Detecting Chronic Critics Based on Sentiment Polarity and User's Behavior in Social Media

Sho Takase[†] **Akiko Murakami**[‡] **Miki Enoki**[‡] **Naoaki Okazaki**[†] **Kentaro Inui**[†]
Tohoku University[†] IBM Research - Tokyo[‡]
{takase, okazaki, inui}@ecei.tohoku.ac.jp
{akikom, enomiki}@jp.ibm.com

Abstract

There are some chronic critics who always complain about the entity in social media. We are working to automatically detect these chronic critics to prevent the spread of bad rumors about the reputation of the entity. In social media, most comments are informal, and, there are sarcastic and incomplete contexts. This means that it is difficult for current NLP technology such as opinion mining to recognize the complaints. As an alternative approach for social media, we can assume that users who share the same opinions will link to each other. Thus, we propose a method that combines opinion mining with graph analysis for the connections between users to identify the chronic critics. Our experimental results show that the proposed method outperforms analysis based only on opinion mining techniques.

1 Introduction

On a social media website, there may be millions of users and large numbers of comments. The comments in social media are related to the real world in such fields as marketing and politics. Analyzing comments in social media has been shown to be effective in predicting the behaviors of stock markets and of voters in elections (Bollen et al., 2011; Tumasjan et al., 2010; O'Connor et al., 2010). Because of their effects on the real world, some complaints may harm the reputation of a corporation or an individual and cause serious damage. Consider a comment such as "Working for *Company A* is really awful" as an example. The complaint gives viewers a negative impression of *Company A* and can increase the number of people who think the company is bad.

Some complaints are expressed by a specific user who is always criticizing a specific target entity (in this example, *Company A*). We call this user a *chronic critic* of that entity, a person who is deliberately trying to harm the reputation of the entity. That is, a chronic critic is trying to run a negative campaign against the entity. If the entity is aware of its own chronic critics, then it is able to take prompt action to stop the malicious complaints. When the complaints are false, the entity can use that defense. In contrast, if the chronic critics are justified, then the entity should address the concerns to limit the damage. Hence, to handle malicious rumors, it is important to detect the chronic critics.

However, it is generally quite difficult for a computer to detect a chronic critic's comments, since especially the comments in social media are often quite informal. In addition, there are complexities such as sarcasm and incomplete contexts. For example, if *Company A* has been involved in a widely recognized fiasco, then some chronic critics might sarcastically write "good job" or "wonderful" about *Company A*. They are using positive words, but in the context they are effectively criticizing *Company A*. Some chronic critics bash a target entity solely with sarcasm, so they damage the target with positive words. It is exceedingly difficult to directly detect these chronic critics based on their comments. In an example of an incomplete context, if one author starts an exchange with a comment such as "The new product from *Company A* is difficult to use" and another user responds with something like "Fool", we cannot easily recognize the meaning of this comment as related to "*Company A* being foolish because the product really is difficult to use" or whether "the user is the fool because the product is easy for other people to use". To find chronic critics for a given entity, we need to identify the actual target of the complaints. Take the comment "*Company B* is much worse than *Company A*" for

Proceedings of the ACL Student Research Workshop, pages 110–116,
Sofia, Bulgaria, August 4-9 2013.

example. This comment is probably complaining about *Company B* but not *Company A*. In contrast, most of the previous work on sentiment analysis in social media does not consider these kinds of problems (Barbosa and Feng, 2010; Davidov et al., 2010; Speriosu et al., 2011).

Switching to the behavior of each user, in social media we often see that users who have similar ideas will tend to cooperate with each other. In fact, previous work suggests that users who have the same opinions tend to create links to each other (Conover et al., 2011b; Yang et al., 2012). Because chronic critics share the purpose of attacking some target's reputation, they may also decide to cooperate. For this reason, to detect chronic critics, we believe that information about the connections among users will be effective.

In this paper, we present a method that combines opinion mining based on NLP and graph analysis of the connections among users to recognize the chronic critics. In the experiments, we demonstrate the difficulty in detecting chronic critics by analyzing only the individual comments. In addition, we investigate the effectiveness of using the connections between users, i.e., using the proposed method. For our experiments, we used Twitter, a popular social media service. In particular, we focus on Japanese comments on Twitter.

This paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed method which applies the opinion mining and graph analysis. Section 4 demonstrates the effectiveness of the proposed method and discusses the experimental results. Section 5 concludes this paper.

2 Related Work

In recent years, an interest in opinion mining in online communities has emerged (Conover et al., 2011a; O'Connor et al., 2010; Speriosu et al., 2011; Murakami and Raymond, 2010; Barbosa and Feng, 2010; Davidov et al., 2010). O'Connor et al. (2010), Barbosa and Feng (2010), Davidov et al. (2010), and Speriosu et al. (2011) proposed methods to predict a sentiment polarity (i.e., positive or negative) of a comment in social media. O'Connor et al. (2010) studied a subjectivity lexicon. Barbosa and Feng (2010) and Davidov et al. (2010) used machine learning approaches. Speriosu et al. (2011) introduced connections between words, emoticons, tags, n-grams, comments and users. These studies did not identify the target of the polarized sentiment of each comment.

Conover et al. (2011a) proposed a method that predicts the political polarity of a social media user based on the connections between users and tags. They demonstrated that label propagation on the graph representing the connections between users is effective. However, this method is not guaranteed to obtain the optimal solution. In contrast, our research uses graph analysis that converges on the optimal solution.

Murakami and Raymond (2010) proposed a method that uses the connections between users to predict each user's opinion, i.e., support or oppose a topic in online debates. They analyzed the content of the discussions to infer the connections. However, in social media, it is difficult to infer connections based on content because of such complexities as incomplete contexts. To address these problem, we analyzed the behavior of the users to predict the connections between users.

Our task is similar to spammer detection (Wang, 2010; Yang et al., 2012). Wang (2010) proposed a method using a classifier to detect spammers. They used the content in the comments and the number of linked users as features. Yang et al. (2012) analyzed spammer communities and demonstrated that spammers closely link to each other in social media. They also proposed a method that extracts spammers using the connections between users. While Wang (2010) and Yang et al. (2012) required manually annotated data for training or as seeds, we extract the seeds for the graph analysis automatically through opinion mining.

3 Proposed Method

Figure 1 presents an overview of the proposed method. The proposed method has two phases, opinion mining and graph analysis. First, we extract a few chronic critics by analyzing the opinions of many users referencing the target entity. For the opinion mining, we are initially looking for users who strongly criticize the target entity. In Figure 1, given *Company A* as a target entity, we find users "b" and "e" since they said "Working for *Company A* is really awful" and "This product from *Company A* is useless". However, we may miss the other chronic critics since they used sarcasm and incomplete contexts.

Next, we find the users who are linked to the

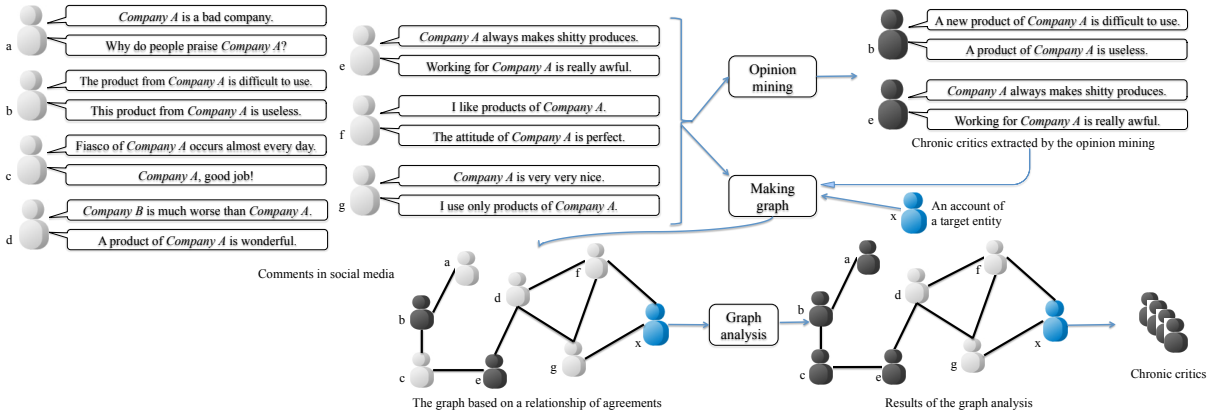


Figure 1: Overview of proposed method

chronic critics that were detected through opinion mining. We built a graph in which the users are represented by nodes and the links between the users are represented by edges.We recognize additional chronic critics based on the graph analysis. In the example of Figure 1, we find more chronic critics not recognized by the opinion mining, such as "a" and "c", because they are linked to the chronic critics "b" and "e". In this section, we explain the opinion mining and graph analysis. Since a comment in Twitter is called a *tweet*, we use the term tweet below.

