Sentiment Analysis in Social Media Platforms: The Contribution of Social Relationships

Research-in-Progress

Shaokun Fan[1]

West Texas A&M University

2501 4th Ave, Canyon, Texas 79016, USA

sfan@wtamu.edu

Noyan Ilk[1]

Florida State University

College of Business, 821 Academic Way, Tallahassee, Florida 32306, USA

nilk@business.fsu.edu

Kunpeng Zhang[1]

University of Maryland, College Park

4316 Van Munching Hall, College Park, Maryland 20742, USA

kzhang@rhsmith.umd.edu

Abstract

The massive amount of data in social media platforms is a key source for companies to analyze customer sentiment and opinions. Many existing sentiment analysis approaches solely rely on textual contents of a sentence (e.g. words) for sentiment identification. Consequently, current sentiment analysis systems are ineffective for analyzing contents in social media because people may use non-standard language (e.g., abbreviations, misspellings, emoticons or multiple languages) in online platforms. Inspired by the attribution theory that is grounded in social psychology, we propose a sentiment analysis framework that considers the social relationships among users and contents. We conduct experiments to compare the proposed approach against the existing approaches on a dataset collected from Facebook. The results indicate that we can more accurately classify sentiment of sentences by utilizing social relationships. The results have important implications for companies to analyze customer opinions.

Keywords: Sentiment analysis, social media, social networks, bag-of-words, big data

[1] Authors are ordered alphabetically.

Introduction

Over the past decade, social media has become one of the most popular communication mediums that allow users to share their opinions on various topics with people in their social networks. For example, Facebook allows companies to create fan pages and customers may comment on or like campaigns posted by a company. Twitter users could send tweets with a maximum length of 140 characters to share their opinions on various topics. Ecommerce platforms such as Amazon.com allow users to post product reviews. Consequently, social media driven information contains opinions and sentiments on various topics of interest and is extremely valuable for companies to design marketing strategies (Tan et al. 2011).

Sentiment analysis (SA) is one of the emerging technologies in the effort to devise such strategies and to help people understand the massive amount of data available online (Pang and Lee, 2008). A major goal of SA is to determine the overall sentiment orientation of a text document at the sentence level. This problem has been extensively investigated in literature. Most of the existing SA approaches predict the sentiment of a sentence by solely using the textual contents of that particular sentence. Although these methods are relatively effective for sentences collected from traditional platforms, they were built on the assumption that these sentences are fully representative of their authors' genuine opinions and are the only available information to infer author sentiments. However, such assumptions might not hold for social media platforms. For example, users in social media platforms are from all over the world and they may use multiple languages. In addition, social media postings are commonly prone to misspellings and may contain special characters (e.g. emoticons or abbreviations) that are harder to analyze (Hu et al. 2013). As a result, current SA systems that solely rely on the textual content may be ineffective in analyzing sentiment in social media platforms.

In this paper, we tackle this problem by proposing a framework for SA based on social relationships in social media platforms. Inspired by the attribution theory (Kelley 1967), we propose to use two types of social relationships (e.g., user-topic and user-user relationships) to capture the internal and external causes of user sentiments. Then, we combine traditional textual analysis with social relationships to improve the SA process. We evaluate our framework by comparing it with a traditional SA method on a data set collected from Facebook. The results show that introducing social relationships features into SA models can help increase the model performance.

Literature Review

There are two main streams of research on sentiment analysis. The first stream studies the application of state-of-the-art sentiment identification algorithms to practical problems such as summarizing customer reviews or predicting product sales (Pang and Lee, 2008). For example, Archak et al. (2011) used SA techniques to extract opinions of customers about products and studied the economic impact of the extracted opinions. In another recent study, researchers used sentiment extracted from Twitter posts to predict movie box revenues (Asur and Huberman 2010).

The second research stream aims to design new sentiment algorithms, which can be split into three categories: 1) Bag-of-Words; 2) Rule-based approaches; and 3) Machine learning techniques (Zhang et al. 2011). Bag-of-Words approaches classify sentences or documents as positive or negative based on the occurrences of sentiment phrases (Hu and Liu 2004). Rule-based approaches try to classify sentiments based on the assumption that the meaning of a compound expression (e.g. sentence) is a function of the meaning of its parts (e.g. word or phrase) (Choi and Cardie 2008). Supervised machine learning methods such as Naive Bayes and support vector machines have also been widely used in SA (Wiebet and Bruce 1999; Mullen & Collier 2004). In order to reduce the cost of labeling training data set, researcher also proposed to use unsupervised learning methods for SA (Fang et al. 2014).

