IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

Cyber Bullying Detection on Social Media based on Denoising Auto-Encoder

Ruksar Fatima[1*], Umme Khadija[2]

[1]CSE Dept., KBNCE, Kalaburagi
[2]M. Tech Student KBNCE, Kalaburagi

[*]*Corresponding Author: viceprincipalkbnce@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 15/Sept/2018, Published: 30/Sept/2018

Abstract- As a signal of more and more distinguished on-line networking, cyberbullying has developed as a big issue harassing kids, adolescents and vernal grown-ups. Machine learning procedures build programmed recognition of harassing messages in web-based social networking doable, and this might build a solid and safe web-based social networking condition. During this important analysis zone, one basic issue is powerful and discriminative numerical portrayal learning of instant messages. During this paper, we tend to propose another portrayal learning strategy to handle this issue. Our technique named Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) is created by means that of linguistics enlargement of the notable profound learning model stacked denoising autoencoder. The linguistics enlargement includes of linguistics dropout commotion and meagerness limitations, wherever the linguistics dropout clamor is planned in sight of area learning and therefore the word inserting system. Our planned strategy will misuse the hid part structure of tormenting knowledge and soak up a full of life and discriminative portrayal of content.

Keywords- NLP, cyberbullying, Social Network, Mining, collaboration.

I. INTRODUCTION

Opinion extraction, sentiment mining, affect analysis, analysis of emotion, review analyzing or mining, etc. are numerous comparative names with somewhat unique errands and with slightly different tasks. However, they are presently considered within the shades of opinion mining or sentiment analysis. This can be considered as an important application of NLP (Natural Language Processing). Some of social media sites, includes social networking sites such as Facebook, MySpace, and Twitter; video sites such as YouTube; photo sharing such as Flicker, Photobucket, or Picasa; gaming sites and virtual worlds such as Kaneva, Club Penguin, Second Life, and the Sims; live casting such as Upstream or Twitch; instant messaging like Google talk, yahoo messenger or skype and blogs. Authors in and use reciprocally the terms Big Social Data and Social Big Data to refer about the data which is generated by social media. To show the tremendous measure of data that is generated by social networking, affirms that via Facebook (the most popular social media [11]), 10 million photos are getting uploaded almost every day. focus attention to highlight the aspect that more than 250 million tweets are sent by Twitter every day consistently as well, also 3000+ photos are getting uploaded across Flicker every minute consistently without overlooking above 150 million web blogs posted daily. The expansion of Social Big Data is extremely valuable in numerous fields such as sociology, human science, psychology, governmental issues, politics and the very important commercial area.

Nonetheless, social media over and above have some consequential side effects as cyberbullying, that might have intense unfavorable impact and can transform the life of people, especially the children, adolescent and teenagers. Cyberbullying detection can be administrated from social media as it can be particularized as a supervised learning problem. A classifier is initially trained on a cyberbullying corpus that is labeled with mark by humans, and then later the learned classifier is then used as a result to recognize & perceive bullying message. Three categories of information are commonly used inclusive of user demography, text, and social network features of cyberbullying detection. Since the text content substance is the most definitely dependable as well as reliable, our work here spotlights on the text-based content cyberbullying detection.

II.RELATED WORK

A. Survey on Sentiment Analysis

Paper [40], Nasukawa et al. demonstrates an approach to sentiment analysis in which they extract sentiments for particular subjects associated with its polarities i.e. negative or positive from a document, in spite of classifying the

whole document. Initial work on sentiment analysis focused on identifying polarity of reviews of product Opinions and movie reviews from IMDB (Internet Movie data base) at entire document level [7]. Later work handles sentiment analysis at sentence level. Recent studies are focus was shifted from sentence level to phrase-level [2] and short-text forms in response to the popularity of micro blogging services such as Twitter [6,1,3,5]. In, Ding et al. proposed an effective method for identifying semantic orientations of opinions expressed by reviewers on product features.

