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

[1]Maw Maw and [2]Vimala Balakrishnan

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

Address For Correspondence:
Maw Maw, Department of information System, Faculty of Computer Science and Information System, University of Malaya, Box 50603. Lembah Pantai, Kuala Lumpur, Malaysia.
E-mail: hanifa@um.edu.my +60-3-7967 6300

ARTICLE INFO

Article history:
Received 10 December 2015
Accepted 22 January 2016
Available online 30 January 2016

Keywords:
Hate speech, Offensive language, Online Detecting techniques

ABSTRACT

Background: Detecting hateful contents on social media becomes a broad and important research area along with the popularity of social media. **Objective:** This paper aims primarily 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 a solid and concrete reference to future researchers and practitioners. **Methodology:** In this paper, we investigated previous researches done in the domain of hateful contents detection on social media. We selected relevant published journal articles and conference proceedings from 2010 to 2015. **Results:** We observed that Support Vector Machine (SVM) algorithm is the most frequently applied for text classification. Data ambiguity problem, classification of sarcastic sentences and lack of necessary resources are identified as the difficulties for researchers in detecting the use of hateful contents. **Conclusion:** Future researchers must pay more attention on developing techniques to perform a deep analysis of sentences in order to detect the hateful contents.

INTRODUCTION

Billions of internet users from different countries upload their discussions and opinions on social media daily with different languages. However, it might lead to the dispute and hatred if they use hateful terms and usages which affect a person or a specific group of people in a bad way. Hateful contents are malicious contents such as hate speech, use of abusive terms and offensive languages, and cyber-bullies. 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 of people based on race, skin color, ethnicity, gender, sexual orientation, nationality and religion.

Social media becomes a virtual life for the people and a spot to throw out feelings, beliefs and opinions. Concurrently, it also becomes a place where hateful contents can be seen most frequently. Different types of campaign are being promoted to reduce hateful contents involvement on social media. Similarly, researchers in computer science field have been finding the best techniques to detect automatically of the hateful contents on social media.

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 opinions upon the purpose of doing researches in the area of interest. Poor content quality, a mixture of abusive, malicious and bullied contents, on social media give users bad online experience and might lead to severe problems to outside community (Sood, Churchill, & Antin,

To Cite This Article: Maw Maw and Vimala Balakrishnan., An Analysis on the Hateful Contents Detection Techniques on Social Media. ***Aust. J. Basic & Appl. Sci.,*** *10(3): 25-31, 2016*

2012). Manual checking and monitoring by human beings are the most flawless detection (Ravi, 2012) but they need time, energy and money (Ismail & Bchir, 2015; Singh, 2015). In addition, unstructured text format, lack of specific labeled corpus (Chen *et al*., 2012) and technical flaws of a particular technique make researchers explore competitively to obtain better techniques in detecting online hateful languages.

Detection of hateful contents process fundamentally includes: preprocessing, feature extraction, feature selection, and classification (Chen *et al*., 2012; Ravi, 2012). Data preprocessing is process of data cleaning which transforms the raw data into 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 processing time, to get data which are more comprehensive for further processes, and to improve system performance. Finally, classification techniques are applied to distinguish desired results from undesired ones through preprocessed datasets based on selected features (Chen *et al*., 2012; Ismail & Bchir, 2015).

Though detecting hateful contents on social media becomes a noticeable problem for researchers, no previous studies have conducted upon the techniques applied in this area of research. Therefore, it is worth to investigate existing techniques applied in detection of hateful contents on social media, and to identity advantages and disadvantages of those techniques. We believe our analysis will be beneficial to the future researchers in same field of research.

In this paper, we will investigate the various techniques applied in each step and will discuss most applicable techniques and challenges which are currently faced by researchers. Therefore, this paper aims to identify the detection tools and techniques that have been applied for identifying hateful contents involvement on social media, their strengths and challenges.

