Identification of Flaming and Its Applications in CGM
Case Studies toward Ultimate Prevention

Yuki Iwasaki, Ryohei Orihara, Yuichi Sei,
Hiroyuki Nakagawa, Yasuyuki Tahara and Akihiko Ohsuga
Graduate School of Information Systems, University of Electro-Communications, Chofu-city, Tokyo, Japan

Keywords: Flaming, Microblogging, Reputation Mining, Topic Extraction, Sentiment Analysis.

Abstract: Nowadays, anybody can easily express their opinion publicly through Consumer Generated Media. Because of this, a phenomenon of flooding criticism on the Internet, called flaming, frequently occurs. Although there are strong demands for *flaming management*, a service to reduce damage caused by a flaming after one occurs, it is very difficult to properly do so in practice. We are trying to keep the flaming from happening. Concretely, we propose methods to identify a potential tweet which will be a likely candidate of a flaming on Twitter, considering public opinion among twitter users. We divide flamings into three categories: *criminal episodes*, *struggles between conflicting values* and *secret exposures*. The first two represent the vast majority of flaming cases. As for the CEs, a Naïve Bayes-based method has been promising to identify the cases. As for the SBCVs, we propose *a dynamic P/N analysis* based on *daily polarity*, which represents the strength of the polarity of public opinion on a given topic. An experiment using a past flaming case has shown that the method has successfully explained the case as one caused by a gap between the polarity of the tweet and that of public opinion.

1 INTRODUCTION

In recent years, thanks to the spread of Consumer Generated Media, anybody can easily express their opinion publicly. Because of this, a phenomenon called flaming frequently occurs on the Internet. There is an increasing risk of suffering damage for not only a celebrity, but also ordinary people. A flaming is defined as a situation where a remark gets a flood of critical comments against it (Tashiro, 2008). Existing research has a limitation that a flaming has to be prevented by hand essentially. In order to detect the flaming that has already occurred, it is sufficient to find a situation where critical comments are flooded (NAVER, 2010). However, it is not possible to prevent the flaming by those approaches.

The ultimate goal of this study was to prevent the flaming. To achieve that, it is necessary to predict the future flaming. One possible approach is to identify potential situations and remarks which will be likely candidates of flamings by means of machine learning.

The rest of this paper is organized as follows: Section 2 presents identification of flamings based on flaming keywords and its experiment. Section 3 describes three typical flaming patterns. In section 4, we propose our method to identify flamings. Section 5 proposes a system that aims to visualize the likelihood of flaming. Section 6 evaluates P/N classifiers used in our method. Section 7 summarizes related work. Finally, section 8 concludes the paper with directions for future research.

2 FLAMING KEYWORDS

An approach to identify flaming is employing flaming keywords: words that will likely cause flamings. Here we are examining a claim: **it is possible to extract flaming keywords from flaming recidivists' remarks**.

2.1 Experiment regarding the Claim

We conduct preliminary experiments to verify the claim. We extract flaming keywords from recent 3,000 tweets by three politicians who have more

Iwasaki Y., Orihara R., Sei Y., Nakagawa H., Tahara Y. and Ohsuga A..
Identification of Flaming and Its Applications in CGM - Case Studies toward Ultimate Prevention.
DOI: 10.5220/0004916606390644
In *Proceedings of the 6th International Conference on Agents and Artificial Intelligence* (ICAART-2014), pages 639-644
ISBN: 978-989-758-015-4

than two flaming records, based on the assumption that it must be easier to recognize the characteristic of the flaming keywords when the range of topics is limited. An example of flaming-causing remarks by them is "You are less than a cockroach." by Mayor Hiwatashi of Takeo, Saga (Hiwatashi, 2013).

2.2 Results and Discussion

In order to evaluate the effectiveness of the method, we compare frequent words in the politicians' remarks with the words extracted by Ishino's method (Ishino *et al.*, 2012). The method makes use of the difference between the word frequencies of two corpora. One is a *target* corpus, from which we wish to extract keywords. The other is a *reference* corpus, which is a superset of the target and provides baseline frequencies of words. We pick 50 most frequent words from each set and compare them each other by their average *webidf* value. Its small value indicates that the word is a common one. Therefore, the generality of the extracted words can be evaluated through this. As a result, the method has successfully removed common words, shown by the fact that webidf increased by 34% (Table 1). Words intended to mean criticism, such as *stupid*, *foolish*, *unreasonable*, *meaningless* are extracted as flaming keywords. We have tried to extract Mayor Hiwatashi's flaming cases from all of his remarks using the keywords. However, only 20% of the flaming cases are recalled. Namely, a flaming can be caused by remarks without these violent words. The extracted words are poor as flaming keywords.

