Spam Comments Detection with Self-Extensible Dictionary and Text-Based Features

Qiang Zhang, Chenwei Liu, Shangru Zhong, Kai Lei*
Institute of Big Data Technologies
Shenzhen Key Lab for Cloud Computing Technology & Applications
School of Electronics and Computer Engineering(SECE)
Peking University, SHENZHEN 518055 P.R.CHINA
Email: {zhangqiang,liuchenwei,shangru}@sz.pku.edu.cn, Corresponding Author: *leik@pkusz.edu.cn

***Abstract*—The new social media have become popular for information spreading, allowing online users to publish latest events and personal opinions. However, massive spam comments seriously decrease users' reading experience. To detect spam comments in Chinese social media, we employ semantic analysis to build the self-extensible dictionary which updates and extends itself with new cyber words automatically. The Semantic analysis brings extra semantic features which helps in text classification. Based on the statistical analysis of microblogging comments, we select four text-based features, which basically represent characteristics of Chinese spam comments. We use spam dictionary and text-based features to construct classifiers for detecting spam comments. Finally, we achieve an average detection accuracy of 93.6%, which is preferable to existing spam comments detection methods. Experimental results demonstrate that our method can effectively detect spam comments in Chinese microblogging field.**

***Index Terms*—spam comments, spam dictionary, text-based features**

I. Introduction

Online social media, such as Twitter and Sina Weibo, have become popular for information sharing. However, the huge popularity makes social media into ideal places to publish spam comments for illicit purpose. Spam comments refer to the unwanted comments with rude words, advertisement, political or religious views. Massive spam comments seriously decrease users' reading experience and hinder the healthy development of social media. Thus, it is essential to detect and filter spam comments.

Spam comments detection has been studied for several years, and various methods have been proposed. For example, some researchers detected spam comments by recognizing spammers. They thought that spammers have unreasonable social networking relations and abnormal behavior[1–3]. Some researchers proposed to detect spam comments by analyzing the content of comments. They selected text-based features and used machine learning methods to classify microblogging comments into spam or not[4–6].

Word vector method is used to convert natural language into mathematical symbols. Distributed representation is a word vector representation method. This method converts the word into a short vector, which is related to the context of this word. Distributed representation provides abundant semantic information which helps in text classification[7, 8]. In this paper, we introduce semantic analysis of microblogging comments to build the self-extensible spam dictionary, which is constructed by two steps. First, we use the Skip-Gram model to transform words into vectors. Then, we calculate the cosine similarity of vectors and expand spam dictionary with plenty of cyber new words. Based on the statistical analysis of massive Chinese microblogging comments, we find the text-based features' distributions of spam comments are significantly different from normal comments. In particular, we choose four text-based features which represent the characteristics of the Chinese microblogging field, including duplicate comments, noun proportion, hyperlink amount and emotional score. We extract spam words and text-based features from each comment and quantify them. We regard spam comments detection as a classification problem and use several classifiers to detect spam comments in testing data. We achieve a better accuracy of 93.6% compared with other methods on the same datasets. Experimental results demonstrate that our method could effectively detect spam comments in Chinese microblogging field.

The contributions of this paper are three-fold.

- Different from the existing methods which mainly use social networking or users' behavior, we achieve preferable results by introducing sematic analysis based on the distributed vector presentation and cosine similarity calculation. Semantic analysis provides us extra features which helps in text classification.
- We build the self-extensible dictionary which expands itself from basic seed spam words. With the extension of spam dictionary, the accuracy of spam comments detection will also increase overtime. Therefore, our approach applies very well to online social media with massive cyber new words created daily.
- Most existing researches focused on Twitter or other popular English microblogs, while we consider the characteristics of Chinese microblogs in our paper. By statistical analysis of labeled comments, we select four text-based features including duplicate comments, noun proportion, hyperlink amount and emotional score.

978-1-5386-1629-1/17/$31.00 ©2017 IEEE

The reminder of this paper is organized as follows. Section II reviews the development of spam comments detection on social media. Section III describes the details of building self-extensible dictionary. Section IV describes the selection and evaluation of four text-based features. Section V describes data collection and discusses our experimental results and analysis. Finally, Section VI outlines conclusion and future work.

II. Related Work

Due to the popularity of social media, such as Twitter and Sina Weibo, many research works have been conducted on spam comments detection. The existing research mainly focused on two aspects: detecting spammers and analyzing content features of spam comments.