3.1 Opinion Mining

As defined in Section 1, we defined a user who frequently criticizes a target entity as a chronic critic. Therefore, we classify the tweets of each user into critical or non-critical and label any users who complain about the target entity many times as chronic critics. Because we want to investigate the opinions of each user in public, we analyze public tweets, excluding the private conversations between users. In Twitter, this means we ignore a *reply* that is a response to a specific user named *username* (written in the format "@*username* response") and *QT* that is a mention in a quoted tweet from *username* (written in the format "mention RT @*username*: quoted tweet").

We assume a phrase representing negative polarity or profanity to be critical phrases. The proposed method determines whether a tweet complains about the target entity by investigating a critical phrase and the target of the phrase.

Note that a negative polarity is represented by declinable words or substantives. We used the sentiment analyzer created by Kanayama and Nasukawa (2012) to detect a phrase representing negative polarity by using declinable words. We used the lexicon collected by Higashiyama et al. (2008) to find negative polarity in substantives. For detecting profanity, we use a profane lexicon collected by Ogino et al. (2012).

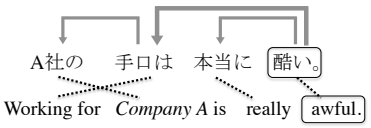


Figure 2: Example of critic tweet

The sentiment analyzer can find not only sentiment phrases but the targets of the phrases based on syntactic parsing and the case frames[1]. However, because there are many informal tweets and because most users omit the grammatical case in tweets, the sentiment analyzer often fails to capture any target. To address this problem, in addition to a target extracted by the sentiment analyzer, we obtain a target based on the dependency tree. We extract nouns in parent and child phrases within distance 2 from a critical phrase in the dependency tree.

Figure 2 shows an example of a Japanese tweet criticizing *Company A* and its English translation. The Japanese tweet is split into phrase-like units (*bunsetsu*). Each English phrase is linked to the corresponding bunsetsu by a dotted line. The dependency relationships among the bunsetsu are expressed by the arrows. In the tweet, the black-edged phrase "awful" is a critical phrase. We extract the nouns in "Working for" and "*Company A* is" as targets of the critical phrase since these

[1] A case frame is a list which represents grammatical cases of a predicate.

phrases are parents within distance 2 of the critical phrase. Therefore, we decide that the tweet is criticizing *Company A*.

Since a chronic critic frequently complains about the target entity, we can predict that most of the tweets written by a chronic critic of the target entity will be critical tweets. Therefore, we can calculate a ratio of critical tweets for all of the tweets about the target entity. We score the user u_i with equation (1).

$$score_i = \frac{n_i}{N_i} \tag{1}$$

N_i is the number of all tweets about the target entity and n_i is the number of critical tweets about the entity by that user [2]. We extract the top M users based on $score_i$ as chronic critics.

3.2 Graph Analysis

In social media, it is often very difficult to determine whether a tweet is critical since many tweets include sarcasm or incomplete contexts. The opinion mining may miss numerous complaints with sarcasm or incomplete contexts. To resolve this problem, we apply user behaviors. In social media, we assume that users having the same opinion interact with each other in order to demonstrate the correctness of their opinion. In particular, since the purpose of chronic critics is to spread the bad reputation, we assume that they want to assist each other. We supplement the opinion mining by a graph analysis using this assumption. Thus, we make a graph representing connections among the users and use label propagation on the graph based on the results of the opinion mining as the seeds.

In addition, we believe that a user will try to spread user matching opinions. This implies that a user who spreads the opinion of another of agrees with the author of that opinion. In Twitter, a user can spread an opinion as an ***RT***, which is a reposting of a tweet by a *username* (written in the format "RT @*username*: tweet"). Conover et al. (2011b) demonstrated that they can make a graph representing the connections among users who support each others opinions by using RTs. Hence, an RT expresses a relationship of endorsement. We also created a graph based on this feature.

