International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 6, Issue 3, Apr-May 2017; 21-30

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

CYBERBULLYING DETECTION IN TWTTER USING LANGUAGE EXTRACTION BASED SIMPLIFIED SUPPORT VECTOR MACHINE (SSVM) CLASSIFIER

SHERLY T.T[1] & B. ROSILINE JEETHA[2]

[1]Research Scholar, PG and Research Department of Computer Science,
Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India

[2]Research Guide, PG and Research Department of Computer Science,
Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India

ABSTRACT

Text mining is the thrust research area in the field of data mining and knowledge engineering. The communication data commencing online social networks is capable enough to offer new insights for building societies that was earlier thought as impossible in terms of scale and extent. Cyberbullying is a common thing in social networks like twitter which is described as the use of information and communication technology by an individual or a group of users to annoy other users. This research work presents cyberbullying detection in twitter using language extraction and simplified support vector machine classifier. Around 4556 tweets are taken from the Twitter. The proposed SSVM classifier allowed to train with 3000 tweets. The SSVM is compared with existing SVM classifier. Simulations are carried out using MATLAB 2012. The result shows that the proposed language extraction based SSVM outperforms than that of the existing classifier.

KEYWORDS: Cyber Bullying, Language Extraction, Support Vector Machine, Stop Word Removal Text Mining

INTRODUCTION

Text mining is the branch of data mining which is the form of text analytics. Text mining is the process of obtaining information from text. Many techniques are involved in text mining. In the real time scenario, many texts are in the form of semi – structured and leads to more scope of research in text data mining area. One such research problem is cyberbullying detection in twitter. Cyber criminals have utilized social media as a new platform in committing different types of cybercrimes, such as phishing (Aggarwal, Rajadesingan, & Kumara guru, 2012), spamming (Yardi, Romero, & Schoenebeck, 2009), spread of malware (Yang, Hark reader, Zhang, Shin, & Gu, 2012), and cyberbullying (Weir, Toolan, & Smeed, 2011).

In particular, cyber bullying has emerged as a major problem along with the recent development of online communication and social media (O'Keeffe & Clarke-Pearson, 2011). Cyberbullying can be defined as the use of information and communication technology by an individual or a group of users to harass other users (Salmivalli, 2010). Cyberbullying has also been extensively recognized as a serious national health problem (Xu, Jun, Zhu, & Bellmore, 2012), in which victims demonstrate a significantly high risk of suicidal ideation.

Twitter is a common online social network service that enables users to send and read 140-character messages. The Twitter network currently includes over 560 million users, of which 288 million are keenly exchange a few words through this network and create approximately 580 million tweets every day. Around 80% of these keen users of twitter are

posting their tweets using their smart phones and tablets. Even though twitter turned into an important, near real-time communication channel (Kavanaugh et al., 2012), a study determined that Twitter is turning into a “cyberbullying playground” (Xu et al., 2012).

Twitter which is one of the social networks turned into new targets for cybercrime, and malicious users attempt to perform illegal activities such as cyber attacks, bullying, fraudulent information, organized crimes, and even terrorist attack planning on these systems (Yu et al., 2015). In addition Twitter is more prone to malwares, spam messages and other offensive materials (Akoglu et al., 2010; Gao et al., 2012; Hassanzadeh and Nayak, 2013; Rahman et al., 2012; Shrivastava et al., 2008). It is obvious that cyber bulling causes not only monitory loss and also affects a person’s behavior patters. Such activities of the cyberbulliers in Twitter necessitate the scope of cyber forensics in social networks arena.

This research work aim to make use of essential information in tweets for improving the cyberbullying detection performance. The stem and branch words are initially removed from the collected tweets dataset. Next certain features in tweets are utilized to train the detection model and improve its performance.