Some researchers detected spammers by analyzing social networking. Stringhini et al. considered that spammers have unreasonable social networking relations compared with normal users, such as following lots of users but having less fans[9]. However, it's not difficult for spammers to get lots of ossified fans. Thus, these methods based on social networking may not work well. Some researchers detected spammers by analyzing the users' attributes and representative behaviors, such as registration date, repeated reposting and aggressive following[10, 11]. Lin et al. assumed that spammers' attributes and behavior are different from those of normal users. However, social spammers may continue to change their behaviors and try to behave like normal users[10]. These two methods focus on analyzing social networking, the attributes and behavior of users, detecting spam comments by recognizing spammers. However, from our observation, normal users occasionally have abnormal behavior, but they post normal comments most of the time. Simply taking them as spammers and blocking them may not be effective.

Some researchers proposed to detect spam comments by analyzing the comments content. Rdulescu et al. detected spam comments by post-comment theme coherence. They calculated the words co-occurrence frequency between microblogs and comments, then get the post-comment similarity. They thought that low similarity means that comment might be spam[4]. However, from our observation, lots of normal comments are short and inconsistent with microblogs' theme. Liu et al. proposed spam dictionary and Proportion-Weight Filter model to detect two kinds of spam comments (advertisement and vulgar comments), and achieved an average accuracy value of 87.6%[12]. Their results have much room for improvement because they ignored other text-based features. Romero et al. performed a comparative study using four classifiers (Naive Bayes, K-Nearest Neighbors, Neural Networks and Support Vector Machines) in spam comments detection. Support Vector Machines got the highest performance of 84.6%[13, 14]. By extracting several useful features, these machine learning methods get decent results. However, with lots of new cyber words created daily, these methods may not understand the ever-changing expression.

Our approach differs from the existing approaches in two aspects. First, we introduce semantic analysis to expand spam dictionary. Semantic analysis provides us with additional useful features to detect spam comments. Second, using statistical analysis, we select four text-based features, which present the characteristics of Chinese microblogs. Finally, we train classifiers using spam dictionary and text-based features. Those distinctive features make our approach achieve a more preferable accuracy of 93.6%.

III. Self-Extensible Dictionary

In this paper, spam words refer to the words decreasing users' reading experience in microblogging comments. Therefore, spam words are critical for spam comments detection. with lots of new cyber words created daily, artificially building spam dictionary could not expand itself with new cyber words. In this section, we introduce the details of building the self-extensible spam dictionary.

A. Building Seed Spam Dictionary

To get the seed spam dictionary, we organize several people to select spam words from Chinese microblogging comments. The selecting strategy is that a word will be added into seed spam dictionary if two-thirds people select it. Finally, we get 675 seed spam words including 5 categories: vulgar, advertising, erotic, politically sensitive and criminal words.

B. Expanding Spam Dictionary with New Cyber Words

With massive new cyber words created daily, there is a potential limitation of seed spam dictionary. It's difficult to update and expand itself with new cyber words timely. To overcome this problem and improve the result of spam comments detection, we use the Skip-Gram model to convert the words in comments to vectors, and calculate the cosine similarity between them[15–17].

The Skip-Gram model adopts distributed representation, which convert words into short vectors. The dimension-reduction representation makes it easy to calculate cosine distance between two words. The Skip-Gram model predicts context words by middle word[18]. Mathematically, it maximizes the objective function:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

In Eq.(1), w_t and w_{t+j} means the middle and context words separately in training corpus and c means the size of training window. We compute the log probability of predicting word w_{t+j} from -c to c in the training window. The value of $p(w_{t+j}|w_t)$ is calculated by Eq. (2).

$$p\left(w_{t+j}|w_t\right) = \frac{exp({v_{w_{t+j}}}^T v_{w_t})}{\sum_{w=1}^{W} exp({v_w}^T v_{w_t})} \tag{2}$$

In Eq.(2), v_{w_t} and $v_{w_{t+j}}$ means vectors of middle and context words. The probability of every two words in training window will be calculated, even though they are separated by some words. In order to accelerate the training speed, we adopt Hierarchical Softmax algorithm in our solution. This accelerating algorithm calculates conditional probability values in Eq.(2) by building the Huffman tree.

