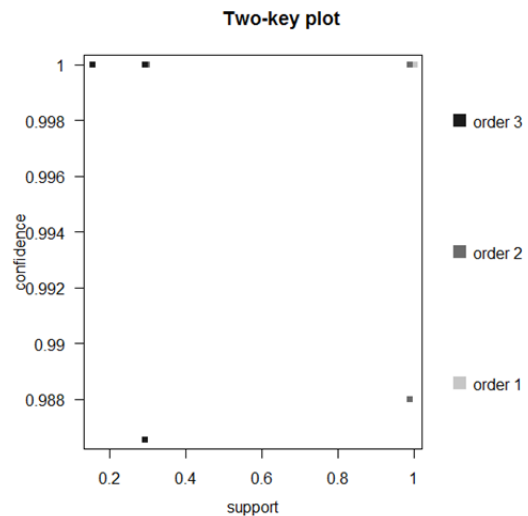


```
> head(quality(rules));
```

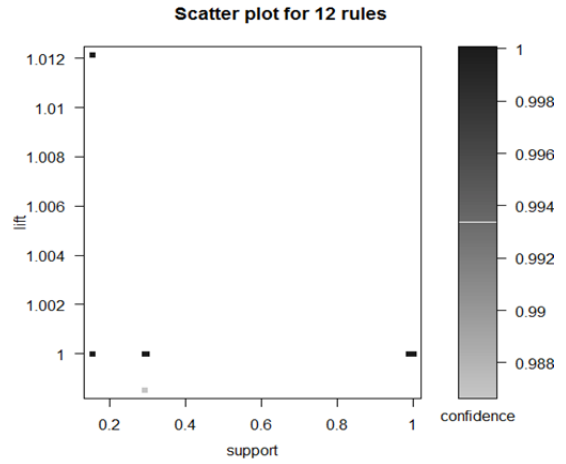
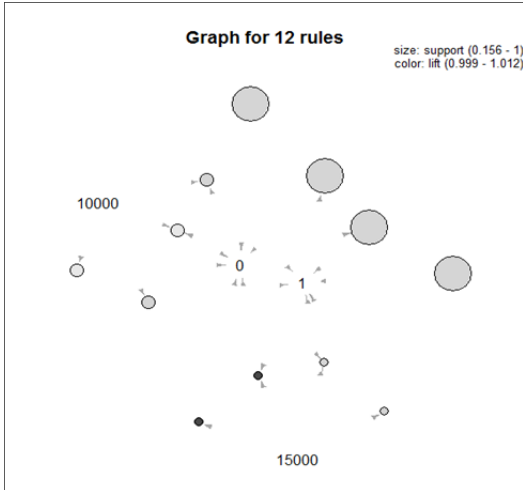
```
> plot(rules, measure=c("support","lift"),
shading="confidence");
```



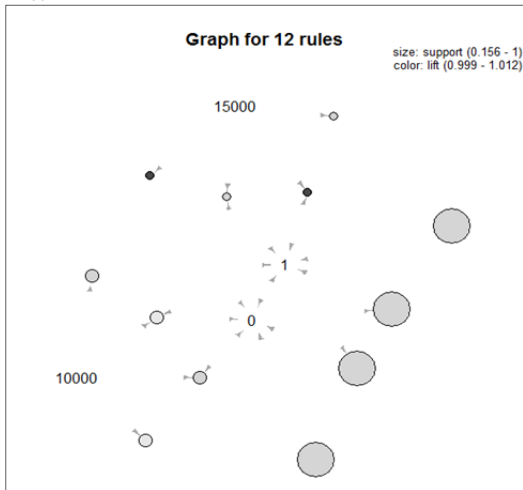
```
>plot(rules, shading="order", control = list(main
="Two-key plot");
```