Our graph has m users ($U = \{u_1, ..., u_m\}$) as nodes, where u_i connects with u_j via an edge that has weight w_{ij} ($0 \leq w_{ij} \leq 1$) and w_{ij} corresponds to the degree to which u_i supports u_j. We calculate w_{ij} by using Equation (2).

$$w_{ij} = \frac{1}{2}\left(\frac{r_{ij}}{R_i} + \frac{r_{ji}}{R_j}\right) \tag{2}$$

r_{ij} is the total RT tweets of u_j by u_i and R_i is the number of RTs by u_i. Therefore, the more u_i and u_j RT each other, the more weight w_{ij} is close to 1. In contrast, if u_i and u_j rarely RT each other, the value of w_{ij} will approach 0. In addition, this w_{ij} definition is symmetric means (i.e., $w_{ij} = w_{ji}$).

We find more new chronic critics by label propagation on the graph. We use the chronic critics obtained by the opinion mining as seeds. It is assumed that a user who supports the target entity is not a chronic critic. Using this knowledge, we use the account of the target entity as a seed.

The label propagation assigns a confidence score $\boldsymbol{c} = (c_1, ..., c_m)$ to each node $U = u_1, ..., u_m$, where the score is a real number between -1 and 1. A score close to 1 indicates that we are very confident that the node (user) is a chronic critic. A score close to -1 indicates that we are sure that the node is not a chronic critic. In addition, the scores of seeds are fixed and cannot be changed. The scores of chronic critics obtained by the opinion mining are 1 and the score of the target entity is set to -1. To formulate the label propagation as an optimization problem, we used the loss function proposed by Zhu et al. (2003), because $w_{ij} \geq 0$ for all i, j.

$$E(\boldsymbol{c}) = \frac{1}{2}\sum_{i,j} w_{ij}(c_i - c_j)^2 \tag{3}$$

To minimize $E(\boldsymbol{c})$, c_i is close to c_j when w_{ij} is greater than 0. That is, if the users support each other, the scores of the users are close to each other. Thus, by minimizing $E(\boldsymbol{c})$, we assign the confidence scores considering the results of the opinion mining and agreement relationships among the users. We find the users that have scores greater than the threshold.

We believe that if the distance between users on the graph is large, then users slightly support each other. However, we can assign a score of 1 to each node in any subgraph that has chronic critics extracted by the opinion mining to minimize $E(\boldsymbol{c})$ if the subgraph does not include the account of the target entity, no matter how far away a node

[2] The formula (1) assigns a high score to a user if the user only produces one or two tweets about the target entity and those tweets are negative. To prevent this, we disregard the users whose the number of tweets are fewer than 5.

Table 1: Properties of the experimental datasets

Target entity	Tweets	Critics	Kappa
Company A	35,807	112	0.81
Politician A	45,378	254	1.0

is from the seeds. To avoid this problem, Yin and Tan (2011) introduced a *neutral fact*, which decreases each confidence score by considering the distance from the seeds. The neutral fact has a fixed confidence score 0 and connects with all of the nodes except the seeds. Suppose u_1 is the neutral fact, $U_l = \{u_2, ..., u_l\}$ is the set of seeds and $U_t = \{u_{l+1}, ..., u_m\}$ is the set of all nodes except seeds. To assign the weight of the edge between u_1 and other nodes considering the degrees of the nodes, we calculate the weight by as:

$$w_{1i} = \begin{cases} 0 & i = 1, ..., l \\ \mu \sum_{j>1} |w_{ij}| & i = l+1, ..., m \end{cases} \quad (4)$$

where μ is a small constant. Thus, the weight is proportional to the total weight of the edges from each node.

4 Experiments

4.1 Experimental Setting

For our experiment, we gathered tweets by using the *Twitter search API*. The twitter search API returns the tweets that contain an input query. We used the name of a target entity, words related to the entity[3], and the account name of the entity as queries. In this research, there were two target entities, *Company A* and *Politician A*. We found many critical tweets about these target entities. The entities have their own accounts in Twitter. We collected the Japanese tweets for one month. We want to extract the users who frequently express a public opinion related to a target entity. For this reason, we eliminated users whose the number of tweets except conversation (i.e., reply, QT, RT) are fewer than 5. In addition, to eliminate bots that automatically post specific tweets, we eliminated users whose conversational tweets were fewer than 2. We selected some of the remaining users for the experiment. To satisfy our definition, a chronic critic must tweet about the target entity many times. Therefore, we focused on the top 300 users based on the number of tweets as our experimental users. Table 1 shows the total numbers of tweets by the top 300 users, excluding the account of the target entity.