Table 1: Average webidf of 50 frequent words.

Flaming politicians	Frequent words	After application of (Ishino *et al.*)	Rate of increase
Kawakami	1.93	2.38	0.23
Niwayama	1.92	2.23	0.38
Hiwatashi	1.85	2.33	0.41

The result can be explained by the fact that we have indiscriminately treated all the flaming cases as the target corpus. The method by Ishino *et al*. relies on uniformity of the characteristic of the target corpus. The failure of the experiment could be caused by lack of the condition. In the next section we will try to classify the cases into categories.

3 FLAMING CATEGORIES

A literature reports that flamings can be classified into categories (Kobayashi, 2011). In order to verify the claim, we have actually classified 100 flaming cases by hand. As a result, most of them are classified into the following three categories: *criminal episodes (CEs)*, *struggles between conflicting values (SBCVs)* and *secret exposures (SEs)* (Table 2). A CE is a remark that makes one's own criminal behavior public. Most of them are caused by ordinary people. In those flamings striking words indicate crimes such as *unlicensed*, *drunk driving*, *shoplifting* and *planted a bomb* are commonly seen. A SBCV is a remark that forces one's own opinion about a topic on others. Most of them are caused by celebrities. It is easy to cause this type of flaming if there are many people have opinions differ from speaker's one. Let us remark that this study does not cover flames as expert topic conflicts (e.g. Windows vs. Linux). A SE is a remark that makes celebrities' or organization's privates public. They can be caused by either ordinary people or celebrities.

Based on the observation, we put the following assumption: **it is possible to classify flamings into three categories**. Furthermore, we propose identification methods for two categories represent the vast majority of flaming cases: CEs and SBCVs.

Table 2: Result of classifying flaming cases.

CE	SBCV	SE
0.51	0.41	0.08

3.1 Criminal Episodes

3.1.1 Automatic Identification for CEs

In CEs crime-related words frequently occur. Based on the observation, we use Bayesian filter, the same extraction method for spam email (Graham, 2002), to distinguish a CE from the other. We have prepared 100 tweets representing CEs and 300 general tweets as experimental data. With the data, we perform 5-fold cross-validation.

3.1.2 Experimental Results and Discussion

Table 3 shows accuracy of classification of CEs obtained through the experiment. The result is good enough to say that it is possible to identify CEs by means of the Bayesian filter.

Table 3: Accuracy of detecting CEs.

Recall	Precision	F-Measure
0.97	0.70	0.81

3.2 Struggles between Conflicting Values

In order to determine whether a remark is SBCV or not, it is required to extract the following three elements: the remarks' topic, the remarks' polarity and a reputation of the topic which is defined by polarity of public opinion toward the topic. There are two types of the reputations. One is stable, namely its polarity does not change over the time, such as historical events with established evaluation. Another is dynamic, namely its polarity may change over the time by external factors such as news. Let us examine a flaming case caused by Hollywood actor Ashton Kutcher. He got flamed when he made a sympathetic tweet about longtime Pennsylvania State University football coach Joe Paterno, without knowing a child sex-abuse scandal involving the coach (L.A. times, 2011). Considering his achievement with the team, we can assume that the reputation of Joe Paterno had been positive before the scandal was reported. However, it changed to be deeply negative after the scandal. Mr. Kutcher's failure was caused by his insensitivity to the change. In this case we say that the reputation of Joe Paterno is dynamic. Among the 41 cases of SBCV of Table 2, 32 cases are related to the dynamic reputation. Although for the stable reputation it is possible to describe its polarity using a checklist, for the dynamic reputation it is not trivial how to identify it. We propose a method for it in the next chapter.

4 IDENTIFICATION OF SBCV

4.1 Terminology

In this study, we call a topic dynamic if its reputation is dynamic. We propose a dynamic P/N analysis as a method of detecting the reputation of a dynamic topic.

P/N analysis is to categorize topics' polarity into positive and negative. A dynamic topic is called a dynamic P/N topic if its polarity is positive or negative. Dynamic P/N analysis is to analyze the reputation of a dynamic P/N topic considering influences of good/bad news and the passage of time.

4.2 A Dynamic P/N Topic

Figure 1 shows an example of a dynamic P/N topic. The vertical axis of the graph represents polarities of positive, even and negative to the topic. The horizontal axis represents the time.