LITERATURE REVIEW

Mohammed Ali Al-garadi et al., proposed a set of unique features; such as network, activity, user, and tweet content derived from Twitter. A supervised machine learning solution has been proposed based on the feature for cyberbullying detection in the Twitter. The evaluation results of the authors work provided a feasible solution to detecting Cyberbullying in online communication environments through their proposed detection model. The authors used data collected from Twitter between January 2015 and February 2015 for their evaluation process. 2.5 million geo-tagged tweets within a latitude and longitude boundary of the state of California have been fetched using the sampled API service of Twitter. The authors categoriesed the features as network, activity, user, and content, to detect cyberbullying behavior, and used NB, SVM, random forest, and KNN for machine learning. All the four classifiers have been evaluated in four various settings, namely, basic classifiers, classifiers with feature selection techniques, classifiers with SMOTE alone and with feature selection techniques, and classifiers with cost-sensitive alone and with feature selection techniques. AUC has been considered for the measure of performance. AUC has high robustness for evaluating classifiers. Precision, recall, and f-measure were also used as reference measures. Random forest using SMOTE alone proven the best AUC (0.943) and f-measure (0.936).

R. Forssell investigated the prevalence of cyberbullying and face-to-face bullying in Swedish working life and its relation towards gender and organisational position. A large sample of 3371 respondents has been involved in the study. A cyberbullying behaviour questionnaire (CBQ) has been used in the study; 9.7% of the respondents have been labelled as cyberbullied in accordance with Leymann's cut-off criterion, 0.7% of the respondents as cyberbullied and 3.5% of the respondents as bullied face-to-face. Their study also revealed that men when compared with women were exposed to a high degree of Cyberbullying. Individuals with a supervisory position were observed with more exposure on cyberbullying than persons with no managerial responsibility.

Manuel Gámez-Guadix et al, examined the possibility of the presence of an identifiable group of stable victims of cyberbullying. The author analysed the stability of cyber victimization associated with the perpetration of cyberbullying and bully–victim status. The psychosocial problems of non-stable victims and noninvolved peers have been compared with stable victims. The authors used a sample of 680 Spanish adolescents which includes 410 girls in completing the self-report

measures on cyberbullying perpetration and victimization, depressive symptoms, and problematic alcohol use at two time points that were separated by one year. The cluster analyses results suggested the existence of four distinct victimization profiles. Stable-Victims (5.8% of the sample) were observed with victimization at both Time 1 and Time 2.

The authors also found that the stable victims were more likely to fall into the bully–victim category and presented more cyberbullying perpetration than the rest of the groups. Time1-Victims (14.5% of the sample) and Time 2-Victims (17.6% of the sample) presented victimization only at one time. Non-Victims (61.9% of the sample) presented minimal victimization at both times. Overall, the authors observed that the Stable victims group with higher scores of depressive symptoms and problematic use of alcohol over time than the other groups, whereas, the Non-Victims with the lowest of the scores. Their findings have been observed with major implications for prevention and intervention efforts intended at reducing cyberbullying and its consequences.

A previous study proposed an approach for offensive language detection that was equipped with a lexical syntactic feature and demonstrated a higher precision than the traditional learning based approach (Chen, Zhou, Zhu, & Xu, 2012). A YouTube databased study (Dadvar, Trieschnigg, Ordelman, & de Jong, 2013) applied SVM to detect cyberbullying, and determined that incorporating user-based content improved the detection accuracy of SVM. Using data sets from MySpace, Dadvar et al. developed a gender-based cyberbullying detection approach that used the gender feature in enhancing the discrimination capacity of a classifier. Dadvar et al. and Ordelman et al. included age and gender as features in their approach; however, these features were limited to the information provided by users in their online profiles. Moreover, most studies determined that only a few users provided complete information about themselves in their online profiles. Alternatively, the tweet contents of these users were analyzed to determine their age and gender (D. Nguyen, Gravel, Trieschnigg, & Meder, 2013).

Several studies on cyberbullying detection utilized profane words as a feature (Kontostathis, Reynolds, Garron, & Edwards, 2013), thereby significantly improving the model performance. A recent study (Squicciarini, Rajtmajer, Liu, & Griffin, 2015) proposed a model for detecting cyberbullies in MySpace and recognizing the pairwise interactions between users through which the influence of bullies could spread. Nalini and Sheela proposed an approach for detecting cyberbullying messages in Twitter by applying a feature selection weighting scheme (Nalini & Sheela, 2015). Chavan and Shylaja included pronouns, skip-gram, TFeIDF, and N-grams as additional features in improving the overall classification accuracy of their model (Chavan & Shylaja, 2015).