The semantic similarity of two words is calculated by Eq.(3). $Sim(A, B)$ means the similarity of words A and B. The similarity is calculated by cosine value of vectors x_i and y_i. The higher the cosine value is, the greater the similarity of two words will be.

$$Sim(A, B) = cos(\theta) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (3)$$

We use segmented microblogging comments to train Skip-Gram model and convert these words into vectors. For each word in seed spam dictionary, we calculate the similarity of this word and other word in microblogging comments and select 20 results with high similarity scores. By our observation, the result with higher similarity is not necessary more similar with seed spam word. Therefore, it is not reasonable to add these words into spam dictionary directly.

In order to solve the problem above, The Iterative Algorithm is proposed for expanding spam dictionary. The detailed process is shown in Algorithm 1.

Algorithm 1 The Iterative Algorithm for expanding spam dictionary.

Input: Seed spam words.
for each word s in Input file **do**
  Acquire 15 most similar words $(s'1...s'15)$ and their similarity weights $(w(s'1)...w(s'15))$;
  for each word s' in $(s'1...s'15)$ **do**
    if s' in seed spam dictionary or $w(s') < 0.4$ **then**
      Drop it;
    else
      Add the word s' into the candidate spam dictionary;
    end if
  end for
end for
for each word s' in candidate spam dictionary **do**
  Acquire 15 most similar words $(s''1...s''15)$;
  if there are three or more words of $(s''1...s''15)$ exist in seed spam words **then**
    add this candidate word s' into dictionary;
  else
    Drop it;
  end if
end for
Output: New cyber spam words.

As is shown in Algorithm 1, we use the Iterative Algorithm to expand spam dictionary with new cyber words. First, for each word in seed spam dictionary, we select 15 most similar words and add them into candidate spam dictionary if the similarity is greater than 0.4. Second, for each word in candidate spam dictionary, we find 15 most similar words and add this candidate word into spam dictionary if there are three or more of them included by seed spam dictionary. Finally, we get the self-extensible spam dictionary.

IV. Text-based Features

Detection spam words is effective to detect spam comments, but not all spam comments include spam words. For example, massive duplicate comments also decrease users' reading experience, and harmful hyperlinks have potential hazards to users. By observation and statistical analysis, we select and evaluate several text-based features.

A. Feature Selection

Duplicate Comments: Spammers are more inclined to publish plenty of duplicate comments to attract more attention, whereas normal users tend to publish different comments to express unique opinions. Duplicate comments could be measured by the Edit Distance, which is defined as the minimum cost of transforming one string into another through several edit operations, such as deletion, insertion and substitution of individual symbol. We find that spammers usually change few words to avoid complete replication, therefore, we set the threshold as 2. After removing special symbols from original comments, we calculate the Edit Distance of two comments. If the Edit Distance of two comments is less than or equal to 2, they will be considered as duplicate comments.

Noun Proportion: By analyzing plenty of spam comments, we find that advertisement comments are inclined to propaganda their products by using lots of noun, such as "New products are on sale every day, such as skin care products, diet pills, luxury, popular shoes, women clothes, scarves". We use HanLP tools to extract Part-of-Speech tags from the comments and calculate noun proportion of each labeled comment.

Hyperlink Amount: Malicious hyperlinks attached in spam comments spread quickly and widely because many famous users have plenty of followers. However, it's difficult for microblogging platforms to distinguish between normal and spam hyperlinks. Different from other media, by our observation, plenty of advertisement comments attach their WeChat ID in Chinese microblogging field. We consider it as another style of hyperlink. Finally, hyperlink amount of each labeled comment is counted if a comment contains the sequence of characters "http://", "www." or "WeChat ID".

Emotional Score: We find that spam comments are more inclined to express passive emotion. The emotional phrase usually consists of negative words, degree adverbs and emotional words, such as "not very good". In order to calculate emotional score of a sentence, we assume that emotional scores of all

phrases are linearly additive. The average emotional score of a sentence is calculated by following equation.

$$EmotionalScore = \frac{1}{n}\sum_{i=1}^{n} w(n)_i w(d)_i w(s)_i \quad (4)$$

In Eq.(4), n means amount of phrases in a comment. In ith phrase, $w(n)_i$ means the weight of negative word, and $w(d)_i$ means the weight of degree adverb, and $w(s)_i$ means the weight of emotional word.

