AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES

ISSN:1991-8178 EISSN: 2309-8414
Journal home page: www.ajbasweb.com

An Analysis on the Hateful Contents Detection Techniques on Social Media

Maw Maw[1], Vimala A/P Balakrishnan [2]

[1]University of Malaya, Department of Information System, Faculty of Computer Science and Information Technology, Box.50603, Kuala Lumpur, Malaysia

[2]University of Malaya, Department of Information System, Faculty of Computer Science and Information Technology, Box.50603, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received X X 201X
Received in revised form X X X 201X
Accepted X X 201X

Keywords:
Hate speech, Offensive language, Online Detecting techniques

ABSTRACT

Background: Detecting abusive contents on social media becomes a broad research area along with the popularity of social media. **Objective:** This paper mainly aims to understand the different techniques applied within the scope of detecting the use of hateful language on social media, their strengths and challenges to provide the future researchers and practitioners with a solid and concrete reference in this research area. **Methodology:** In this paper, we analyzed previous researches done in the domain of hateful language detection in the social media. We selected relevant published journal articles and conference papers from 2010 to 2015.**Results:** We observed that Support Vector Machine (SVM) algorithm is the most frequently applied for classification. Data ambiguity problem and classification of sarcastic sentences are identified as the challenges for the researchers in this area of research **Conclusion:** The future researchers should be paid attention on the identified challenges.

To Cite This Article: Names of authors, Paper title. ***Aust. J. Basic & Appl. Sci.,*** *7(13): x-x, 2014*

1. Introduction

Billions of internet users from different countries with different languages upload their discussions and opinions on social media daily. However, it might lead to the dispute and hatred if they use hate speech which affects a specific group of people in a bad way. In (Chen, Zhou, Zhu, & Xu, 2012) and (Warner & Hirschberg, 2012), hate speech is defined as any kind of expression which deviates the law and which discredits a person or a group based on race, skin color, ethnicity, gender, sexual orientation, nationality and religion.

Though there are several reasons why the use of hateful language on online community needs to be detected, most of the previous researchers had common opinion upon the purpose of doing research in the area of interest. Poor content quality which is a mixture of abusive, malicious and bullied contents on social media give users bad online experience and lead to the severe problems in outside community(Sood, Churchill, & Antin, 2012). Manual checking and monitoring is the most flawless detector(Ravi, 2012) but it needs time, energy and money(Ismail & Bchir, 2015; Singh, 2015). Also, due to the unstructured text format, lack of specific labeled corpus(Chen et al., 2012) and technical flaws of a certain applied technique, researchers are competitively exploring better techniques to detect online hateful languages.

Detection of hateful language strategies fundamentally includes 1) preprocessing 2) feature extraction 3) feature selection and 4) classification (Chen et al., 2012; Ravi, 2012). **Data**

Corresponding Author:. Clearly indicate who will handle correspondence at all stages of refereeing and publication process. **Ensure that** Name of University, Name of Department, Name of Faculty, Box.3030. City. Country. **Phone numbers (with country and area code) are provided in addition to the e-mail address and the complete postal address**

AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES

ISSN:1991-8178 EISSN: 2309-8414
Journal home page: www.ajbasweb.com

preprocessing in other words is data cleaning process which transforms the raw data into the machine readable format by removing non-printable characters, special characters and html tags, reducing the duplicate words and labeling the data. **Feature extraction** is reducing of redundant features and dimensionality whereas **feature selection** techniques are applied to reduce the processing time, to get data which are more comprehensive for further processes, and to improve the system performance. Finally, **classification** techniques are applied to distinguish the desired results from undesired ones through preprocessed data sets based on the selected features(Chen et al., 2012; Ismail & Bchir, 2015).

In this paper, we will investigate the various techniques applied in each of the steps and will discuss the most applicable techniques. Therefore, this paper aims to identify the detection tools and techniques that have been applied for identifying hateful terms involvement on social media, their strengths and challenges.