In sentiment analysis, Pandey et al. presents a system, which can extract micro messages relevant to any specific topic from a blogging service such as Twitter and then analyze the messages to determine sentiments they carry and to classify them as neutral, positive or negative.

B. Survey on social media big data

According to indication in, the Social Big Data, produced represent all the data as well as information generated through the social media. This data is then perceived by: the substantial volume, the commotion that can be propagated (spam) and the dynamic characteristic property (the frequent changes day by day). They can likewise be perceived by an arrangement set of connections or links (due to connections between users), a structure or frame which is not aligned pattern in nature (as a result of the length of messages which is required by particularly some kind of micro-blogging, the existence of spelling inaccuracies or other) and the absence of fulfillment (as a result of satisfying the user requirements for privacy of their data). The mentioned attributes of Social Big Data qualify them unconventionally discrete from other version data on which elementary data mining techniques are enforcedly applied. For mining such sort of data, various research issues, methodologies, approaches, procedures and techniques are derived.

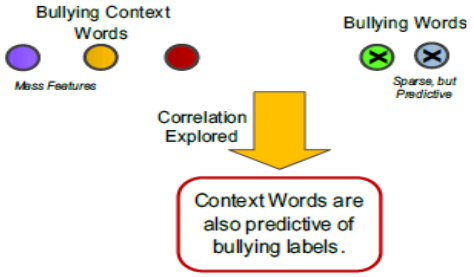
C. Survey on Effect of Social Media

Sentiment Analysis is used for the prediction of polarity of textual data into positive, negative and neutral classes. This textual data can be gathered from social media (e.g. twitter). Following Table 1[9] shows how social media creates its impact in various situations.

D. Survey on Cyberbullying

To originate a productive potent efficient practical cyberbullying identification model, we suggest expanding on discoveries within the psychology community. There are various reviews investigating the psychological dimensions of social collaborations that can be used to determine and analyze the cyber bullying risk elements, risk factors, risk aspects, which a model ought to consider. A large portion of the work in determining bullying among youth has concentrated on traditional bullying, conventional tormenting, or cyberbullying by the means of versatile mobile or visit chat-based scenes, e.g., [10]. Previous contributions have concentrated different aspects of bullying and cyberbullying in their studies, e.g., whether guardians' viewpoint of adolescents' online conduct is causational with teenagers' vulnerability to cyberbullying [10], probabilities of exploitation and victimization and passionate effect as well as emotional impact in light of age and gender, and measuring the correlation between seriousness of online hostility and the quantity of spooks or bullies included. While the outcomes about pervasiveness and determinants of cyberbullying change in the psychology literature, there are some critical patterns and regions of understanding among these outcomes.

III. SYSTEM ARCHITECTURE:



Algorithm:

- **Input:**
  - training examples $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)$
  - Test examples $\vec{x}_1^*, \vec{x}_2^*, \ldots, \vec{x}_k^*$
- **Parameters:**
  - C, C^*: parameters from OP(2)
  - num_+ : number of test examples to be assigned to class +
- **Output:** predicted labels of the test examples $y_1^*, y_2^*, \ldots, y_k^*$

$(\vec{\varpi}, b, \vec{\xi}, \_) = solve\_svm\_qp(\{(\vec{x}_1, y_1) \ldots (\vec{x}_n, y_n)\}, [\,], C, 0, 0)$;

Classify the test examples using $<\vec{\varpi}, b>$ The num_+ test examples with the highest value of $\vec{\varpi} * \vec{x}_j^* + b$ are assigned to the class $+(y_j^* = 1)$ the remaining test examples are assigned to class $-(y_j^* = -1)$

$C_-^* = 10^{-5}$;

$C_+^* = 10^{-5} * \frac{num_+}{k - num_+}$;