Most of the existing SA techniques rely solely on the textual information in sentences and they are not very effective in social media platforms. To address this issue, researchers have recently started to use social network relationships to improve the performance of SA. For instance, Hu et al. (2013) used message-message relationship to improve the performance of SA for twitter content. Tan et al. (2011) used follower network relationship in Twitter to analyze user-level sentiments. Besides relationships in social media platforms, researchers also tried to use some special features of social media platforms to improve

SA accuracy. For example, Hu et al. (2014) incorporated emotional signals in Twitter into the SA algorithm and achieved significant performance improvement.

While following the same goal, our framework is different from previous work in three aspects: (1) The framework design is based on a research theory from the psychology literature and considers social relationships for both internal and external causes of opinions. (2) The framework is not constrained to a particular platform. (3) the proposed framework combines bag-of-words, rule-based and machine learning techniques at different stages of SA.

A Framework for Sentiment Analysis Based on Social Relationships

Major Components in Social Media Platforms

Figure 1. Major Components in Facebook

We first define the terms that will be used in our framework based on the major components in real life social media platforms as follows: A **"*user*"** is a person who subscribes to social media platforms. A **"*comment*"** is a post by a user to express his/her opinion. It is the target of SA. An **"*entity*"** is the container of comments in social media platforms. It can be a web page or a thread in social media. We use entity to represent both threads and web pages because some platforms only have one thread per product (e.g. Amazon.com). Finally, a **"*topic*"** is the high level of category that represents the focused subject of discussion in social media. It describes the content of an entity. For example, the category of a thread can be certain sports news, a promotion, a video game, etc.Figure 1 demonstrates the web page components that relate to these terms in the context of Facebook, a popular social media platform. We note that our proposed terms and definitions are generalizable to different social media platforms that have similar structures, such as Twitter and Amazon.com.

Social Relationships and the Framework

Attribution theory, which was originally studied in social psychology, provides a solid foundation for our framework. It explains a person's behavior using two categories of causes: internal attributions (e.g. attitudes or personality) and external attributions (e.g. other people) (Kelley 1967). Based on the attribution theory, we argue that a person's sentiment towards an entity is influenced by two causes: ***personal causes*** (how does a person like entities of similar topics) and ***environmental causes*** (how do people connected with this person in social networks like the entity). Personal causes are determined by the inherent attitude of a person. For example, personality, knowledge background, past experience may influence one's opinion towards a topic. Environmental causes are people who are interconnected in social networks. Theoretical studies in social networks have found that people who are connected in social networks tend to behave similarly. There are three major causes for the influence of social networks: homophily, induction and confounding. People's expression of emotion is one of the behaviors that are affected by social networks. For example, Fowler and Christakis (2008) studied how happiness spreads in a social network.

We consider two types of social relationships in the framework: ***user-topic relationship*** and ***user-user relationship***. User-topic relationship captures the personal causes in the sense that a user tends to have similar sentiment towards similar entities of the same topic. User-user relationship captures the environmental causes because a user's sentiment may be influenced by people in their social networks. Below are two fundamental assumptions that are derived from attribution theory for SA in social media platforms.

- **Assumption 1 (User-topic relationship assumption):** A user tends to have similar sentiment towards entities of similar topics. For example, if a person makes positive comments on all previous promotions of Walmart, it is likely that he/she will make a positive comment on future promotions of Walmart.
- **Assumption 2 (User-user relationship assumption):** A user's sentiment towards an entity tends to be influenced by people whom the user is connected with in social networks. For example, if social friends of a user give positive comments to promotion of an item, it is likely that he/she will give positive comments to the same promotion.

Based on the two assumptions, Figure 2 describes the framework for SA in social media platforms. We mainly consider two major data sources: text in sentences and relationships in social networks. Textual information in sentences will be first used to identify the initial sentiment based on word sentiment and linguistic rules. We also use clustering techniques to identify topics for the sentences. Next, we build user-topic relationships and user-user relationship based on the social networks in social media platforms. Finally, we combine the initial textual sentiment, user-topic relationship, and user-user relationship and use machine learning based approaches such as support vector machine (SVM) to predict the sentiment.