2. Related Works

One of the problems in text classification is lack of labeled dataset in a specific domain. Reynolds et al.(Reynolds, Kontostathis, & Edwards, 2011) developed their own dataset for detection of cyberbullying in 2011. By applying language-based approach, the bully contents can be detected till 78.5%. As the social networks become the virtual life for the people, works on the detection of cyberbullying, hidden groups and insults have increased. In 2012, Fu et.al(Fu, Peng, Kuo, & Lee, 2012) found the way to detect the hidden community on Facebook using the topic identification module. Due to the noisy nature of the data, the accuracy of the classification are usually affected. To overcome this problem, an unsupervised possibilistic based local approach for automatic insult detection in the comments of social networks was proposed(Fu et al., 2012). Based on the experiments, their approach reduced the noise involvement in feature space and yielded the optimal results with high accuracy. Another remarkable research is detecting hate speech in web forums and blogs(Gitari, Zuping, Damien, & Long, 2015) which was focusing not only on racist speech but on the hate speech of religion and nationality as well. The authors developed their own lexicon of hate speech and detected hate speech in three levels: no hate, weakly hate and strong hate, by applying rule learning approach.

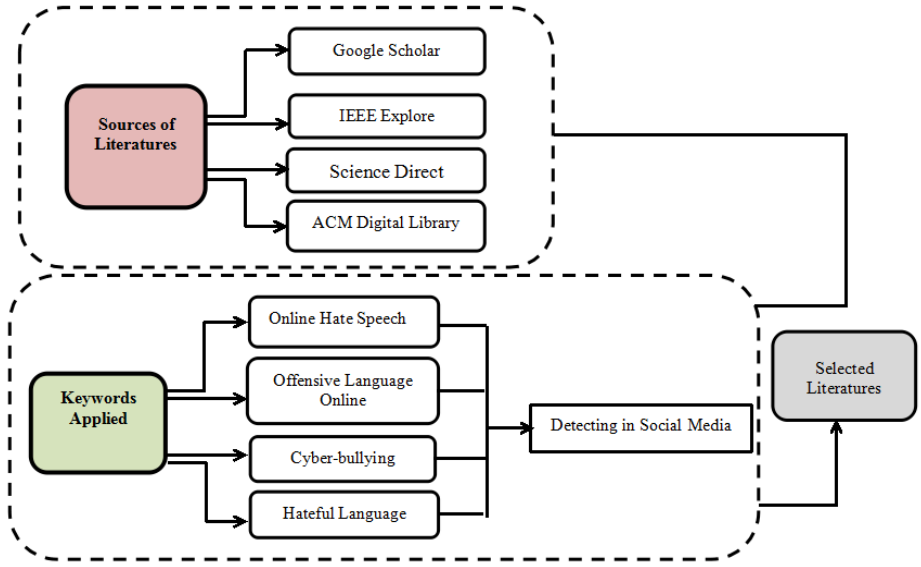

Figure 1: Illustration of steps of literature selection

Corresponding Author:. Clearly indicate who will handle correspondence at all stages of refereeing and publication process. **Ensure that** Name of University, Name of Department, Name of Faculty, Box.3030. City. Country. **Phone numbers (with country and area code) are provided in addition to the e-mail address and the complete postal address**

AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES

ISSN:1991-8178 EISSN: 2309-8414
Journal home page: www.ajbasweb.com

To be specific, four keywords were used as shown in the figure above. The term "cyber-bullying" was also included as it is defined as insulting or attacking others by use of awful language(Kansara & Shekokar, 2015). Approximately, 35 relevant literatures including journal articles and conference proceedings were found, and these were filtered to obtain the most relevant studies based on a set of criteria. Only articles written in English were included, those which context do not exactly fall under our domain, were removed and resulting in seven journal articles and seven conference proceedings.

4. Results and Discussion

By analyzing the previous efforts, we found that most of the detection techniques are based on two approaches: supervised learning and unsupervised learning, however unsupervised approaches have been applied very rarely. Table 1 summarizes the techniques in each step applied by the previous literatures.