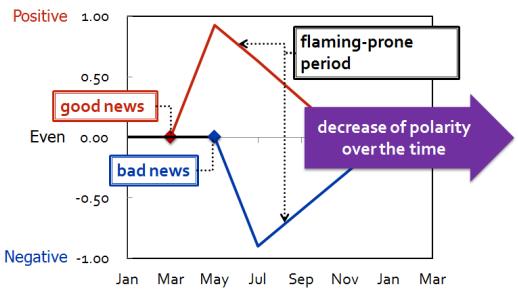

Figure 1: A dynamic P/N topic.

At first, the polarity of the topic is even. Upon hearing good news such as a victory in a tournament series, or bad news such as a tragic accident, its polarity changes positive or negative. At this moment the topic is hot and prone to cause a flaming. The news will be forgotten over the time, and the polarity decreases accordingly.

Definitions of positive-negative-even in this case are as follows (Pak and Paroubek, 2010): *Positive* includes general positive emotions such as happiness, amusement or joy. In addition, it includes concepts such as encouraging, sympathizing, supporting or viewing optimistically. *Negative* includes negative emotions, such as general sadness, anger or disappointment. *Even* includes statements of a fact without any emotional expression.

4.3 Real World Example of SBCV

Let us analyze a Japanese flaming case, which is analogous to Ashton Kutcher's. Olympic judo gold medallist Ryoko Tani got flamed when she made a remark *coach Sonoda is a wonderful person* about former head coach of Japanese national judo team Ryuji Sonoda, after the news that he had been involved in violence and harassment toward female judo wrestlers (Tani, 2013). This is an example of flaming involving a dynamic P/N topic. The reputation of Sonoda had been even before the news. If Mrs. Tani had made the remark before the news, she should have been fine. We explain this because the remark does not conflict against the reputation of the topic, namely, *Sonoda*.

On the other hand, the reputation of Sonoda became negative after the news, just like Joe Paterno. According to our interpretation, she got flamed because she made the remark right at this moment, when the remark conflicted against the reputation.

We have mentioned that the majority of flamings by celebrities belongs to SBCV in section 3.

Although most celebrities are supposed to be attentive to their remarks, they cannot prevent this type of flamings from happening. The fact can be explained if the detection of the reputation of a dynamic topic is difficult. It also shows a potential demand for flaming prediction.

4.4 Experiments regarding SBCV

In order to verify the effectiveness of dynamic P/N analysis we proposed, we carry out an experiment to see if the Sonoda case can be explained by dynamic P/N analysis.

The news on Sonoda's violent behavior was reported on January 29, 2013 (Asahi, 2013) and, Tani made the sympathetic remark about the coach on February 6. In order to analyze the time series of the reputation of coach Sonoda, we collect tweets regarding coach Sonoda for two months from January 24, to March 25. Namely, we prepare 12,825 tweets after removing inappropriate ones by hand from the result to a query "coach Sonoda OR Ryuji Sonoda". As a method of dynamic P/N analysis, we propose daily polarity (dp). dp is a value defined by formula (1), which is the difference of the number of daily P/N tweets, normalized by the number of total tweets.

$$dp_{I,T}(t) = \frac{P_I(t) - N_I(t)}{\sum_{t \in T}(P_I(t) + N_I(t) + E_I(t))} \quad (1)$$

$dp_{I,T}(t)$: daily polarity of topic I at time t in time segment T
$P_I(t), N_I(t), E_I(t)$: number of P/N/E tweets on topic I at time t
T: time segment in which topic I is involved

It shows the strength of the polarity of the reputation that can be read from daily tweets, considering the overall upsurge of the topic. Figure 2 is a graph with the number of P/N tweets and dp. The dp represents the transition of P/N tweets distribution very well.

4.5 Results and Discussion

We discuss three periods in Figure 2.
January 29 to February 1: The negative dp *valley* coincides with the time the news was initially reported and Sonoda announced his resignation.
February 6 to February 8: Events happened during the period include resignation of Yoshimura, a board member of All Japan Judo Federation (AJJF). Tani made the positive remark on Sonoda during the period when Sonoda's reputation was negative. Namely, it is a conflict between Tani's positive view and the negative reputation regarding topic *Sonoda* that causes the flaming.

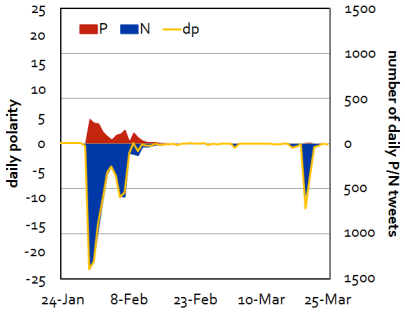

Figure 2: Transition of reputation.