PROPOSED WORK

Pre Processing

This step aims to renovate raw tweet messages into a computer-readable form. There are four tasks involved such as information extraction, noise removal, slang normalization and language extraction.

Information extraction: Tweets are in several forms such as text, video or image. This research work considers only texts collected form the dataset. These tweets are made avilable which contains tweets from twitter. Taking into account all information are extracted from the HTML files. These contains hash tag (#) at symbol (usually the username), tweet messages and their posting time. From all these information, only the tweet messages are deficient in a standard format.

Noise removal: The most frequent form of noise in tweet messages is the pointless repeated use of punctuation

marks or letters. Such repeatedness probably happens at any position such as start, middle, or end of a tweet message. For example, “okkkk” and “really?????” have single character repeat at the end, “hahaha” has double character repeats, and “wowwowwow” has triple character repeats. Regular expressions are used to normalize all such kind of repeats. To normalize the words containing punctuation marks and digit-letter mixing, it is presumed that no character is possibly repeated more than once continuously, and consequently the excessive characters are dropped, such that “f99” becomes “f9” which is a slang version of “fine”. As far as letters are concerned, it is presumed that are not repeated endlessly more than twice, and as a result drop the extra repeats, such that “oookkkk” becomes “ok”, “okk” remains as it is, “freakkkky” becomes “freaky”, “freaeakkkky” becomes “freaky”, and “add” also remains as it is.

Slang normalization: Slang expressions are generally employed in tweet messages which does not present in standard dictionaries. For that reason, list of slang expressions are compiled and their equivalent standard terms are replaced. A table-lookup process is employed in order to scan the complete set of chat messages to recognize slang expressions and replace them with the equivalent standard terms. For example, “⟨ lol, Laugh out loud⟩” replaces each occurrence of “lol” by “Laugh out loud”.

Language extraction: Commonly, language consists of a set of terms that fully wrap a user's or communicative knowledge. In context of tweet, we define language as a set of important information containing keyterms exchanged among participating users during its complete life of communication. Thus, the language extraction process intends to recognize language of the community involved in tweets, and it comprises sub-tasks such as x-gram extraction, stop-word removal, stemming, and case-folding. The x-gram is described as a series of x successive words in a portion of text. Based on the value of x, it can be a 1-gram containing single word if n is 1; 2- gram containing two consecutive words if n is 2; and so on. In view of the fact that a keyword generally consists of three words at most and exceeding this is an exception. Our proposed technique generates chunks of texts from tweet messages by a process called vouchering and then extracts all possible 1-, 2-, and 3-grams from them. Usually such beginning or ending with a stop-word probably may not be finite in their information. Thus, all such n-grams are filtered out from the list. Consequently, a 1-gram is discarded if it is a stopword, a 2-gram is discarded if any of its constituting words is a stop-word, and a 3-gram is discarded if any of the boundary words is a stop-word.

Furthermore, a single word also contains various forms (e.g., “laugh” and “laughing”), that will carry the same meaning but does not equal exactly. In order to counteract this, all x-grams are stemmed in their base forms. Likewise, a word can be differently cased, that probably results in mismatch for same words. To normalize this all x-grams are case-folded.

Simplified Support Vector Machine Classifier

SVM is a supervised machine learning classifier which is applied for categorization. SVM finds the best possible surface to separate the positive samples from the negative samples. SVM is comparatively better than that of text classification when compared to Naive Bayes (NB) classifier and maximum entropy based classifiers. The fundamental aim of SVM during the training process is to hit upon a maximum margin hyperplane to solve the feature review’s classification task. There exist limitless possible boundaries in order to break up the two different classes. For choosing the best class, it is significant to prefer a decision boundary which contains a maximum margin between any points from both classes.

The decision boundary with a maximum margin would be less likely to make prediction errors, which is close to the boundaries of one of the classes. In this part of research a simplified SVM that is capable enough to classify multi-class and performs dual roles. In the beginning, making a model for the training data set and then using that model to conclude facts of a testing data set.

The proposed SSVM procedure includes the following steps.