Table 1: Summary of the Proposed Systems of Selected Literatures

No.	Year	Summary of the Proposed System	Reference
1	2011	A model which detects cyberbullying by creating own labeled dataset by applying language-based method of detecting cyberbullying.	(Reynolds et al., 2011)
2	2012	A machine learning approach which detects inappropriate contents such as profanity, insults and objects of the insults automatically.	(Sood et al., 2012)
3	2012	Their approach includes four main steps: preprocessing, feature extraction, feature selection and classification. For classification, they used 4 classifiers including, SVM, Naïve Bayes Multinomial, Random Forest and AdaBoostM1. But they discussed only for the classifier with the highest accuracy.	(Ravi, 2012)
4	2012	A Lexical Syntactic Feature (LSF) architecture which detects the involvement of inhumane, profane and offensive terms. The system contains two components: sentence offensiveness prediction and user offensiveness estimation. Former includes constructing lexical feature, syntactic feature and generation of sentence offensiveness value. Latter includes sentence offensiveness aggregation, additional features extracted from user's language profile which is based on style features, structural features and content specific feature.	(Chen et al., 2012)
5	2012	A system which detects the hate speech on online by categorizing seven groups which are related to race and nationality by adapting template-based strategy of previous researcher from 1994.	(Warner & Hirschberg, 2012)
6	2012	A hierarchical approach that exploits the co-occurrence of vulgar language via statistical topic modeling techniques and detects profane language with automatically generated features using a machine learning framework. They explored the predictive value of highly expressive topical features and reliable lexical features and combined them into single compact feature space.	(Xiang, Fan, Wang, Hong, & Rose, 2012)
7	2012	A flame detector model which retrieve the written notes of the users on social networking sites and detect the flaming words and calculate the intensity level of those words.	(Shukla, Singh, Parande, Khare, & Pandey, 2012)
8	2013	An improved cyberbullying system which classifies the users' comments on YouTube using content-based, cyberbullying-specific and user-based features by applying support vector machine.	(Dadvar, Trieschnigg, Ordelman, & de Jong, 2013)
9	2015	An automatic flame detection method which extracts features at different conceptual levels and applies multi-level classification for flame detection.	(Razavi, Inkpen, Uritsky, & Matwin, 2010)
10	2015	A framework to detect abusive text messages or images on the social network by applying SVM and Naïve Bayes classifiers.	(Kansara & Shekokar, 2015)

Corresponding Author:. Clearly indicate who will handle correspondence at all stages of refereeing and publication process. **Ensure that** Name of University, Name of Department, Name of Faculty, Box.3030. City. Country. **Phone numbers (with country and area code) are provided in addition to the e-mail address and the complete postal address**

11	2015	A system which work through Soft Text Classifier approach using various machine learning algorithms. It is type of a screening mechanism which alerts the users about the presence of profanity and insults. The messages are also labeled according to the subject matter.	(Singh, 2015)
12	2015	A two-step method to detect hate speech using Continuous Bag of Word (CBOW) neural language model.	(Djuric et al., 2015)
13	2015	A lexicon-based approach (classifier) to detect hate speech using semantic and subjectivity features.	(Gitari et al., 2015)
14	2015	A novel approach for automatic detection of offensive comments on social network based on local multi-classifier fusion method.	(Ismail & Bchir, 2015)

4.1 Preprocessing Techniques

Cleaning the data before they are fed into the classifier is an essential task. We identified that automatic pre-processing software including Regex(Ravi, 2012), Hadoop(Xiang et al., 2012),WordNet corpus and spell-correction algorithm(Chen et al., 2012) have applied for parsing, checking for grammar and spelling mistakes, stemming, removing symbols or unwanted characters and excluding duplication. Table 2 shows the preprocessing techniques applied in the selected studies.