We basically use existing dictionaries, including positive emotional dictionary, passive emotional dictionary, negative dictionary and degree adverb dictionary, which are published by HowNet. We assume that all positive emotional words have equal weights 1, and all passive emotional words are -1, and all negative words are -1. We divide degree adverbs into 3 levels and assign weights as 0.5, 1 and 2 respectively.

B. Feature Evaluation

In order to discover the differences between spam and normal comments, we count the amounts of text-based features in 2000 labeled comments. The proportions of text-based features in spam and normal comments are shown in Fig.1. We find the text-based features' distributions of spam comments are significantly different from normal comments. For example, there are 20 duplicate comments in 1000 normal comments, while 430 in 1000 spam comments.

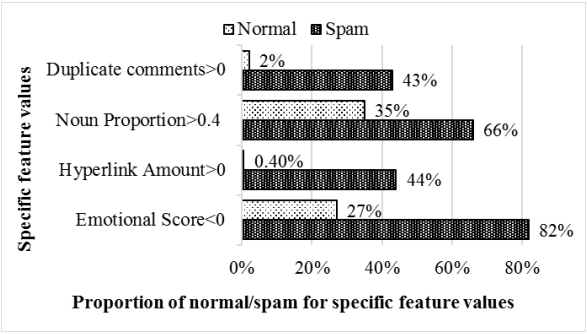

Fig. 1: Proportion of feature in normal and spam comments.

Not only that, we also use Chi-Square value to calculate whether text-based features are relevant to spam comments. As to the independence between every feature and spam comments, we suppose each feature is independent of the spam comments. The Chi-Square threshold value is 3.84 when α = 0.05 and d = 1. For the independence test, we reject hypothesis above if the Chi-Squared value is greater than 3.84. The Chi-Square value of each feature are shown in Table 1.

We find that the Chi-Square value of each feature is much greater than the threshold 3.84. Therefore, we reject the hypothesis above, in another word, each feature has strong relationship with spam comments and can be used as a feature in spam comments detection.

TABLE I: The Chi-Square value of each feature.

Feature	Chi-Square Value
Duplicate comments	*210*
Noun Proportion	*41*
Hyperlink Amount	*274*
Emotional Score	*195*

V. Experiment Evaluation and Related Analysis

In this paper, we primarily collect the comments using the Sina Weibo API. Considering the amount of comments, we continuously crawl the comments from microblogs released by ten famous users between 12/15/2015 and 02/14/2016.

We artificially labeled 1000 spam comments and 1000 normal comments and preprocessed labeled data using the Jieba Word-Segmentation tool and remove punctuation and stop words using stop words list. Then, we extract spam words and text-based features from labeled comments, then quantify and normalize these features. We get labeled data, such as "1, 0, 1, 0.33, 0.37, Y", where the first five numbers are features' values and the last symbol is the label. We randomly select 70% labeled data as training data, the rest 30% as testing data.

We conduct four sets of experiments, First, we expand spam dictionary using Iterative Algorithm. Second, we evaluate the performances of spam dictionary and text-based features. Third, we use three classifiers to detect spam comments in testing data based on the different features. Finally, we compare our model with other models on the same datasets.

To evaluate the performance of a classification algorithm, we define the confusion matrix as follows:

	Predicted Spam	Not Spam
Real Spam	*a*	*b*
Not Spam	*c*	*d*

In confusion matrix, a represents the number of spam comments that are correctly classified, b represents the spam comments that are falsely classified as non-spam, c represents the number of non-spam comments that are falsely classified as spam, and d represents the number of non-spam comments that are correctly classified.

We consider the following evaluation parameters:

$PrecisionRate(P)$ = a / $(a + c)$

$RecallRate(R)$ = a / $(a + b)$

$F1$ is the balanced value of P and R used to evaluate the overall result of classification.

$$F1 = 2 * P * R/(P + R)$$

A. Experiments of Expanding Spam Dictionary

In order to implement Skip-Gram model and Hierarchical Softmax algorithm, we use an open source toolkit, which is

published by Gensim. We set word vector dimension value as 200 and training window value as 10, and use default values as other parameters. The training corpus comes from microblogging comments we crawled from Sina Weibo. After words segmentation and removing stop words and punctuation, we train the Skip-Gram model using these words and convert all these words into vectors.

We use Iterative Algorithm to expand spam dictionary with 6673 new cyber spam words based on 675 seed spam words. Since the Iterative Algorithm is an unsupervised method, there is no testing data to evaluate the result. We artificially evaluate the accuracy of Iterative Algorithm by randomly selecting 600 words from 6673 new cyber spam words, and find that there is 79 words not belonging to spam words, then we get the accuracy of 86.8%, which is acceptable.

