CYBERBULLYING DETECTION USING MACHINE LEARNING

Literature Survey Report submitted in the partial fulfilment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

by

D. Niharika Naik **19K91A05E5**

P. Ram kumar **19K91A05F5**

P. Aryan Raj **19K91A05F7**

P. Bhargavi **19K91A05F8**

R. Sai Varsha **19K91A05G3**

UNDER THE GUIDANCE OF

Ms. Y. Latha

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE &ENGINEERING

TKR COLLEGE OF ENGINEERING & TECHNOLOGY

(AUTONOMOUS)

(Accredited by NBA and NAAC with 'A' Grade)

Medbowli, Meerpet, Saroornagar, Hyderabad-500097

2022-2023

CERTIFICATE

This is to certify that the main project report entitled **CYBERBULLYING DETECTION USING MACHINE LEARNING,** being submitted by Ms. **D.NIHARIKA NAIK**, bearing ROLL.NO:**19K91A05E5**, Mr. **P.RAMKUMAR**, bearing ROLL.NO:**19K91A05F5**, Mr**. P. ARYANRAJ** bearing ROLL.NO:**19K91A05F7**,Ms **P.BHARGAVI**, bearing ROLL.NO:.**19K91A05F8 ,** Ms. **R. SAI VARSHA** bearing ROLL NO:**19K91A05G3** in partial fulfilment of requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering, to the TKR College of Engineering and Technology is a record of Bonafide work carried out by them under my guidance and supervision.

Name and Signature of the Guide

Ms. Y. Latha

Assistant Professor

Name and Signature of the HOD

Dr. A.Suresh Rao

Professor

CONTENTS:

1. INTRODUCTION

Social media networks such as Facebook, Twitter, Flickr, and Instagram have become the preferred online platforms for interaction and socialization among people of all ages. While these platforms enable people to communicate and interact in previously unthinkable ways, they have also led to malevolent activities such as cyber-bullying. Cyber bullying is a type of psychological abuse with a significant impact on society. Cyber-bullying events have been increasing mostly among young people spending most of their time navigating between different social media platforms. Particularly, social media networks such as Twitter and Facebook are prone to CB because of their popularity and the anonymity that the Internet. In India, for example, 14 percent of all harassment occurs on Facebook and Twitter, with 37 percent of these incidents involving youngsters. This article, we mainly focus on the problem of cyber-bullying detection on the Twitter platform. As cyberbullying is becoming a prevalent problem in Twitter, the detection of cyberbullying events from tweets and provisioning preventive measures are the primary tasks in battling cyberbullying threats. There is a greater need to increase the research on social networks-based CB in order to get greater insights and aid in the development of effective tools and approaches to effectively combat cyberbullying problem. Manually monitoring and controlling cyberbullying on Twitter platform is virtually impossible. Furthermore, mining social media messages for cyberbullying detection is quite difficult. For example, Twitter messages are often brief, full of slang, and may include emojis, and gifs, which makes it impossible to deduce individuals' intentions and meanings purely from social media messages. Moreover, bullying can be difficult to detect if the bully uses strategies like sarcasm or passive-aggressiveness to conceal it Despite the challenges that social media messages bring, cyberbullying detection on social media is an open and active research topic. Cyberbullying detection within the Twitter platform has largely been pursued through tweet classification and to a certain extent with topic modelling approaches. Text classification based on supervised machine learning (ML) models are commonly used for classifying tweets into bullying and non-bullying tweets. Deep learning (DL) based classifiers have also been used for classifying tweets into bullying and non-bullying tweets. Supervised classifiers have low performance in case the class labels are unchangeable and are not relevant to the new events. Also, it may be suitable only for a pre-determined collection of events, but it cannot successfully handle tweets that change on the fly. Topic modelling approaches have long been utilized as the medium to extract the vital topics from a set of data to form the patterns or classes in the complete dataset. Although the concept is similar, the general unsupervised topic models cannot be efficient for short texts, and hence specialized unsupervised short text topic models were employed. These models effectively identify the trending topics from tweets and extract them for further processing. These models help in leveraging the bidirectional processing to extract meaningful topics. However, these unsupervised models require extensive training to obtain sufficient prior knowledge, which is not adequate in all cases. Considering these limitations, an efficient tweet classification approach must be developed to bridge the gap between the classifier and the topic model so that the adaptability is significantly proficient.