Table 2: Preprocessing Techniques Applied in Selected Literatures

No.	Preprocessing Techniques	Reference
1.	Bootstrapping Method	(Gitari et al., 2015; Xiang et al., 2012)
2.	Customized Ready-to-use Tool	(Chen et al., 2012; Gitari et al., 2015; Kansara & Shekokar, 2015; Ravi, 2012)
3.	Spell-correction Algorithm	(Chen et al., 2012; Kansara & Shekokar, 2015)
4.	Crowdsourcing service	(Reynolds et al., 2011; Shukla et al., 2012; Sood et al., 2012)
5.	Manual Preprocessing	(Dadvar et al., 2013; Ismail & Bchir, 2015; Warner & Hirschberg, 2012)
6.	Not specifically discussed	(Djuric et al., 2015; Razavi et al., 2010; Singh, 2015)

Labeling the data with the aid of human resource can be an optimal solution as subjectivity analysis can be done best by the human being(Ravi, 2012). There are many good points of using crowdsourcing such as saving the time and internal resources, and scalability when working on large amount of dataset. Sood et al.(Sood et al., 2012) discussed that crowdsourcing service is suitable for analyzing texts and coding the contents with high efficiency. On the other hand, it cannot be cost effective when a large amount of data is to be handled. One advantage of this task is anonymity of the coders to the requestors (Reynolds et al., 2011), therefore the results will be less biased.

4.2 Feature Extraction Techniques

Feature extraction techniques are applied for reducing redundant features and dimensionality. Table 3 summarizes the feature extraction techniques applied by the studies.

Table 3: Feature Extraction Techniques Applied in Selected Literatures

No.	Feature Extraction Techniques	Reference
1.	Local Binary Pattern (LBP)	(Kansara & Shekokar, 2015)
2.	Bag-of-Words (BoW)	(Chen et al., 2012; Kansara & Shekokar, 2015; Sood et al., 2012)
3.	Bag-of-Visual-Words (BoVW)	(Kansara & Shekokar, 2015)
4.	Term Frequency Inverse Document Frequency (TF-IDF)	(Ismail & Bchir, 2015; Singh, 2015)
5.	N-gram	(Chen et al., 2012; Ravi, 2012; Singh, 2015; Sood et al., 2012; Warner & Hirschberg, 2012)
6.	Skip gram	(Kansara & Shekokar, 2015; Singh, 2015)
7.	K-means Clustering	(Kansara & Shekokar, 2015)
8.	Customized Ready-to-use Toolkit	(Ravi, 2012; Xiang et al., 2012)
9.	Not specifically discussed	(Dadvar et al., 2013; Shukla et al., 2012)

Bag-of-Words (BoW) approach detects offensive sentences regardless of grammar mistakes and word order and this approach has the highest recall rate(Chen et al., 2012). Many researches in natural language processing applied BoW due to its common use(Kansara & Shekokar, 2015). Though it is one of the most popular text categorization approaches, Guang et.al(Xiang et al., 2012) found the fact that BoW did not

work properly for the detection of profane tweets due to the noisy nature of tweets. Many of previous researches applied BoW approach but some terms and usages cannot be detected well if the offensive words and terms are replaced by other terms which does not seem to be offensive although the sentence means in negative. That leads to the sparsity and overfitting problem(Djuric et al., 2015). In (Singh, 2015), the author highlighted that using BoW approach gives a high-false positive rate.

N-gram approach is the one which detect the offensive sentences based on the n words of sequence(Chen et al., 2012). Though(Warner & Hirschberg, 2012)discussed the effectiveness of n-gram approach in feature extraction, the bigram and trigram methods decrease the efficiency of the classifier and they do not work well in finding related words which are far away from each other(Chen et al., 2012; Warner & Hirschberg, 2012). This problem can be tackled by increasing the number of n, but the system processing time will be longer(Chen et al., 2012). One advantage is that the words which are previously neglected but important can be added to the attribute list with the n-gram approach(Ravi, 2012).

The authors did not discuss the reasons why the customized software toolkit such as WEKA was chosen to apply in their researches to perform preprocessing steps and classification tasks. In our opinion, the main reasons might be the easiness for the use, availability of all necessary software in on place and popularity of the tools.

4.3 Feature Selection Techniques

Feature selection techniques are necessary to reduce the redundant features from the dataset and for selecting the most relevant subset of the feature for more accurate classification results(Ravi, 2012). In Table 4, the feature selection techniques applied in the selected literatures are described.