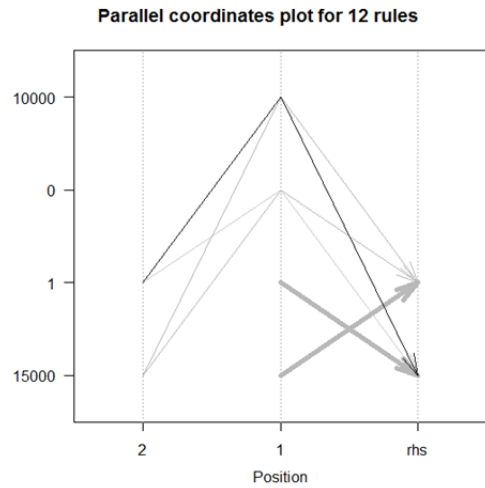
```
> plot(rules, method = "graph")
```



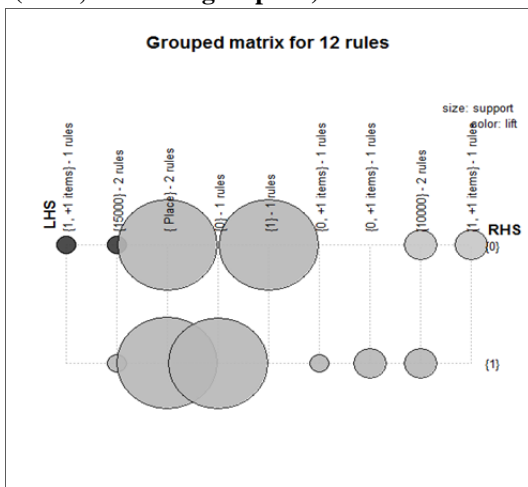
```
> plot(rules, method = "graph", control = list(type = "items"))
```



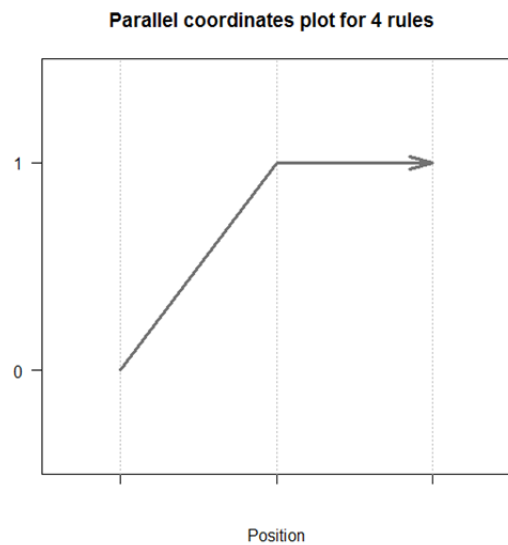
```
> plot(rules, method = "paracoord", control = list(reorder = TRUE))
```



```
> plot(rules, method="grouped")
```



```
> plot(subrules2, method="paracoord");
```

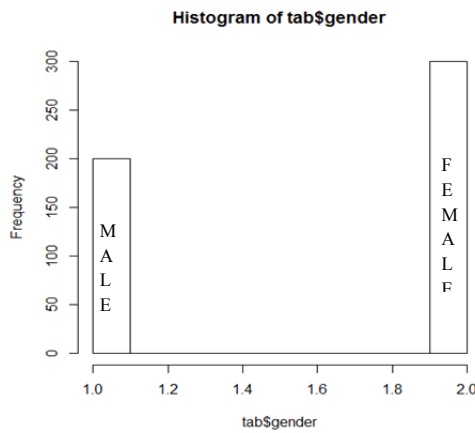


```
> head(quality(rules));
> plot(rules, measure=c("support", "lift"), shading="confidence");
```

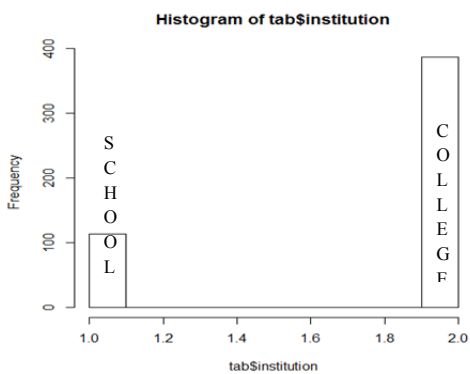
4. FINDINGS AND INTERPRETATIONS:

Information collected by the researchers and analyzed by using R tool of Data mining. In a day and age where teens are sending Snapshots instead of passing handwritten notes and "selfie" has become a regular part of our vocabulary, there's no denying that social media is impacting the way teenagers view themselves. Fortunately, studies show that 20 percent of students say social media makes them feel more confident, compared with 4 percent who said it makes them feel less so. Unfortunately, sometimes teenagers forget that their identity is more than what they project on Facebook

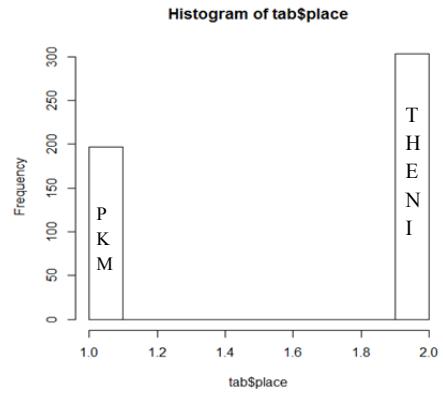
- **Gender Based Impact using SNS :** As we go through the research carefully, we can easily trace that the female users are greatly addicted for the SNS irrespective of their educational qualification.