2. LITERATURE SURVEY REPORT

2.1 Bully Net: Unmasking Cyber bullies on social networks.

Title: Bully Net: Unmasking Cyber bullies on Social Networks

Published by: Aparna, Sankaran Srinath, Hannah Johnson, Gaby G. Dagher and Min Long

The Internet has created never before seen opportunities for human interaction and socialization. In the past decade, social media in particular , has had a popularity explosion. From My Space to Facebook, Twitter, Flickr, and Instagram, people are connecting and interacting in a way that was previously impossible. The widespread usage of social media across people from all ages created a vast amount of data for several research topics, including recommender systems, link predictions, visualization, and analysis of social networks.

We study the problem of cyber bullying in social media in an attempt to answer the following research question: Can tweet contexts (conversations) help improve the detection of cyber bullying in Twitter? Our intuition is that each tweet should be evaluated not only based on its contents but also based on the context in which it exists. We call such a context a conversation, which is a set of tweets between two or more people exchanging information about a certain subject. Thus, our solution consists of three parts. First, for each conversation, a conversation graph is generated based on the sentiment and bullying words in the tweets. Second, we compute the bullying score for each pair of users in a conversation graph and then combine all graphs to create an SSN called bullying SN (B). The inclusion of negative links can bring out information that would otherwise be missed with only positive links. Finally, we propose a centrality measure called attitude and merit (A&M) to detect bullying users from the SNB.

Related Work:

In this section, we review the literature on areas related to cyber bullying detection and SSNs.

A. Cyber bullying Detection There is not a lot of works in the literature that utilizes SNs to detect cyber bullies. We are aimed at detecting trolls in an SN.

B. SSNs This section reviews the previous work done on SNs. The idea of SNs is not new, but its application and analysis of them were only developed in recent years. We extended its application to establish node classification in our model.

Proposed Work:

1) Collected, pre-processed, and labelled the Twitter data set.

2) Proposed a novel efficient algorithm for detecting cyberbullies on Twitter.

a) Built conversation.

b) Constructed bullying SN.

c) Proposed A&M centrality.

3) Experimented on 5.6 million tweets collected over six months. The results show that our approach can detect cyberbullies with high accuracy while being scalable with respect to the number of tweets.

Results:

We exploit bullying tendencies by proposing a robust method for constructing a cyberbullying signed network (SN). We analyse tweets to determine their relation to cyberbullying while considering the context in which the tweets exist in order to optimize their bullying score. We also propose a centrality measure to detect cyberbullies from a cyberbullying SN and show that it outperforms other existing measures. We experiment on a data set of 5.6 million tweets, and our results show that the proposed approach can detect cyber bullies with high accuracy while being scalable with respect to the number of tweets. Index Terms Cyberbullying, signed networks (SNs), social media mining.

Conclusion:

Although the digital revolution and the rise of social media enabled great advances in communication platforms and social interactions, a wider proliferation of harmful behaviour known as bullying has also emerged. This article presents a novel framework of Bully Net to identify bully users from the Twitter social network. We performed extensive research on mining SNs for better understanding of the relationships between users in social media, to build an SN based on bullying tendencies.

2.2 Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection

Title: Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection

Published by: Zheng Lin Chia a, Michal Ptaszynsl Fumito Masui a, Gniewosz Leliwa b,Michal Wroczynski b.

Introduction:

Irony and sarcasm detection has been increasingly recognized as an important task in the field of Natural Language Processing, with many related studies conducted on this topic in recent years (Barbieri, 2017; Burfoot & Baldwin, 2009; Ghosh & Veale, 2017; Reyes, Rosso, & Buscaldi, 2012). Irony, often used together or interchangeably with sarcasm, is considered an important component of human communication as one of the most prominent and pervasive figurative and creative language tools widely used dating back from ancient religious texts to modern time (Ghosh & Veale, 2017).There has been little agreement on the correct definitions of irony and sarcasm despite their popularity as figurative means of expression in everyday communication, especially on the Internet (Hancock, 0000; Li & Ma, 2016; Papapicco & Mininni, 2020). A common and thorough clarification of differences between irony and sarcasm accepted by the whole research community does not seem to exist yet even with a considerable amount of related literature published in the past decades. Moreover, Hee (2017), in her attempt to define the concept, pointed out that many studies have also struggled to distinguish between irony, in particular verbal irony, and sarcasm. Thus, she suggested not to distinguish between the terms due to the lack of clear quantifiable distinctions between those two phenomena despite numerous efforts. The difficulty of this task comes from the fact that, similarly to other types of figurative language, ironic texts should not be interpreted in its literal sense due to requiring a more complex understanding based on associations with context and external world knowledge