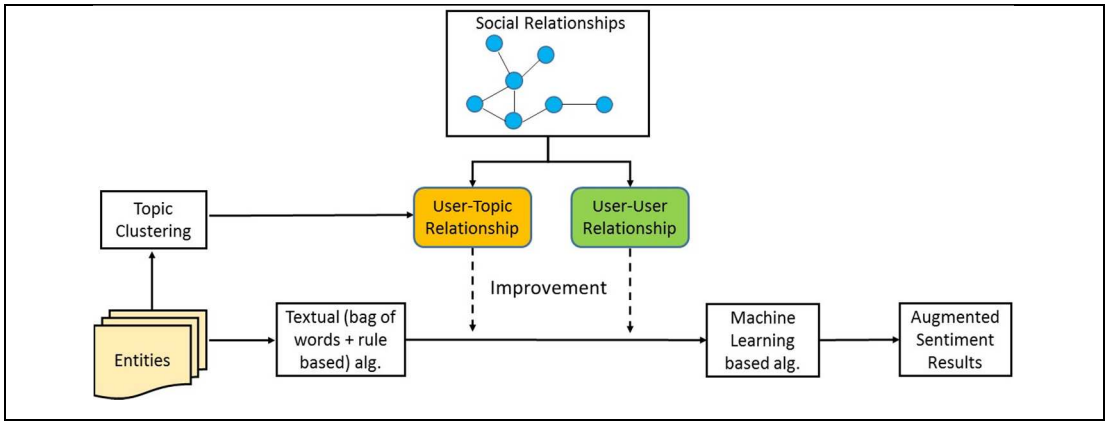

Figure 2. Framework for Sentiment Analysis Based on Social Relationships

Textual Sentiment Analysis

The baseline of our framework is textual sentiment analysis (SA). Based on the literature on SA, we propose to combine bag-of-words and rule-based approaches to produce an initial sentiment label. We use a state-of-art emotional dictionary (Choi et al. 2014) and add some special notations that are frequently used in social media (such as emoticons). We count the number of positive and negative words in a sentence. If a sentence contains more positive words than negative words, it will be labeled positive. If the numbers of positive and negative words are the same, the sentence will be labeled as neutral. Otherwise, it is a negative sentence. For rule-based approach, we define some rules based on linguistic characteristics. One example of rules is that we negate the sentiment of a word following a negation word. Details of the algorithm are discussed in a previous paper (Zhang et al. 2011).

User – Topic Relationships

The user-topic relationship analysis employs a hybrid clustering and item based collaborative filtering approach to improve text based sentiment identification. Our main approach follows a three-step procedure. First, we compute the text based sentiments for all user-entity pairs using the textual sentiment algorithm. Then, we use the standard K-means algorithm to group the entities into a

pre-defined number of topics and aggregate the individual entity scores into the topic scores. Finally, we employ topic-based collaborative filtering to predict the sentiment score of a user on a certain topic. Below, we discuss each step in more detail.

***Individual Sentiment Score Identification**:* As discussed earlier, our eventual goal is to improve text based sentiment identification from a single user by incorporating user similarity knowledge within a social network. To achieve this, we should first compute the individual text based sentiment scores of all entities by each user in the network. For individual scoring, we use the textual sentiment algorithm. Three possible outcomes of the textual sentiment algorithm are "-1" (i.e. negative sentiment), "+1" (i.e. positive sentiment) and, "0" (i.e. neutral sentiment). The outcome of this step is a "user-entity" matrix $A \in \mathbb{R}^{n \times m}$, where n is the number of users in the network and m is the number of entities observed. The element a_{ij} in A represents the sentiment score of the i^{th} user (u_i) on the j^{th} entity (e_j).