- **Education Level Based Impact using SNS :** Our research shows that teenagers especially college students are greatly addicted for the SNS than the school children.



- **Geographical Based Impact using SNS:** Our research shows SNS users are more in cities. That is more number of SNS users are in Theni than Periyakulam.



5. RECOMMENDATIONS

After getting all findings and discussed the conclusion of collected data the researcher recommend some measures to use of social media in right direction and utilize social media favorable and appropriate manner to its users.

- Positive use of social media can develop the youth's academic career, their skills, better living style, to adopt new trends, fashion, and anthropology so on.
- Social media is recent and most favourite form of media. It is a useful tool for youth so its use is essential to get information and knowledge when youth going to connect the social media should keep in mind that basic purpose to usage and always remember that they are going to share the information or links are not only for their gratification and interest but also for all their contacts and friends community, so be carefully utilize with social responsibility, ethically, religiously and politically appropriate links should be share.
- While sharing among groups, keep in mind the society standards, social norms and different religion values. Do not share the links that create hatred different communities among the different segments, groups, sects, religions, cultures and races. The relevant information should be preferred on social media sites.
- There should extra-curriculum and awareness forum in the educational institution regarding the instruction about the positive usage of social media networking portals.
- Since the last few years, the use of social has significantly increased to make perception regarding the socio-political images. The users should be aware about the right to information which is provided to them from their respective states and societies. They should avoid from defamation and hate speech on the social media forums.
- Healthy mind get progress in healthy physique. The youth should avoid from excessive use of social media. They should be balanced in their life and give proper time to the co-curricular activities in their daily life routine.

6. SUGGESTIONS TO REMAIN SAFE IN SNS:

From the result we conclude, if there is awareness about the risk of social media in our community it will not lead to anything bad but there is always lack of public awareness and as mentioned above that the graph of internet users is getting higher and higher while we are still far behind in the field of education. Thus public awareness is very difficult in societies with lack of education.

Unfortunately, most teenagers don't realize what they post or share on Facebook. The best thing for us to do is to teach the younger generation, that the ripple effect of social media can make and how to put their best foot forward online.

The users of SNS must be aware of certain things when discussing their private life in public. These are some of the tips that can make them and their family safe on the networking sites.

- Change the profile as privacy.
- Don't accept friendship request from strangers.
- Don't post very personal information on the profile.
- Be cautious while posting your photo.
- Don't post your current location when on a tour.
- Don't post negative things about your life.
- Make distance from your ex's profile.
- Don't substitute real friends with virtual friends.
- Avoid using Social networking sites in work hours.

7. CONCLUSION

The usages of the Social Networking Sites are increasing daily. As the results show, students spend a significant amount of their times using the Social Networking Sites. As a consequence of this increase in usage, students are establishing communication with people in their environments. Students increase their knowledge and communication skills by sharing photographs, links that they like, and news with their friends. Future studies should concentrate on integrating the Social Networking Sites into education and teaching, which is important in students' everyday working lives. In addition, Social Networking Sites provide individuals with a way of maintaining and strengthening social ties, which can be beneficial in both social and academic settings.

BIBLIOGRAPHY

- [1] G.Vinodhini and RM.Chandrasekaran, Sentiment Analysis and Opinion Mining: A Survey, Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering
- [2] Zhongwu Zhai, Bing Liu, Hua Xu and Hua Xu, Clustering Product Features for Opinion Mining, WSDM'11, February 9-12, 2011, Hong Kong, China. Copyright 2011 ACM 978-1-4503-0493-1/11/02...\$10.00
- [3] Singh and Vivek Kumar, A clustering and opinion mining approach to socio-political analysis of the blogosphere, Computational Intelligence and Computing Research (ICIC), 2010 IEEE International Conference.
- [4] Alexander Pak and Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining,
- [5] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.