Related Work:

While most of the previous research in irony detection within the field of AI focused on binary classification between ironic or non-ironic contents, two types of irony have been widely distinguished in most of the previous linguistic and communication studies on irony: verbal irony

and situational irony (Barbieri, 2017; Hee, 2017; Sulis, Fariaz, Rosso, & Patti, 2016; Van Hee et al.,2018). This distinction has also been acknowledged in the Semantic Evaluation 2018 Workshop, a workshop in the form of a contest where multiple teams attempt to develop a Machine Learning method based on a unified dataset. The workshop's Task B especially focused on multi-class irony classification. The task had the participants compete in predicting one out of four labels describing (i) verbal irony realized though a polarity contrast, (ii) verbal irony without a polarity contrast, (iii) descriptions of situational irony, and (iv) non-irony (Barbieri, 2017; Van Hee et al., 2018). Situational irony is an unexpected or incongruous event in a specific situation that fails to meet an expectation (Barbieri, 2017; Shelley, 2001). Shelley (2001) gives an example of a typically ironic situation regarding firefighters who left something cooking, had a fire in their kitchen while they were out putting down a fire in the other part of the city. As firemen are usually the ones who extinguish fire instead of starting it, this situation is quite unexpected and is considered ironic. This shows that situational irony is usually produced unintentionally and unplanned. As indicated by Grant, Hardy, Oswick, and Putnam (2004) ''Situational irony focuses on the surprising and inevitable fragility of the human condition, in which the consequences of actions are often the opposite of what was expected''. According to ''A glossary of literary terms'' by Abrams and Harpham (2009), verbal irony is a statement in which the meaning that a speaker employ is sharply different from the meaning that is ostensibly expressed. An ironic statement usually involves the explicit expression of one's attitude or evaluation but with intended implications being very different, and often opposite, to the literal attitude or evaluation. Verbal irony is considered different from situational irony in that it is produced intentionally by the speakers. On the other hand, sarcasm is defined to be ''a way of using words that are the opposite of what you mean in order to be unpleasant to somebody or to make fun of them'' by the Oxford dictionary (Oxford, 2020) As an attempt to provide explanation on the differences between irony and sarcasm, ''A Dictionary of Modern English Usage'' (Fowler, 1926) points out that sarcasm does not necessarily involve irony and irony has often no touch of sarcasm, even though sarcasm is often expressed with irony as a tool. Therefore, the relationship between verbal irony and sarcasm have been confused in many studies. On the other hand, Kreuz and Glucksberg (1989) argued that sarcasm and irony are similar in that both are forms of a reminder, yet different in that sarcasm conveys ridicule of a specific victim whereas irony does not. Lee and Katz (1998) followed up with an indication that a ridicule of a specific victim plays a more important role in sarcasm than in irony. They also pointed out that a sarcastic utterance brings to mind the expectation of a specific person who is identified by that expectation, whereas irony brings to mind the collective expectation of numerous people. In the same vein, Jorgensen (1996) coined the

term ‘‘sarcastic irony’’ which is typically used to complain to or criticize intimates, who are usually the target of the remarks. Attard (1999) argues that sarcasm is an overtly aggressive type of irony and also claims that there is no consensus on whether sarcasm and irony are essentially the same thing, with superficial difference, and that they do not differ significantly. Many studies also claim that there is no way to distinguish between the terms (Tsur et al., 2010). Barbieri (2017) points out another reason why many researchers do not differentiate between the irony and sarcasm which is due to the observation of a shift in meaning between the two terms. They also conclude that while research efforts on irony and sarcasm are expanding, a formal definition is still lacking in the literature. Therefore, many researchers tend to not distinguish between the terms and consistently use either of them throughout their studies (Bouazizi & Outski, 2016; Busch Meier, Cimiano, & Klinger, 2014; Hee, 2017). On the other hand, in some languages other than English, such as Japanese (where irony/sarcasm is called hiniku), have no distinction between irony and sarcasm. This is because the figurative function of irony, which in English is considered as a sophisticated figure of speech enriching conversation, in Japanese is used only in its aggressive (sarcastic) context information Processing and Management.