Related Works:

One problem in text classification is lack of labeled datasets in a specific domain. Reynolds *et al*. (2011) developed a dataset of their own for detection of cyberbullying in 2011. By applying language-based approach, bully contents can be detected till 78.5%. As social networks become a virtual life for the people, researches on detection of cyberbullying, hidden groups and insults have increased. In 2012, Fu et.al (2012) found a way to detect hidden communities on Facebook using topic identification module. Owing to noisy nature of data, accuracy of classification is 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 noise involvement in feature space and produced optimal results with high accuracy. Another remarkable research is detecting hate speech on web forums and blogs (Gitari, Zuping, Damien, & Long, 2015) which was focusing not only on racist speech but on hate speech regarding 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.

Research Methodology:

As this paper is an analysis of previous studies, we collected the literature in the area of detecting use of hateful language on every type of social media which were published within past five years (i.e. 2010-2015). We believe this analysis will offer a useful collection of techniques and tools for other researchers and practitioners. In Figure 1, keywords applied in the selection of relevant literature and sources are listed.

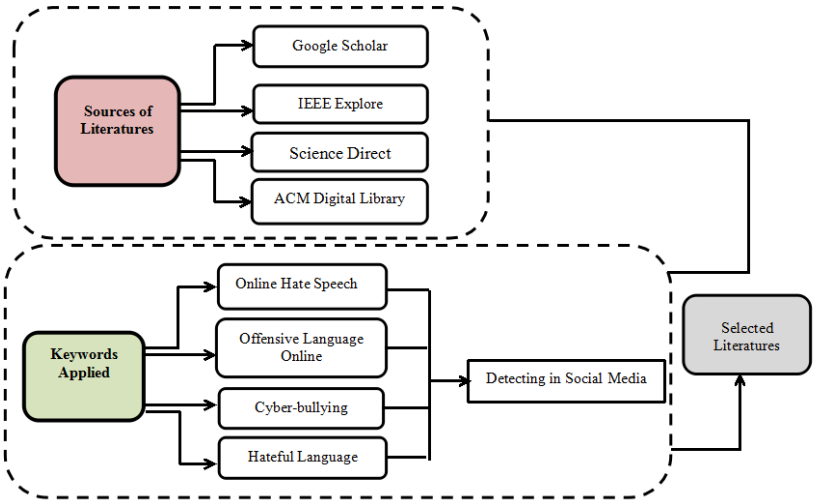

Fig. 1: Illustration of steps of literature selection

To be specific, four keywords were used. 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 literature, both journal articles and conference proceedings, were found from online databases listed in the figure, and these were filtered to obtain the most relevant studies based on a set of criteria. Only articles written in English were included, the context do not exactly fall under our domain were removed and resulting in seven journal articles and seven conference proceedings.

RESULTS AND DISCUSSION

By determining the previous efforts, we found that most detection techniques are based on two approaches: supervised and unsupervised learning; however unsupervised approaches have applied very rarely in this domain. Table 1 summarizes the brief description of proposed systems in selected studies.

Table 1: Summary of proposed systems from selected literature

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)
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 data before they are fed into 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 literature

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 aid of human resource can be an optimal solution as subjectivity analysis can be done best by human beings (Ravi, 2012). Crowdsourcing is data labeling service provided by trained people from different countries. There are many good points of using crowdsourcing such as saving time and internal resources, and scalability when working on large amount of datasets. 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 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 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 most popular text categorization approaches, Xiang *et al*., (2012) found the fact that BoW did not work properly for detection of profane tweets owing to the noisy nature of tweets. Many of previous researches used BoW approach but some terms and usages cannot be detected well if offensive words and terms are replaced by other terms which do not seem to be offensive although the sentence means in negative. That leads to sparsity and overfitting problem (Djuric *et al*., 2015). In (Singh, 2015), the author highlighted that using BoW approach gives a high-false positive rate. Hence, BoW approach has more weaknesses than strengths.

N-gram approach detects offensive sentences based on n words of sequence (Chen *et al*., 2012). Though Warner and Hirschberg, (2012) discussed the effectiveness of n-gram approach in feature extraction process, the bigram and trigram methods decrease the efficiency of 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 number of n, but system processing time will be longer (Chen *et al*., 2012). One advantage is that words which are previously neglected but important can be added to attribute list with n-gram approach (Ravi, 2012).