B. Experiments of Evaluating Performances of Features

We use Pointwise Mutual Information(PMI) to evaluate performances of features. The PMI value is used for calculating correlation of two things. The PMI value is zero if two things are not relevant. The greater the PMI value is, the more relevant two things will be. The PMI value is calculated by Eq. (5):

$$I(X,Y) = \log \frac{P(X,Y)}{P(X)P(Y)} \tag{5}$$

In Eq.(5), $I(X,Y)$ means the PMI value of X and Y, $P(X,Y)$ means the joint probability of X and Y. The PMI values of spam dictionary and four text-based features are shown in Table 2.

TABLE II: The PMI value of each feature.

Feature	PMI Value
Duplicate comments	*0.93*
Noun Proportion	*0.38*
Hyperlink Amount	*0.94*
Emotional Score	*0.58*
Spam Dictionary	*0.98*

We find that the PMI value of each feature is significantly greater than zero. In particular, we think the feature "Spam Dictionary" has stronger relation with spam comments compared with other features. Therefore, the spam dictionary and text-based features we extracted are significantly relevant with spam comments.

C. Experiments of Three Classifiers with Different Features

It is insufficient to use only spam dictionary or one of text-based features to determine whether a comment is spam or not. In order to comprehensively measure every feature, we regard spam comments detection as a classification problem and use several classifiers to classify microblogging comments. We use Scikit-learn toolkit to construct classifiers, which provides simple and efficient tools for data mining and data analysis.

We conduct two sets of experiments at the feature level. First, in order to study how well text-based features perform in spam comments detection, we train three classifiers using proposed text-based features as baseline. Second, in order to study the effect of semantic analysis, we train three classifiers using combinations of self-extensible spam dictionary and text-based features. We calculate the average F1 value of each experiment based on the testing data. The experimental results are shown in Fig.2.

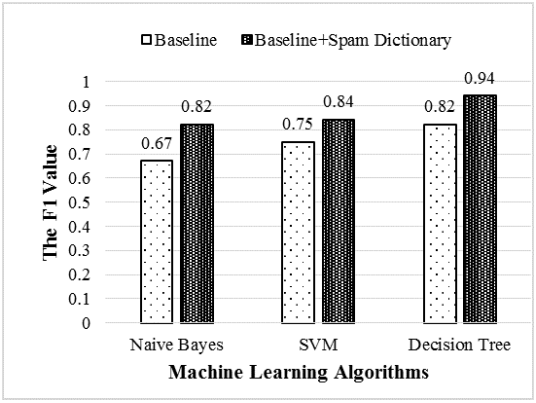

Fig. 2: The F1 value of three machine learning algorithms.

Compared with only using text-based features, the results of three classifiers improve 15%, 9% and 12% separately after adding self-extensible spam dictionary. The result demonstrates the effectiveness of self-extensible dictionary for spam comments detection. Taking the performances of three classifiers into condition, we find that the decision tree model has the better F1 value. Therefore, we choose the decision tree model based on the four proposed text-based features and self-extensible dictionary as classifier to detect spam comments.

D. Experiments of Comparing with Other Researches

There are other researches focusing on the problem of spam comments detection. Considering Sina Weibo is the most popular microblog platform in Chinese, we compare our model with others, which also performed experiments on the Sina Weibo platform. Liu et al. proposed spam dictionary and Proportion-Weight Filter(PWF) model to detect advertisement and vulgar comments[12]. Wu et al. used Social Spammer and Spam Message Co-Detection(S3MCD) model to detect spam comments[1]. In order to compare the performance of different models, we test PWF and S3MCD models on our labeled comments datasets. The average F1 values of three models are shown in Fig.3.

Our results have a slight improvement compared with the best results of models above. We owe the improvement to the self-extensible dictionary, which expanding itself with plenty of new cyber words timely. Therefore, with the capacity of spam dictionary increasing, the accuracy of spam comments detection will also increase overtime.