March 19 to March 22: In this period there is news of suspension of grants to AJJF from Japanese Olympic Committee. The negative reputation reflecting public reaction to the news is clearly shown by the dp's movement.

Along with the dynamic P/N analysis, we have also managed to visualize the transition of the reputation by extracting feature words (Ishino *et al.*, 2012) from the periods corresponding to the valleys.

Our method has successfully analyzed the Sonoda case. Although we have shown that another case is similarly explained (Iwasaki *et al.*, 2013), it is necessary to analyze more cases in order to verify the generality of the technique. It is a future work.

5 PROPOSED SYSTEM

We propose a system whose goal is to visualize a remark's likelihood to cause flaming by digitizing it (Figure 3).

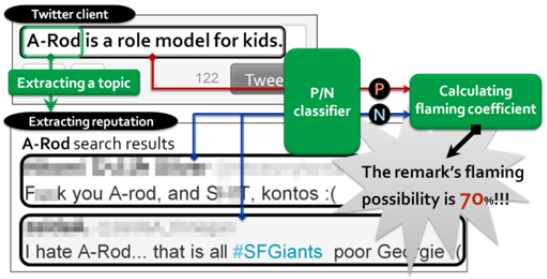

Figure 3: System overview.

Our system's inputs and outputs are as follows:
Input: a remark S, which includes an evaluation on a topic
Output: flaming coefficient, a value calculated from the difference between the polarity of topic I in remark S and that of the reputation on topic I

For example, suppose there is remark S such as *A-Rod is a role model for kids* and recent tweets

including *A-Rod*, which is topic I. The latter acts as the reputation on I. The system calculates the flaming coefficient from the difference between the P/N polarities for each. The daily polarity can be used to determine the polarity of the reputation. It is a future work to determine how to calculate the flaming coefficient from the daily polarity and remark S.

6 PRELIMINARY EXPERIMENT

The experiment aims to evaluate the accuracy of the P/N classifier based on a Japanese polarity dictionary and to investigate the possibility of its improvement by combining it with a machine learning technique. We compare three methods: *MATCH*, which is based on the dictionary, *BAYES*, which is based on Naïve Bayes and *M+B*, which combines MATCH and BAYES using pseudo-words. We explain how to combine MATCH and BAYES.

The polarity information is added to data if the words contained in the dictionary appear in the data, before learning and inference on the data are performed.

Let us take a sentence "I feel rock bottom, but let's do my best" for example. In the sentence we find nouns and verbs contained in the P/N dictionary, such as "rock bottom" and "do my best". Then we add appropriate pseudo words, in this case noun negative (!NN) and verb positive (!VP) respectively, at the end of the sentence (Figure 4).

Figure 4: adding pseudo words to the training data using the P/N dictionary.

MATCH, the method based on the polarity dictionary, has yielded the highest accuracy among the three P/N classifiers. Furthermore, M+B has improved the accuracy comparing to BAYES (Table 4). In order to combine a technique based on machine learning and one based on a polarity dictionary, a method using support vector machine (Mullen and Collier, 2004) is known. Here we tentatively use a method based on Naïve Bayes mostly for simplicity.

Taking these results, we have decided to conduct our study based on MATCH whose F value is over 70% at this stage. Building a near-perfect P/N classifier would be out of our scope.

Table 4: Accuracy of the three P/N classifiers.

Classifiers	Recall	Precision	F-Measure
M+B	0.51	0.51	0.51
Bayes	0.47	0.46	0.47
Match	0.74	0.68	0.71

7 RELATED WORKS

7.1 Researches Related to Flamings

Researches dealing with flamings include the followings.

Yamamoto *et al.* made an investigation into flaming cases by tagging and extracting keywords from 150 trouble cases in CGM (Yamamoto *et al.*, 2009).

Tashiro (Tashiro, 2011) categorized Internet troubles into four types—*financial troubles*, *communication troubles*, *information management troubles* and *mental and physical troubles*. Flamings we are dealing with correspond to the communication troubles in this categorization.

Plus Alpha Consulting Co., LTD. (P.A. Consul., 2011) has developed a system to prevent a user from posting a remark that will be likely to cause a flaming. The system automatically sends a manager an email that asks permission to post it before the actual posting is done.

7.2 Feature Word Extraction

We describe a method using the difference of the word frequencies (Ishino *et al.*, 2012), which is used in Chapter 2 and Chapter 4. This method aims to extract feature words of a target corpus by removing frequent words of a reference corpus from frequent words of the target corpus.