While $((C_-^* < C^*) \mid (C_+^* < C^*))$ {

$(\vec{\varpi}, b, \vec{\xi}, \vec{\xi}^*) = solve\_svm\_qp(\{(\vec{x}_1, y_1) \ldots (\vec{x}_n, y_n)\}, \{(\vec{x}_1^*, y_1^*) \ldots (\vec{x}_k^*, y_k^*)\}, C, C_-^*, C_+^*)$;

while $(\exists m, l : (y_m^* * y_l^* < 0) \& (\xi_m^* > 0) \& (\xi_l^* > 0) \& (\xi_m^* + \xi_l^* > 2))$ {

$y_m^* := -y_m^*$

$y_l^* := -y_l^*$

$(\vec{\varpi}, b, \vec{\xi}, \vec{\xi}^*) := solve\_svm\_qp(\{(\vec{x}_1, y_1) \ldots (\vec{x}_n, y_n)\}, \{(\vec{x}_1^*, y_1^*) \ldots (\vec{x}_k^*, y_k^*)\}, C, C_-^*, C_+^*)$;

}

$C_-^* = \min(C_-^* * 2, C^*)$,

$C_+^* = \min(C_+^* * 2, C^*)$,

}

return $(y_1^*, \ldots, y_k^*)$;

IV. IMPLEMENTATION

MODULES:

- OSN System Construction Module
- Construction of Bullying Feature Set
- Cyberbullying Detection.
- Semantic-Enhanced Marginalized Denoising Auto-Encoder.

MODULES DESCSRIPTION:

OSN System Construction Module

- In the first module, we develop the Online Social Networking (OSN) system module. We build up the system with the feature of Online Social Networking. Where, this module is used for new user registrations and after registrations the users can login with their authentication.
- Where after the existing users can send messages to privately and publicly, options are built. Users can also share post with others. The user can able to search the other user profiles and public posts. In this module users can also accept and send friend requests.
- With all the basic feature of Online Social Networking System modules is build up in the initial module, to prove and evaluate our system features.
- Construction of Bullying Feature Set:
- The bullying features play an important role and should be chosen properly. In the following, the steps for constructing bullying feature set Zb are given, in which the first layer and the other layers are addressed separately.
- For the first layer, expert knowledge and word embeddings are used. For the other layers, discriminative feature selection is conducted.
- In this module firstly, we build a list of words with negative affective, including swear words and dirty words. Then, we compare the word list with the BoW features of our own corpus, and regard the intersections as bullying features.
- Finally, the constructed bullying features are used to train the first layer in our proposed smSDA. It includes two parts: one is the original insulting seeds based on domain knowledge and the other is the extended bullying words via word embeddings
- Observe Attentively Over A Period Of Time.

Cyberbullying Detection:

In this module we propose the Semantic-enhanced Marginalized Stacked Denoising Auto-encoder (smSDA). In this module, we describe how to leverage it for cyberbullying detection. smSDA provides robust and discriminative representations The learned numerical representations can then be fed into our system.

In the new space, due to the captured feature correlation and semantic information, even trained in a small size of training corpus, is able to achieve a good performance on testing documents.

Based on word embeddings, bullying features can be extracted automatically. In addition, the possible limitation of expert knowledge can be alleviated by the use of word embedding

Block The Accounts:

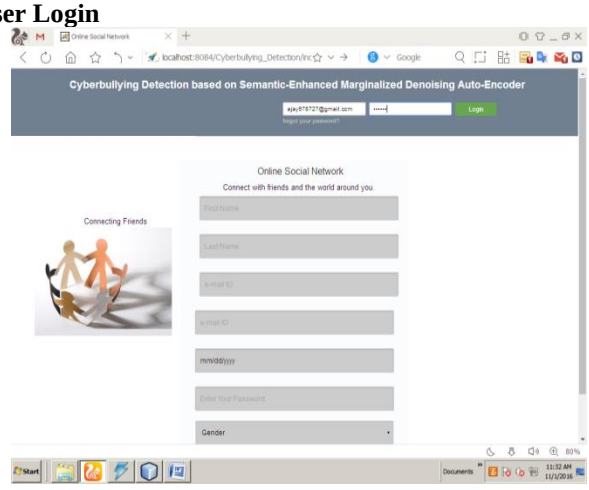
- **Abnormal user.**
- **Cyber- Crime user.**