***Entity Clustering and Score Aggregation*:** A potential problem with using the basic entities for collaborative filtering is the high sparsity of the "user-entity" matrix, since the number of individual entities (e.g. pages, threads, etc.) in a social media platform may be particularly large. This situation is expected to cause computational performance and scalability issues, as well as reducing the prediction performance. To alleviate these issues, we propose to group low-level entities into higher-level topics, with topic dimensionality being significantly smaller than the entity dimensionality. Specifically, we employ a standard document clustering algorithm – the K-means algorithm, to group individual entities into a smaller number of topics. The end result of the clustering process would be an "entity-topic" matrix $V \in \mathbb{R}^{m \times k}$, where k is the number of topics, defined in advance. In this matrix, the final assignment of an entity onto a topic is represented by *1*. Based on the entity-topic matrix, we compute the sentiment scores of users on topics (i.e. "user-topic" matrix M) as the average scores of entities that belong to each topic.

***Topic-based Collaborative Filtering*:** The "user-topic" matrix, M, enables us to utilize the scores of other similar users in the network for sentiment prediction of any active (i.e. unscored) topic-user pair. To incorporate this network specific knowledge, we employ an item based collaborative filtering approach. This approach works by computing the similarity between the topics and then taking a weighted sum of the topic scores for a given (i.e. active) user, while using the similarity values as weights. For similarity computations, we can employ Euclidean-based similarity measure. For example, given two topics $z_1 = a$ and $z_2 = b$, the similarity between a and b is computed as: $sim_{a,b} = \frac{1}{\sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}}$, where n is the number of users in the network. Once the similarity computations for all possible topic pairs are completed, we can use a weighted sum strategy to approximate the sentiment score of the active user on a given topic. A weighted sum score is formally defined as: $P_{u,z} = \frac{\sum_{j \in Z \setminus z} sim_{z,j} * m_{uj}}{\sum_{j \in Z \setminus z} |sim_{z,j}|}$

***Illustrative Example for User-Topic Relationship*:** To illustrate the implementation of the approach, we use an illustrative example with 4 users (n = 4) and 5 entities (e = 5) and an arbitrary set of individual scores (User – Entity Matrix in Figure 3). We further provide the B matrix for k = 3 and the set of entity – topic assignments (Entity – Topic Matrix in Figure 3). Based on A and B, we compute the sentiment scores of users on topics (User-Topic matrix in Figure 3) as the average scores of entities that belong to each topic.

Let's assume that we'd like to predict the sentiment score of User 3 on Topic 1 – i.e. u = 3 and z = 1. We compute the similarity scores between Topic 1 and the rest of the topics as: $sim_{1,2} = \frac{1}{\sqrt{(-0.5+1)^2+(-1-1)^2}} = 0.48$ and $sim_{1,3} = \frac{1}{\sqrt{(-0.5-1)^2+(-1-0)^2}} = 0.55$. Based on the similarity scores, the final sentiment is computed as: $P_{3,1} = \frac{0.48*(-1)+0.55*0.5}{0.48+0.55} = -0.2$.

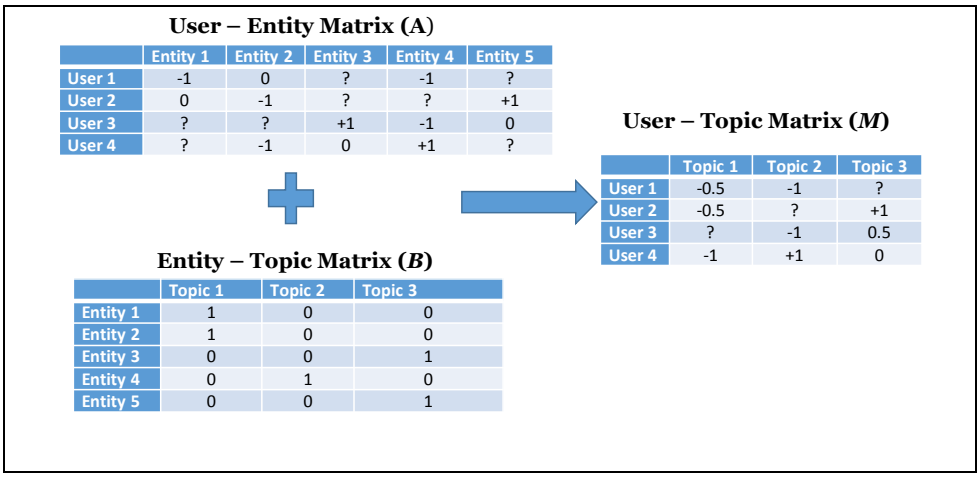

Figure 3. Example of User-Topic Relationships

User-User Relationships

In addition to user-topic relationships, user-user relationships also play a very important role in SA in social media platforms. While it is relatively easy to identify who wrote which comments (e.g., user-topic relationship) in social media platforms, the social relationship between users is not always readily available. For example, in Facebook, we usually do not know whether two people are friends because of the privacy policy of Facebook. By examining some popular social media platforms, we identify two types of social relationships: *direct* and *indirect* among users that allow us to capture indications of connections in social media.