We created an evaluation set by manually dividing the experimental users into chronic critics and regular users. A chronic critic actively complained and tried to harm the reputation of the target entity. We also regarded a user who frequently reposted a critic's tweets and unfavorable news about the target entity as a chronic critic. For the experimental users tweeting about *Company A*, we asked two human annotators to judge whether a user was a chronic critic based on one month of tweets. The Cohen's kappa value was 0.81 which inter-annotator agreement was good. We selected the arbitrarily annotating by one of the annotators as our evaluation set. Table 1 expresses the number of chronic critics for each target entity in the evaluation set. For the experimental users tweeting about *Politician A*, we randomly extracted 50 users randomly to calculate Cohen's kappa, which is displayed in Table 1.

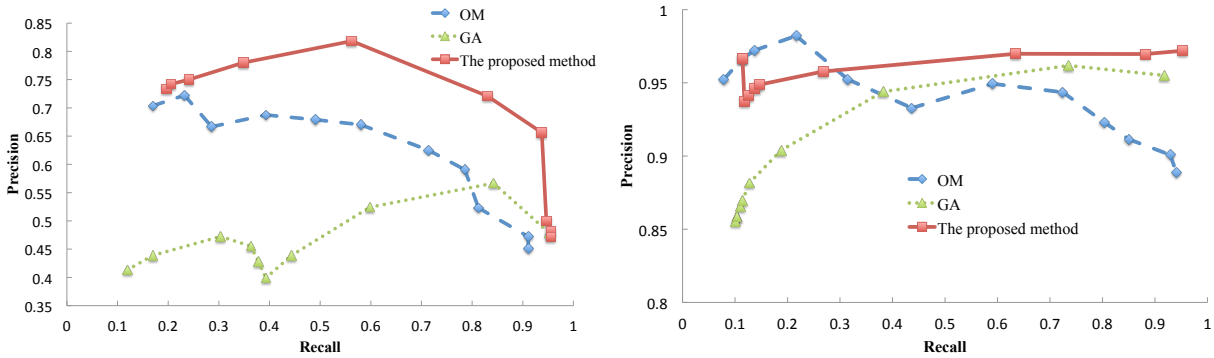
We evaluated the effects of combining the opinion mining with the graph analysis. We compared opinion mining (OM), graph analysis (GA), and the combination of opinion mining and graph analysis (our proposed method). GA randomly selected M users from experimental users as seeds and takes the average of the results obtained by performing label propagation three times. The number of chronic critics extracted by the opinion mining (i.e., the valuable M) was set to 30. The parameter μ, that we use to calculate the weight of the edges connected to neutral fact, was set to 0.1.

4.2 Results

Figure 3 represents the precision and recall of each method for each target entity. In OM, we varied the threshold from 0 to 0.2 in increments of 0.02 and accepted a user with a score over the threshold as a chronic critic. In GA, we varied the threshold from 0.35 to 0.8 in increments of 0.05.

In Figure 3, the results for *Company A* and *Politician A* are quite different, though there are some similar characteristics. Figure 3 shows that OM achieved high precision but it was difficult to improve the recall. In contrast, GA easily achieved high recall. The proposed method achieved high precision similar to OM and high recall. In other words, the proposed method found many

[3]We manually prepared the words that have a correlation with the entity. In this paper, we only used the name of the political party of *Politician A* as the related word.



(a) *Company A* (b) *Politician A*

Figure 3: Precision and recall of each method for each target entity

Table 2: Users connected with the target entity

Target entity	Users	Non-critics
Company A	45	39
Politician A	74	35

chronic critics while retaining high precision of OM. Therefore, the combination of the opinion mining and the graph analysis improved the performance of recognizing the chronic critics.