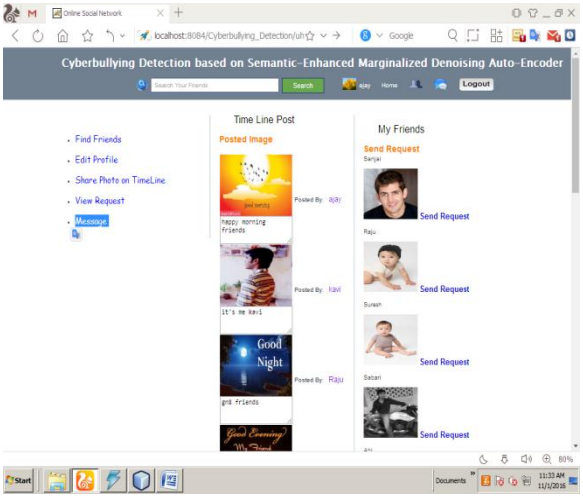
Semantic-Enhanced Marginalized Denoising Auto-Encoder:

- An automatic extraction of bullying words based on word embeddings is proposed so that the involved human labor can be reduced. During training of smSDA, we attempt to reconstruct bullying features from other normal words by discovering the latent structure, i.e. correlation, between bullying and normal words. The intuition behind this idea is that some bullying messages do not contain bullying words.
- The correlation information discovered by smSDA helps to reconstruct bullying features from normal words, and this in turn facilitates detection of bullying messages without containing bullying words. For example, there is a strong correlation between bullying word fuck and normal word off since they often occur together.
- If bullying messages do not contain such obvious bullying features, such as fuck is often misspelled as fck, the correlation may help to reconstruct the bullying features from normal ones so that the bullying message can be detected. It should be noted that introducing dropout noise has the effects of enlarging the size of the dataset, including training data size, which helps alleviate the data sparsity problem.