7.3 P/N Analysis

P/N analysis is a typical sentiment analysis that classifies topics into general positive and negative attitudes.

Turney provided a method to determine word's semantic orientation based on words' co-occurrence in a corpus (Turney, 2002). The method can yield a large amount of information for P/N analysis from a relatively few language resource.

This study is positioned as a dynamic P/N analysis, an evolved form of the traditional P/N analysis.

7.4 Hybrid Classifier

Mullen *et al.* proposed a hybrid analysis system (Mullen and Collier, 2004). It is a sentiment analysis system based on a SVM classifier, whose features are augmented by Turney's semantic orientations and polarity values extracted from WordNet. Their experimental results showed that the addition of the features improved accuracy.

8 CONCLUSIONS AND FUTURE WORKS

In this paper, we have defined a flaming in CGM. We also posed propositions to identify the flaming. First, we have presumed that it is possible to extract flaming keywords from flaming recidivists' remarks. However the experiment gave poor result and that led us to our next proposition. Namely, we have presumed that it is possible to classify flamings into the following three categories: *criminal episodes (CEs)*, *struggles between conflicting values (SBCVs)* and *secret exposures(SEs)* .CEs have been identified by Bayesian filter with high accuracy. As for SBCV, we have focused on the dynamics of reputation and analyzed cases that were widely reported by the media. We have succeeded in visualizing the flaming process caused by a gap between the polarity of tweets and that of public opinion. It can be said that we are one step closer to identification of the flaming. Based on those discussions, we can say that the most of the flamings are predictable.

Although we can explain the mechanism of past flaming cases by our research results, it is impossible to verify whether a remark causes flaming in the particular past situation. Therefore, it will be necessary to investigate approaches such as a use of flaming bots which causes flaming.

In the future, we will work on an implementation of the system described in Chapter 6, to identify a remark with a dynamic P/N topic that is likely to cause a SBCV-type flaming. We also consider analyzing the flaming rate, the social influence of poster and banned words.

ACKNOWLEDGEMENTS

This research is subsidized by JSPS 24300005, 23500039, 25730038. The authors would like to express their deepest gratitude to associate all the staff of professor Honiden's lab. of the University of Tokyo and professor Fukazawa's lab. of Waseda University who provided helpful comments.

REFERENCES

Asahi (2013). Female judo wrestlers accused their coach, *Asahi Shimbun*, 29 Jan, in Japanese.

Graham, P. (2002). A Plan for Spam, Available at: http://www.paulgraham.com/spam.html (Accessed at: 11 Nov 2012).

Hiwatashi, K. (2013). (hiwa1118). "You are less than a cockroach." 5 Feb 2013, 0:19 am. *Tweet*, in Japanese.

Ishino *et al.* (2012). Support for Video Hosting Service Users using Folksonomy and Social Annotation, *Proc. of WI-IAT-2012*, pp.472-479.

Iwasaki *et al*. (2013). Identification of Flaming and Its Applications in CGM, *Proc. of JSAI-2013*, in Japanese.

Kobayashi, N. (2011). The Flaming Case File in Social Media, *NIKKEI Digital Marketing*, in Japanese.

L.A. times (2011). Ashton Kutcher prematurely defends fired Penn State coach Joe Paterno, *Los Angeles Times*, Available at: http://latimesblogs.latimes.com/showtracker/2011/11/ashton-kutcher-prematurely-defends-fired-penn-state-coach-joe-paterno.html (Accessed at: 19 Nov 2012).

Mullen, T. and Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources, *Proc. of EMNLP-2004,* pp.412-418.

NAVER (2010). Criminal Episode and A Collection of Flamings, *NAVER's Collection*, Available at: http://matome.naver.jp/odai/2132708118341913001, (Accessed at: 9 Sep 2013), in Japanese.

P.A. Consul. (2011). Customer Rings, *Plus Alpha Consulting*, Available at: http://www.pa-consul.co.jp/LP_rings_mail/ (Accessed at: 20 Apr 2013), in Japanese.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining, *Proc. of the 7th Conf. on International Language Resources and Evaluation*, pp.1320-1326.

Tashiro, M. (2008). Flaming of Blog, *Tokyo Denki University Press*, in Japanese.

Tashiro, M. (2011). Proposal of classification method of Internet related troubles, *JASI Japan*, Vol.6, No.1, pp.101-114, in Japanese.

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, *Proc. of ACL-2002,* pp.417-424.

Yamamoto *et al*. (2009). A study of CGM troubles and a method for their investigation, *IPSJ SIG Technical Reports*, in Japanese.