***Direct Social Relationship*:** Users are explicitly connected in social media platforms. Direct social relationships are personal relationships generated by the users themselves. For example, the friend relationship is a direct social relationship because the friend list of a user is the explicit representation of friendship. Such direct social relationships can further be categorized into one of the two types: symmetric or asymmetric. Symmetric relationships are those, in which there exists a dual way undirected link between two users. For example, in order to become friends in a social media platform such as Facebook, one person needs to apply to the other person and get approved. On the other hand, asymmetric relationships are those, in which a one way directed link would suffice to establish the relationship. Examples of asymmetric direct relationships are follow or @ relationships in Twitter.

***Indirect Social Relationship*:** Users are connected with each other through one or many common intermediaries. For example, people who come from the same location are considered as connected in a social network. Such social relationship is built on the common location intermediary. Other examples could be common age, common preference, etc.

The easiest way to predict one person's sentiment based on social relationships is to use the average sentiment of people who are connected with this person. For a social network with N users, we use a matrix $U \in \mathbb{R}^{N\times N}$ to represent the social networks among the N users. $u_{ij}=1$ when the i^{th} user and the j^{th} user have a (direct or indirect) relationship in social networks. Assume that S_i is the user-user relationship sentiment score of user i. $S_i = \frac{\sum_j S_j * u_{ij}}{\sum_j u_{ij}}$.

***Illustrative Example for User-User Relationships*:** Assume that we know the sentiment of Users 2, 3, 4, and 5 towards Entity 1 and User 1 is connected to Users 2, 3, and 4. The user-user relationship is shown in the network diagram in Figure 4. The predicted sentiment of User 1 towards the entity is the average sentiment of Entity 1 from Users 2, 3, and 4, which is $\frac{(-1)+(-1)+(0)}{3}$=-0.667.

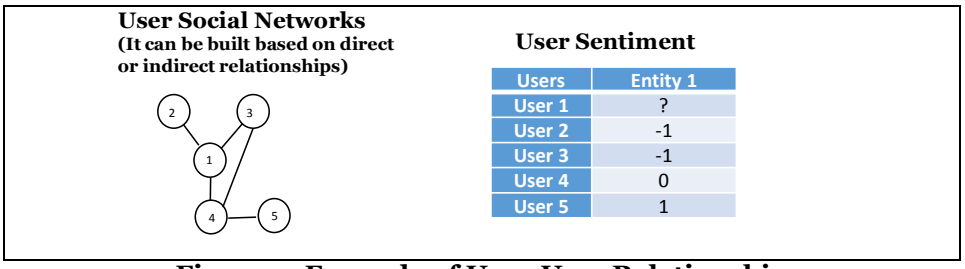

Figure 4. Example of User-User Relationship

On-going Experiment

To demonstrate the applicability of the framework, we have applied it in the context of Facebook social media platform. We used Facebook Graph API to download posts, users' comments, users' likes, and corresponding users' profiles (publicly available information, such as gender and locale) for each public Facebook page during the past three years. We randomly selected 10,000 user comments as our experimental dataset and manually label their sentiments. Then, we implemented the framework components and compared sentiment identification accuracy with the textual SA approach. Since this is an on-going experiment, we limit the framework implementation based on the unique features of Facebook. For topical categories, we adopted the natural Facebook defined categories as topics without running the topic clustering process. In addition, we only used indirect relationships in the experiment as Facebook does not allow us to collect data on user friendship. Specifically, we consider two indirect relationships: same gender and same thread.

Gender has been identified as a factor that has profound impact on the formation and outcome of social networks (Fuhrer and Stansfeld 2002). For example, researchers have found that women tend to be more positive in social network communication than men (Thelwall et al. 2009). In our dataset, we find consistent emotional differences between men and women (Table 1). For different topics, the difference between men and women is even more complicated. Thus, we use gender and thread as two factors to build the indirect social relationships among users.