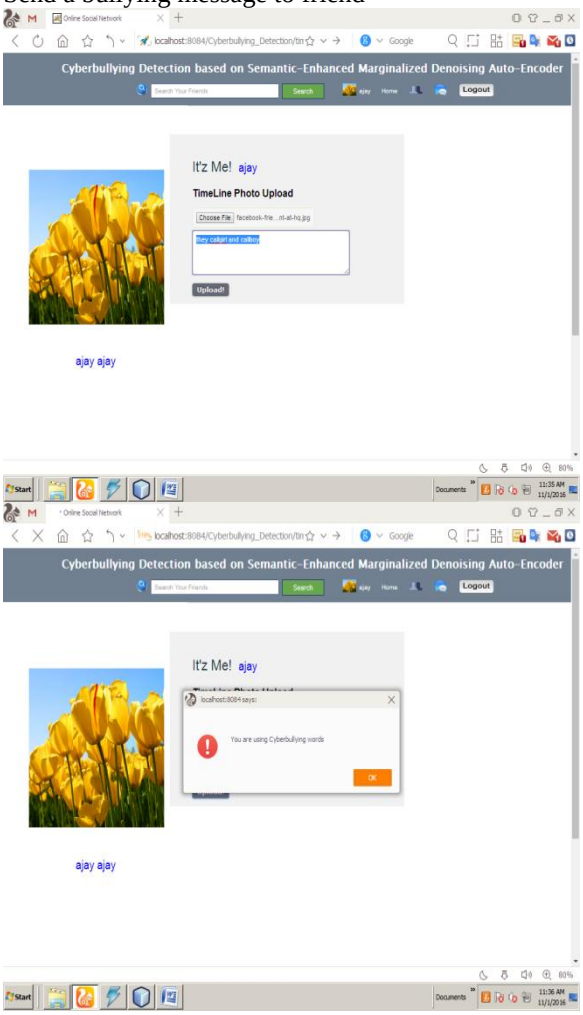
V. OUTPUT

User Login

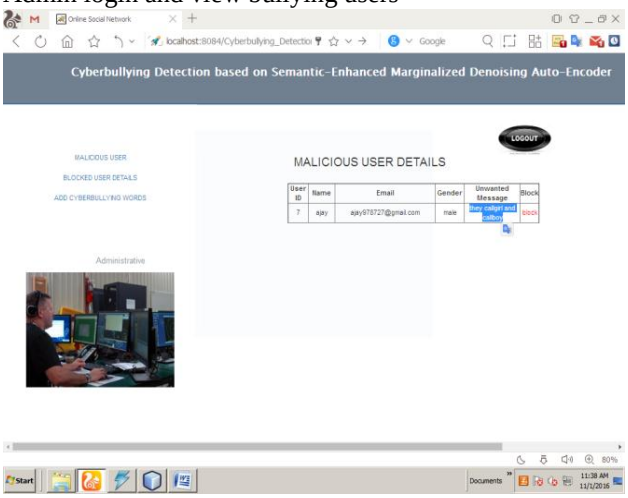




Send a bullying message to friend



Admin login and view bullying users



VI. CONCLUSION

In, This paper addresses the text-based cyberbullying detection problem, where robust and discriminative representations of messages are critical for an effective detection system. By designing semantic dropout noise and enforcing sparsity, we have developed semantic-enhanced marginalized denoising autoencoder as a specialized representation learning model for cyberbullying detection. In addition, word embeddings have been used to automatically expand and refine bullying word lists that is initialized by domain knowledge. The performance of our approaches has been experimentally verified through two cyberbullying corpora from social medias: Twitter and MySpace. As a next step we are planning to further improve the robustness of the learned representation by considering word order in messages.

REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment Analysis of Twitter Data," In Proceedings of workshop on languages of social media, 2011. pp. 30-38.

[2] A. Agarwal, F. Biadsy, and K. R. Mckeown. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009. pp. 24-32.

[3] A. Go, R. Bhayani, L. Huang, "Twitter sentiment classification using distant supervision", CS224N Project Report, Stanford, 2009. pp. 1-12.

[4] A. Harb, M. Plantié, G. Dray, M. Roche, F. Trousset, and P. Poncelet. "Web Opinion Mining: How to extract opinions from blogs?" In Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, ACM, 2008. pp. 211-217.

[5] A. Pak, and P. Paroubek "Twitter as a corpus for sentiment analysis and opinion mining," Proceedings of the seventh International Conference on Language Resource and Evolution (LREC'10), 2010. pp. 19-21.

[6] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. "From tweets to polls: Linking text sentiment to public opinion time series." ICWSM 11, no. 122-129 ,2010. pp. 1-2.
[7] B. Pang, L. Lee, and S. Vaithyanathan. "Thumbs up? : sentiment classification using machine learning techniques."Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002. pp. 79-86.
[8] C. Whitelaw, N. Garg, and S. Argamon "Using appraisal groups for sentiment analysis," presented at the Proceedings of the 14thACM international conference on Information and knowledge management, Bremen, Germany, 2005.
[9] D. M. Haughton, J. J. Xu, D. J. Yates, X. Yan "Introduction to Data Analytics and Data Mining for Social Media Minitrack", 49th Hawaii International Conference on System Sciences, 2016.p. 1414.
[10] D. M. Law, J. D. Shapka, and B. F. Olson. "To control or not to control? Parenting behaviours and adolescent online aggression." Computers in Human Behavior 26.6,2010 pp.1651-1656.
[11] D. Quinn, C. Liming, and M. Maurice "Does age make a difference in the behaviour of online social network users?"Internet of Things (iThings/CPSCom).