Table 4: Feature Selection Techniques Applied in Selected Literatures

No.	Feature Selection Techniques	Reference
1.	Chi-squared Test	(Ravi, 2012; Singh, 2015)
2.	Latent Dirichlet Allocation Algorithm	(Xiang et al., 2012)
2.	Wrapper Supervised Feature Selection Algorithm	(Razavi et al., 2010)
3.	Not specifically discussed	(Chen et al., 2012; Dadvar et al., 2013; Djuric et al., 2015; Ismail & Bchir, 2015; Kansara & Shekokar, 2015; Reynolds et al., 2011; Shukla et al., 2012; Sood et al., 2012; Warner & Hirschberg, 2012)

As large feature space affect processing time and system performance, feature selection techniques are applied to reduce the dimension of the feature set. As shown in Table 4, the chi-squared test technique was applied by two studies while latent Dirichlet allocation algorithm and wrapper supervised algorithm were used by one study each.

The chi-squared test is a statistical tool which is used to find the best features from the large feature set. Prashant (Ravi, 2012), stated that using feature selection tool reduce memory consumption, prevent overfitting and seek more accurate attributes. Chi-squared test tool measures the reliance of two certain variable based on the value(Singh, 2015).

4.4 Classification Techniques

For classification process, supervised learning approaches includes different types of decision tree algorithms(Ravi, 2012; Razavi et al., 2010; Reynolds et al., 2011), Naïve Bayes algorithm(Ravi, 2012; Razavi et al., 2010; Singh, 2015), and Support Vector Machine (SVM) (Chen et al., 2012; Dadvar et al., 2013; Kansara & Shekokar, 2015; Razavi et al., 2010; Singh, 2015; Sood et al., 2012). One study applied a semi-supervised approach in(Xiang et al., 2012) and one more study proposed an unsupervised learning approach (Ismail & Bchir, 2015), others such as proposed novel approaches for classification. Though novel techniques and existing popular techniques were mostly applied, two studies referred back the previous classification methods. (Warner & Hirschberg, 2012) applied template-based strategy which was proposed by other researcher over 20 years ago and (Djuric et al., 2015) applied an unsupervised algorithm named pragraph2vec which was a novel approach of previous researcher proposed in 2014.

Table 5: Classification Techniques Applied in Selected Literatures

No.	Classification Techniques	Reference
1.	Support Vector Machine (SVM)	(Chen et al., 2012; Dadvar et al., 2013; Kansara & Shekokar, 2015; Razavi et al., 2010; Singh, 2015; Sood et al., 2012)
2.	Naïve Bayes (NB)	(Chen et al., 2012; Kansara & Shekokar, 2015;

		Razavi et al., 2010; Singh, 2015)
3.	Regular Expression Pattern Matching Algorithm	(Ismail & Bchir, 2015)
4.	Rule-based Approach	(Razavi et al., 2010)
5.	Decision Tree Approach	(Razavi et al., 2010)
6.	K-Nearest Neighbor Classifier	(Dadvar et al., 2013)
7.	Novel Technique	(Chen et al., 2012; Dadvar et al., 2013; Gitari et al., 2015; Xiang et al., 2012)
8.	Referred back to previous technique	(Djuric et al., 2015; Warner & Hirschberg, 2012)

Surprisingly, Support Vector Machine (SVM) was the most frequently applied classification algorithms as we identified 6 out of 14 studies to have applied this (see **Table 5**). It is a supervised learning algorithm and is traditionally used for the classification tasks. SVM works by enlarging the margin of separation of the data than the feature similarity(Ravi, 2012; Singh, 2015). The researchers agreed on the fact that SVM is highly robust and it could avoid the overfitting problem. SVM was frequently applied due to its efficiency on large volume of data and high performance and it could reduce the error rates(Singh, 2015; Sood et al., 2012).

Though Naïve Bayes (NB) classification techniques are not as frequently applied as SVM, we identified that four of the researches applied in their researches. In(Shukla et al., 2012), the authors discussed that NB methods are simple yet it yields the high performance for complex classifications and they work well even under the limited resources. NB methods are applied to overcome the data sparsity problem(Razavi et al., 2010).