- Transform data to the format of an SSVM package
- Conduct simple scaling on the data.
- Consider the RBF kernel.
- Find the best parameter using cross validation to train the whole training set.
- Test.

After pre-processing, the above procedures are carried out for training the SSVM. The basic form of features and its classification is illustrated in the following equation.

$$\phi = (D_s \times C_s) \rightarrow \{P, N\} \quad \textbf{(1)}$$

where D_S is a set of documents and C_S is a set of categories.

If ∅: ($D_S \times C_S$) = P, then D_{Si} is a positive member of C_{Sj}

If ∅: ($D_S \times C_S$) = N, then D_{Si} is called a negative member of C_{Sj}.

The SSVM method gives a positive value (+1) in the most appropriate holding data points and a negative value (−1) in rest of the places. Furthermore, the non-linear mapping function, that maps the training data can be defined as follows.

$$\emptyset: R^N \rightarrow R^F \quad \textbf{(2)}$$

where R^N is a non-linear mapping that represents training data for feature space R^F. Hence, there is a need for performing optimization in order to segregate the dataset.

The kernel functions provide more decision functions when the data are nonlinearly separable. The kernel functions used the following polynomial function and Gaussian Radial-Basis Function (RBF). The RBF kernels of SSVM are used in our system to build models. These models predict information for the testing data set. The representation points for each feature vector lay on a 1D plane and cannot be separated by a linear hyperplane. Therefore, the system will first use a kernel function that maps the points into feature space and then separates the points by hyperplane. The kernel function that will do the job is k (xi, xj) = ∅ (xi) × ∅ (xj). In addition, the kernel polynomial function maps the feature space points into 2D by multiplying the points to the power of two.

RESULTS AND DISCUSSIONS

The proposed research work is named as SSVM and compared with the existing SVM classifier. Performance metrics such as classification accuracy and time taken for classification are chosen for comparison. 4556 tweets from various topics such as demonetisation, kids, mobile phones, sachin and whatsapp words are searched in Twitter and

analyzed as positive opinion tweets and negative opinion tweets. The analyzed tweets are presented in Table 1.

Table 1: Collected Tweets with Various Search Terms and Opinion Analysis

File Name	Total No. of Tweets	Actual	
		Positive Opinion Tweets	Negative Opinion Tweets
demonetisation.txt	1003	498	505
kids.txt	984	402	582
mobilephones.txt	783	599	184
sachin.txt	994	483	511
whatsapp.txt	792	599	193

- True Positive (TP) → Correctly identified as positive opinion tweets
- False Positive (FP) → Incorrectly identified as positive opinion tweets
- True Negative (TN) → Correctly identified as negative opinion tweets
- False Negative (FN) → Incorrectly identified s negative opinion tweets

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100$$

Table 2: Performance Analysis of the Classifiers

Performance of the Existing SVM Classifier						
File Name	TP	TN	FP	FN	Accuracy (%)	Execution Time (seconds)
demonetisation.txt	358	386	123	136	74.18	228
kids.txt	299	445	123	117	75.61	219
mobilephones.txt	446	137	101	99	74.46	186
sachin.txt	383	386	103	122	77.36	223
whatsapp.txt	421	162	131	78	73.61	199
Performance of the Proposed SSVM Classifier						
File Name	TP	TN	FP	FN	Accuracy (%)	Execution Time (minutes)
demonetisation.txt	448	455	56	44	90.03	161
kids.txt	394	498	46	46	90.65	158
mobilephones.txt	542	149	45	47	88.25	149
sachin.txt	433	461	51	49	89.94	163
whatsapp.txt	554	151	55	32	89.02	142

The Table 2 presents the performance analysis of the proposed SSVM and existing SVM classifiers. It can be observed that the overall accuracy of the SSVM classifier is improved by 15%. The implementation of the proposed SSVM and existing SVM classifiers are implemented using MATLAB. The performance analysis in terms of accuracy is shown in Figure 1. It is to be noted that the execution time of the proposed SSVM classifier is comparatively lesser than that of the existing SVM classifier. The performance analysis in terms of execution time is shown in Figure 2.