Though many previous researches have used customized software toolkits such as WEKA, they did not discuss why those toolkits were chosen to use in performing preprocessing steps and classification tasks. In our opinion, main reasons might be easiness for use, availability of all necessary software in one place, and popularity of the tools.

Feature Selection Techniques:

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

Table 4: Feature selection techniques applied in selected literature

No.	Feature Selection Techniques	Reference
1.	Chi-squared Test	(Ravi, 2012; Singh, 2015)
2.	Latent Dirichl*et al*location 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 affects processing time and system performance, feature selection techniques are used to reduce dimensions of feature set. As shown in Table 4, the chi-squared test technique was applied by two studies while latent Dirichl*et al*location 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 a 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), and novel approaches for classification. Though novel techniques and existing popular techniques were mostly applied, two studies referred back previous classification methods. Warner and 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 most frequently applied classification algorithm 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 classification tasks. SVM works by enlarging the margin of separation of data than feature similarity (Ravi, 2012; Singh, 2015). Researchers agreed on fact that SVM is highly robust and it could avoid overfitting problem. SVM was frequently applied due to its efficiency on large volume of data and high performance and it could reduce error rate (Singh, 2015; Sood *et al*., 2012).

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

Challenges:

When a large amount of data volume is needed to handle, it is possible to encounter data sparsity problem, meaning some data points are missing to observe and hence it could affect the efficiency of system (Singh, 2015). But this can be resolved by applying feature selection techniques. Mocking sentences which use non-offensive words are actually intentionally offensive but they are overlooked and cannot be identified as offensive in word-based detection system (Chen *et al*., 2012). Another challenge in text classification is sarcastic sentences (Singh, 2015). More enhanced techniques are needed to perform a deep analysis of meaning of sentences. Due to the nature of natural language, a word might have more than one meaning and might have

different usages which lead to misassumption of original sentence. This problem is named as ambiguity problem (Chen *et al*., 2012).

Another challenge is lack of resources of hateful terms in different speaking languages. Though there are resources with English language, it is still a challenge to detect the use of hateful terms in another language rather than English.

Conclusion and Future Work:

In this paper, we investigated different techniques and methods employed in each step of detection of hateful language on social media. Moreover, we presented a brief discussion upon strengths of most frequently applied techniques and challenges which have been faced by researchers. We observed that unsupervised machine learning techniques were less frequently applied in the field of detecting hateful contents on social media. Additionally, we investigated that Support Vector Machine (SVM) is the most applied classification technique in this area of research because of its high performance level. We identified important challenges: data sparsity problem, ambiguity problems, classification of sarcastic sentences, and lack of necessary language resources, to pay more attention by the future researchers. As future works, we would like to study in depth of identified challenges and ways to tackle those obstacles. As we mentioned earlier, we also would like to investigate why unsupervised learning approaches were less applicable in this area of research. Also, we would like to look for enhanced techniques to detect hateful sarcasm which is still an open problem for researchers.

ACKNOWLEDGEMENT

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

REFERENCES

Chen, Y., Y. Zhou, S. Zhu and H. Xu, 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., D. Trieschnigg, R. Ordelman and F. de Jong, 2013. Improving cyberbullying detection with user context. In *Advances in Information Retrieval* (pp. 693-696): Springer.

Djuric, N., J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic and N. Bhamidipati, 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., C.-H. Peng, Y.-H. Ku and K.-R. Lee, 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., Z. Zuping, H. Damien and J. Long, 2015. A Lexicon-based Approach for Hate Speech Detection.

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

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

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

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

Reynolds, K., A. Kontostathis and L. Edwards, 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., S.P. Singh, N.S. Parande, A. Khare and N.K. Pandey, 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., S. Nakhare, K. Nair, R. Shetty, 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., E.F. Churchill and J. Antin, 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., and J. Hirschberg, 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., B. Fan, L. Wang, J. Hong and C. Rose, 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.