5. Challenges

When a large amount of data volume is needed to handle, there is a possibility of data sparsity problem, meaning some data points are missing to observe and hence it could affect the efficiency of the system(Singh, 2015). But this can be solved by applying feature selection techniques. Mocking sentences which use the non-offensive words are actually intentionally offensive and such sentences are overlooked and cannot be identified as offensive in word-based detection system(Chen et al., 2012). Another challenge in the text classification is sarcastic sentences(Singh, 2015). More enhanced techniques are needed to perform a deep analysis of the meaning of the sentences. Due to the nature of natural language, a word might have different meanings and usages which can lead to the misassumption of the original sentence. This problem is named as ambiguity problem(Chen et al., 2012).

Another challenge is the lack of resources of hateful terms in different speaking languages. Though there is the resource with English language, it is still a challenge to apply in other language to detect the use of hateful terms.

6. Conclusion and Future Work

In this paper, we investigated the different techniques and methods employed in the each steps of detection of hateful language usage on social media. Moreover, we presented a brief discussion upon the strengths and challenges of most frequently applied techniques. We observed that unsupervised machine learning techniques are less frequently applied in the field of detecting hateful language. Additionally, we found out that Support Vector Machine (SVM) is the most applied classification technique in this area of research because of its high performance level. We identified the challenges such as data sparsity problem, ambiguity problems and classification of sarcastic sentences which the future researchers should be aware of.

As a future work, we would like to study in depth of the challenges and difficulties and ways to tackle those obstacles. As we mentioned earlier, we also would like to investigate why unsupervised learning approaches are less applicable in this area of research. Also, we would like to find out the techniques to detect hateful sarcasm which is still an open problem for the researchers.

7. Acknowledgement

The authors would like to thank and acknowledge the support provided by University Malaya, under research grant reference number: RP28A-14AET.

Reference

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). *Detecting offensive language in social media to protect adolescent online safety.* Paper presented at the Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom).

Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. In *Advances in Information Retrieval* (pp. 693-696): Springer.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). *Hate Speech Detection with Comment Embeddings.* Paper presented at the Proceedings of the 24th International Conference on World Wide Web Companion.

Fu, M.-H., Peng, C.-H., Kuo, Y.-H., & Lee, K.-R. (2012). *Hidden community detection based on microblog by opinion-consistent analysis.* Paper presented at the Information Society (i-Society), 2012 International Conference on.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A Lexicon-based Approach for Hate Speech Detection.

Ismail, M. M. B., & Bchir, O. (2015). Insult detection in social network comments using possibilistic based fusion approach. In *Computer and Information Science* (pp. 15-25): Springer.

Kansara, K. B., & Shekokar, N. M. (2015). A Framework for Cyberbullying Detection in Social Network.

Ravi, P. (2012). Detecting Insults in Social Commentary.

Razavi, A. H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence* (pp. 16-27): Springer.

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). *Using machine learning to detect cyberbullying.* Paper presented at the Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on.

Shukla, S. S. P., Singh, S. P., Parande, N. S., Khare, A., & Pandey, N. K. (2012). *Flame Detector Model: A Prototype for Detecting Flames in Social Networking Sites.* Paper presented at the Computer Modelling and Simulation (UKSim), 2012 UKSim 14th International Conference on.

Singh, S., Nakhare, S. , Nair, K. , Shetty, R. (2015). A System to Detect Inappropriate Messages in Online Social Networks. *World Academy of Science, Engineering and Technology, International Science Index, Mechanical and Mechatronics Engineering*.

Sood, S. O., Churchill, E. F., & Antin, J. (2012). Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology, 63*(2), 270-285.

Warner, W., & Hirschberg, J. (2012). *Detecting hate speech on the world wide web.* Paper presented at the Proceedings of the Second Workshop on Language in Social Media.

Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012). *Detecting offensive tweets via topical feature discovery over a large scale twitter corpus.* Paper presented at the Proceedings of the 21st ACM international conference on Information and knowledge management.