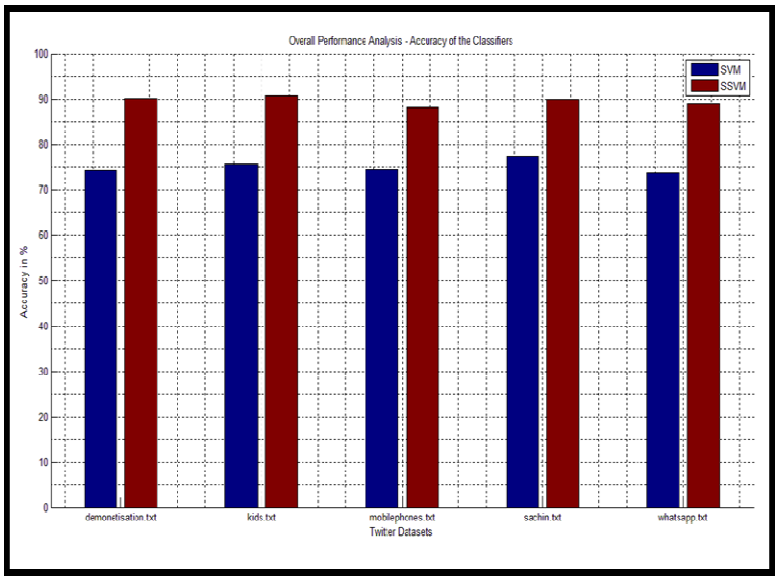

Figure 1: MATLAB Result Graph for Performance Analysis of the Classifiers in Terms of Accuracy (In Percentage)

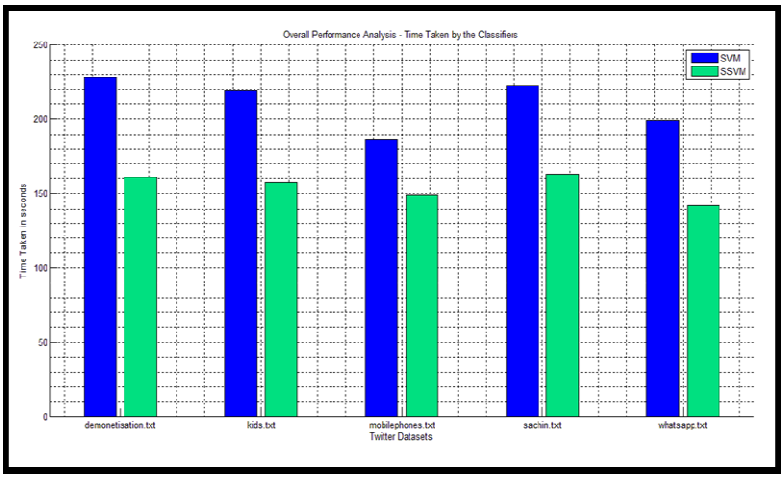

Figure 2: MATLAB Result Graph for Performance Analysis of the Classifiers in Terms of Execution Time (In Seconds)

CONCLUSIONS AND FUTURE WORKS

Text mining is one of the thrust research area in the field of data mining. The large-scale taking up of social media (SM) is one of the most pertinent technological and social trends in the history of the Internet. Cyberbullying is a considerably unrelenting version of traditional forms of bullying with negative effects on the victim. A cyberbully is capable enough to harass his/her victims before an entire online community. This research work aims in classifying tweets under various search criteria. Around 4556 tweets are collected and analyzed. The simplified support vector machine classifier is proposed and the performance metrics classification accuracy and execution time are taken for comparison. Simulations are carried out using MATLAB and the results portray that the proposed SSVM classifier attains better classification accuracy with reduced execution time.