Figure 3 shows that the recall of OM was low, which means that OM missed some of the critical tweets. In this paper, we used domain-independent lexicons to detect the critical phrases. Therefore, OM failed to find domain-dependent critic phrases such as slang words. In addition, some chronic critics do not express criticism clearly in their own tweets. To spread the bad reputation, they reference only a title and link to a webpage that criticizes the target entity such as:

> This shows the reality of *Company A*. Why do you buy products from this company? http://xxx

We believe that is often done because each tweet is limited to 140 characters. It is difficult to classify the tweet as a complaint based only on its content. However, the proposed method recognized most chronic critics that complain with these methods based on the GA.

It cannot reasonably be assumed that a user who supports the account of the target entity is a chronic critic. For this reason, in the graph analysis, we used the entity's account to recognize non-critics. We believe that using the account corrects for mistakes in selecting the seed chronic critics. Table 2 shows the number of users connected with the account. Table 2 also shows the number of non-critics among the users. As seen in Table 2, many non-critics were connected with the account. Especially for *Politician A*, most of the non-critics in the evaluation set were connected with the account. Therefore, incorporating the account into the graph analysis can correct for errors in the seeding of chronic critics. However, some chronic critics were connected with the target's account and reposted tweets from the account. We noticed that they mentioned their negative opinions about the content of such a tweet immediately after reposting that tweet. Hence, we need to analyze the contexts before and after each RT.

For *Politician A*, Table 1 shows that most of the users in the evaluation set criticized the politician. We were able to find most of the chronic critics by extracting the users linked to each other. However, for *Company A*, the precision of GA was low. This means we need high accuracy in selecting the seeds to correctly capture chronic critics. Because we used the users extracted by the opinion mining as the seeds, the proposed method outperformed OM and GA.

5 Conclusion

In this paper, we proposed a method that uses not only opinion mining but graph analysis of the connections between users to detect chronic critics. In our experiments, we found that the proposed method outperformed each technique.

In our study, we used two entities. To improve reliability, we should study more entities. We used a relationship between users that support each other. However, we suspect that the relationship includes adversaries. We hope to address these topics in the future.

Acknowledgments

This research was partly supported by JSPS KAKENHI Grant Numbers 23240018. The authors would like to acknowledge Hiroshi Kanayama and Shiho Ogino in IBM Research-Tokyo for providing their tools for our experiments.

References

Luciano Barbosa and Junlan Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 36–44.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Michael D. Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011a. Predicting the Political Alignment of Twitter Users. In *Proceedings of the 3rd IEEE Conference on Social Computing*, pages 192–199.

Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011b. Political Polarization on Twitter. In *Proceeding of the 5th International AAAI Conference on Weblogs and Social Media*, pages 89–96.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 241–249.

Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto. 2008. Learning Polarity of Nouns by Selectional Preferences of Predicates (in Japanese). In *Proceedings of the 14th Annual Meeting of The Association for Natural Language Processing*, pages 584–587.

Hiroshi Kanayama and Tetsuya Nasukawa. 2012. Unsupervised Lexicon Induction for Clause-Level Detection of Evaluations. *Natural Language Engineering*, 18(1):83–107.

Akiko Murakami and Rudy Raymond. 2010. Support or Oppose? Classifying Positions in Online Debates from Reply Activities and Opinion Expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 869–875, Beijing, China.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the 4h International AAAI Conference on Weblogs and Social Media*, pages 122–129.

Shiho Ogino, Tetsuya Nasukawa, Hiroshi Kanayama, and Miki Enoki. 2012. Knowledge Discovery Using Swearwords (in Japanese). In *Proceedings of the 8th Annual Meeting of The Association for Natural Language Processing*, pages 58–61.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph. In *Proceedings of the 1st workshop on Unsupervised Learning in NLP*, pages 53–63.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 178–185.

Alex Hai Wang. 2010. Don't Follow Me - Spam Detection in Twitter. In *Proceedings of the 5th International Conference on Security and Cryptography*, pages 142–151.

Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. 2012. Analyzing Spammers' Social Networks for Fun and Profit: A Case Study of Cyber Criminal Ecosystem on Twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 71–80.

Xiaoxin Yin and Wenzhao Tan. 2011. Semi-Supervised Truth Discovery. In *Proceedings of the 20th international conference on World Wide Web*, pages 217–226.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. 2003. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *Proceedings of the 20th International Conference on Machine Learning*, pages 912–919.