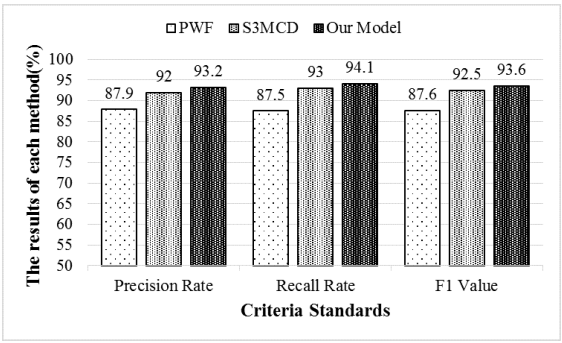

Fig. 3: A comparison of our result with others.

VI. Conclusion and Future Work

In this paper, we conduct deep analysis about spam comments detection. We introduce semantic analysis to build the self-extensible dictionary, which update and expand itself with new cyber words. By statistical analysis, we extract four text-based features including duplicate comments, noun proportion, hyperlink amount and emotional score. These features express the differences between spam and normal comments. Based on features above, we train classifiers to detect spam comments. Our experimental results achieve an average accuracy of 93.6%, which is preferable to other researches for detecting spam comments in Chinese microblogging field.

Although the experimental results of our approach are acceptable for spam comments detection, there are useful suggestions to improve the result. In future work, we will consider combining the spammer recognition with the current approaches to see if the accuracy can be further improved.

VII. Acknowledgement

This work has been financially supported by the National Natural Science Foundation of China (No.61602013), the Shenzhen Key Fundamental Research Projects (Grant No. JCYJ20160330095313861, JCYJ20151030154330711 and JCYJ20151014093505032).

References

[1] Fangzhao Wu, Jinyun Shu, Yongfeng Huang, and Zhigang Yuan. Co-detecting social spammers and spam messages in microblogging via exploiting social contexts. In *Neurocomputing*, volume 201, pages 51–65, 2016.

[2] Lin Liu and Kun Jia. Detecting spam in chinese microblogs-a study on sina weibo. In *Computational Intelligence and Security*, pages 578–581. IEEE, 2012.

[3] Chao Yang, Robert Chandler Harkreader, and Guofei Gu. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *Recent Advances in Intrusion Detection*, pages 318–337. Springer, 2011.

[4] Cristina Rdulescu, Mihaela Dinsoreanu, and Rodica Potolea. Identification of spam comments using natural language processing techniques. In *Intelligent Computer 10 Communication and Processing (ICCP)*, pages 3111–3119. IEEE, 2013.

[5] Yang Shen, Shuchen Li, Xiaoxiao Ye, and Fangping He. Content mining and network analysis of microblog spam. *Journal of Convergence Information Technology*, 5(1):135–140, 2010.

[6] Xin Jin, C Lin, Jiebo Luo, and Jiawei Han. A data mining-based spam detection system for social media networks. *Proceedings of the VLDB Endowment*, 4(12):1458–1461, 2011.

[7] Alex Hai Wang. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference*, pages 1–10. IEEE, 2010.

[8] Joseph Lilleberg, Yun Zhu, and Yanqing Zhang. Support vector machines and word2vec for text classification with semantic features. In *Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 136–140. IEEE, 2015.

[9] Gianluca et al. Stringhini. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 1–9. ACM, 2010.

[10] Chengfeng Lin et al. Analysis and identification of spamming behaviors in sina weibo microblog. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 2013.

[11] Jong Myoung Kim, Zae Myung Kim, and Kwangjo Kim. An approach to spam comment detection through domain-independent features. In *Big Data and Smart Computing (BigComp), 2016 International Conference on*, pages 273–276. IEEE, 2016.

[12] Chenwei Liu, Jiawei Wang, and Kai Lei. Detecting spam comments posted in micro-blogs using the self-extensible spam dictionary. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2016.

[13] Archana Bhattarai, Vasile Rus, and Dipankar Dasgupta. Characterizing comment spam in the blogosphere through content analysis. In *Computational Intelligence in Cyber Security, 2009.*, pages 37–44. IEEE, 2009.

[14] Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. Survey of review spam detection using machine learning techniques. In *Journal of Big Data*, volume 23, 2015.

[15] Zengcai Su, Hua Xu, Dongwen Zhang, and Yunfeng Xu. Chinese sentiment classification using a neural network tool! word2vec. In *Multisensor Fusion and Information Integration for Intelligent Systems, 2014 International Conference on*, pages 1–6. IEEE, 2014.

[16] Tomas Mikolov, Kai Chen, et al. word2vec, 2014.

[17] Tomas Mikolov et al. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[18] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.