REFERENCES

1. Aggarwal, A. Rajadesingan, P. Kumara guru, "Phish Ari: Automatic Real time Phishing Detection on Twitter", e Crime Researchers Summit (e Crime), pp. 1-12, 2012.
2. Kontostathis, K. Reynolds, A. Garron, L. Edwards, “Detecting cyberbullying: query terms and techniques,” Proceedings of the 5th Annual ACM Web Science Conference, pp. 195 – 204, 2013.
3. L. Kavanaugh, E. A. Fox, S. D. Sheetz, S. Yang, L.T. Li, D. J. Shoemaker, A. Natsev, L. Xie, "Social Media Use by Government: From the Routine to the Critical', Government Information Quarterly, vol. 29, no.4, pp.480-491, 2012.
4. A.Squicciarini, S. Rajtmajer, Y. Liu, C. Griffin, “Identification and characterization of cyberbullying dynamics in an online social network,” Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 280 – 285, 2015.
5. Salmivalli, "Bullying and the Peer Group: A Review", Aggression and Violent Behavior, vol. 15, no. 2, pp. 112-120, 2010.
6. Yang, R. Harkreader, J. Zhang, S. Shin, S. Gu, "Analyzing Spammers’ Social Networks for Fun and Profit: A Case Study of Cyber Criminal Ecosystem on Twitter", Proceedings of the International Conference on world wide web, pp. 71-80, 2012.
7. Nguyen, R. Gravel, D. Trieschnigg, T. Meder, “How Old Do You Think I Am?; A Study of Language and Age in Twitter,” Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, pp. 439 – 448, 2013.
8. G. R. Weir, F. Toolan, D. Smeed, "The Threats of Social Networking: Old Wine in New Bottles?", Information Security Technical Report, vol. 16, no. 2, pp. 38-43,2011.
9. G. S. O’Keeffe, K. Clarke-Pearson, "The Impact of Social Media on Children, Adolescents, and Families", Pediatrics, vol. 127, no. 4, pp. 800-804, 2011.
10. H. Gao, Y. Chen, K. Lee, D. Palsetia, A. N. Choudhary, "Towards Online Spam Filtering in Social Networks", Network and Distributed System Security Symposium Conference, 2012.
11. J. M. Xu, K. S. Jun, X. Zhu, A. Bellmore, "Learning from Bullying Traces in Social Media", Proceedings of the Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, pp.656-666, 2012.
12. K. Nalini, L. J. Sheela, “Classification of Tweets Using Text Classifier to Detect Cyber Bullying,” Advances in Intelligent Systems and Computing, vol. 338, pp. 637 – 645, 2015.
13. M. A. Al-garadi, K. D. Varathan, S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," Computers in Human Behavior, vol. 63, pp. 433 - 443, 2016.
14. M. Dadvar, D. Trieschnigg, R. Ordelman, F. Jong, “Improving Cyberbullying Detection with User Context,”

Advances in Information Retrieval, vol. 78, pp. 693 – 696, 2013.

15. M. G. Guadix, G. Gini, E. Calvete, "Stability of cyberbullying victimization among adolescents: Prevalence and association with bully–victim status and psychosocial adjustment," Computers in Human Behavior, vol. 53, pp. 140 - 148, 2016.
16. M. S. Rahman, T. K. Huang, H. V. Madhyastha, M. Faloutsos, "Efficient and Scalable Socware Detection in Online Social Networks", Proceedings of the 21st USENIX conference on Security symposium, pp. 663–678, 2012.
17. N. Shrivastava, A. Majumder and R. Rastogi, "Mining (Social) Network Graphs to Detect Random Link Attacks," IEEE International Conference on Data Engineering, pp. 486-495, 2008.
18. R. Forssell, "Exploring cyberbullying and face-to-face bullying in working life – Prevalence, targets and expressions," Computers in Human Behavior, vol. 58, pp. 454 - 460, 2016.
19. R. Hassanzadeh and R. Nayak, "A Rule-Based Hybrid Method for Anomaly Detection in Online-Social-Network Graphs," IEEE International Conference on Tools with Artificial Intelligence, pp. 351-357, 2013.
20. R. Yu, X. He, Y. Liu, "GLAD: Group Anomaly Detection in Social Media Analysis", ACM Transactions on Knowledge Discovery from Data, vol. 10, no.2, pp. 39 - 61,2015.
21. S. Yardi, D. Romero, G. Schoenebeck, "Detecting Spam in a Twitter Network", First Monday, vol. 15, No. 1, 2008.
22. V. S. Chavan, S. S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2354-2358, 2015.
23. Y. Chen, Y. Zhou, S. Zhu, H. Xu, “Detecting Offensive Language in Social Media to Protect Adolescent Online Safety,” Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, pp. 71 – 80, 2012.