Table 1. Gender Differences on Emotion

Web Page (Category)	Gender	Positive ratio
Barack Obama (Politics)	M	0.61
	F	0.69
Chicago bulls (Sports)	M	0.68
	F	0.79
DKNY (Fashion)	M	0.94
	F	0.96

To see how each component increases the accuracy, we add one component at a time into our model. In this experiment, we pick Support Vector Machines as our machine learning based method to combine textual information, user-topic relationship and user-user relationship. For evaluation purposes, we use 10-fold cross validation. Table 2 demonstrates the results from the on-going experiment. We see that adding more social relationship information into textual SA approach is helpful to improve the accuracy. It is worthy to mention that our framework advance the accuracy of SA significantly. The accuracy of SA in most of the recent work is less than 85% and the accuracy of our approach is 89% (Tan et al. 2011; Hu et al. 2013).

Table 2. Accuracy of Different SA Approaches

Approach	Accuracy
Textual SA approach (Baseline)	0.834
Baseline + User-Topic Relationship	0.861
Baseline + User-User Relationship	0.857
Baseline + all above	0.883

Conclusion

In this paper, we propose a novel framework to address challenges of SA in social media platforms. We draw from the attribution theory and incorporate two types of social relationships into the sentiment identification process. We conduct experiments on a data set collected from Facebook to show that both types of social relationships help improve the accuracy of SA. Our approach can be considered as a big data framework (Fan et al. 2015) in the sense that it combines multiple perspectives of data (textual, user-topic relationship and user-user relationship) to address the SA identification problem. This paper is part of a work-in-progress research. Our on-going efforts aim to improve the framework by considering the interaction between the two types of social relationships. Further, we plan to expand the scope of our experiments and analyses by applying the framework on datasets collected from other social media platforms with different social relationship characteristics.

References

Archak, N., Ghose, A., & Ipeirotis, P. G. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, (57:8), pp. 1485-1509.

Asur, S., & Huberman, B. 2010, August. Predicting the future with social media. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Toronto, Canada, pp.492-499.

Choi, Y., & Cardie, C. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* Waikiki, Hawaii, USA, pp793-801.

Choi, Y., Deng, L., & Wiebe, J. 2014. Lexical acquisition for opinion inference: A sense-level lexicon of benefactive and malefactive events. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), Baltimore, USA,* pp. 107-112.

Fan, S., Lau, R. Y., & Zhao, J. L. 2015. Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix. Big Data Research, (2:1), pp.28-32.

Fang, F., Dutta, K., & Datta, A. 2014. Domain Adaptation for Sentiment Classification in Light of Multiple Sources. *INFORMS Journal on Computing*, (*26*:3), pp.586-598.

Fowler, J. H., & Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *British Medical Journal*, (*337*: a2338), pp. 1-9.

Fuhrer, R., & Stansfeld, S. A. 2002. How gender affects patterns of social relations and their impact on health: a comparison of one or multiple sources of support from "close persons". Social science & medicine, (54:5), pp.811-825.

Hu, M., & Liu, B. 2004. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, Washington, USA, pp. 168-177.

Hu, X., Tang, L., Tang, J., & Liu, H. 2013. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining, Rome, Italy,* pp. 537-546.

Hu, X., Tang, J., Gao, H., & Liu, H. 2013. Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd international conference on World Wide Web, Rio de Janeiro, Brazil, pp. 607-618.

Kelley, H. H. 1967. Attribution theory in social psychology. In *Nebraska symposium on motivation*. University of Nebraska Press.

Mullen, T., & Collier, N. 2004. Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, pp.412-418.*

Pang, B., & Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, (*2*:1-2), pp.1-135.

Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., & Li, P. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, CA, USA,* pp. 1397-1405.

Thelwall, M., Wilkinson, D., & Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in MySpace. Journal of the American Society for Information Science and Technology, (61:1), pp.190-199.

Wiebet J., Bruce R. 1999. Development and use of a gold standard data set for subjectivity classifications. ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, MD, USA, pp.264–253.

Zhang, K., Cheng, Y., Xie, Y., Honbo, D., Agrawal, A., Palsetia, D. & Choudhary, A. 2011. SES: Sentiment elicitation system for social media data. 11th International Conference on Data Mining Workshops (ICDMW), Vancouver, Canada, pp. 129-136.