Predictive Validity While the total victimization scores were significantly and negatively correlated with life satisfaction total scores ($r = -.29$, $p < .05$), it was negatively but not significantly correlated with hope total scores ($r = -.13$, $p > .05$). In addition, the results of the analysis of variance conducted to examine the differences in life satisfaction and hope according to the victimization classification (non-victims, peer victims, and bullied victims) indicated significant group differences.

Conclusion:

In this paper, we set out to explore the sarcasm and irony on Twitter, using various Natural Language Processing and Machine Learning techniques. First, we reviewed and clarified the definitions of irony and sarcasm by discussing various studies focusing on those terms. The review extended our knowledge of the definitions for the terms irony and sarcasm, which indicates an occurrence of a meaning shift between those terms throughout the modern days. The terms were originally strictly distinguishable; however, most researchers and social media users no longer differentiate clearly between them. Therefore, the review suggested that the term irony and sarcasm are being used interchangeably on social media in modern days. Next, we conducted first experiment comparing between various types of classification methods including some popular classifiers for text classification task. The results of this experiment show that machine learning

methods, especially deep learning methods, are rising in the latest trend by having the most potential for classification tasks. However, the downside of deep learning methods is the requirement for having a large dataset in order to achieve the best result. We also observed the importance of the social media markers (e.g., #hashtags in Twitter) which greatly impact the classification results. For the second experiment, we compared between different types of data pre-processing methods with the classifier ranked best from the previous experiment based on the dataset with all hashtags removed, which is the Convolutional Neural Network. The findings from the results enhanced our understanding of data pre-processing where the best result came from the dataset with the least pre-processing methods applied. This is due to oversimplification of data which causes many vital and important features, on which irony and sarcasm detection heavily depended, being manipulated, or removed. However, we observed that further data processing could still be crucial to the sensitivity of the results. We also compared between sarcasm and irony utilizing their respective dataset from previous experiments. We trained models on sarcasm dataset and tested on irony dataset, and vice versa. Interestingly, the highest result attained an F-score of 0.94 which provided additional evidence with respect to the similarity between sarcasm and irony. The results also supported the claim that sarcasm is mostly a type of irony (aggressive irony). Finally, we conducted a small experiment where model trained on sarcasm dataset was tested on cyberbullying dataset. The result attained an F-score of 0.889 which is comparable to the result of sarcastic dataset itself. An implication of this is the possibility that there is preponderance of sarcasm in cyberbullying, and this extends our knowledge of the practical application of sarcasm detection in other tasks.

2.3 Nature-Inspired-Based Approach for Automated Cyber bullying Classification on Multimedia Social Networking.

Title: Nature-Inspired-Based Approach for Automated Cyberbullying Classification on Multimedia Social Networking.

Authors: N. Yuvaraj, K. Srihari, Gaurav Dhiman, K. Somasundaram, Ashutosh Sharma, S. Rajeskannan, Mukesh Soni, Gurjot Singh Gaba, Mohammed A. Alzain and Mehedi Masud.

Introduction:

Cyberbullying is defined as a repetitive, intentional, and aggressive reaction committed by a group or an individual against another group or an individual, which is made by the utilization of Information Communication Technology (ICT) tools such as social media, Internet, and mobile phones. An automated behaviour of social network platforms alerts the moderators to review the reported CB contents. However, most of the frameworks lacks an automated intelligent system that alerts the moderators and detects the contents in an automated way faster than the traditional reporting system. This enables the moderator to respond on the alert and take required action on reporting the user or removing the content.

Related Work:

Nandhini and Sheeba presented a detection technique to combat CB on social media. The study extracts feature such as the noun, pronoun, and adjective obtained from the text and frequency of words occurrences. These features are used to classify various activities such as Harassment, Flaming, Terrorism, and Racism using a Fuzzy logic-based genetic algorithm. employed a Support Vector Machine (SVM) classifier to classify the CB based on various features such as local, sentimental, contextual, and gender-specific language features. The SVM classifier combined with a TF-IDF measure and linear kernel identifies the online harassment.

Proposed Work:

In the present research, the entire focus is not on a specific CB word, but the vulgarity is determined based on weight score calculation and harmfulness index estimation for the entire word sequence (optimal words chosen by the feature selection method) of the collected tweets. This reduces well the cost of training data construction and further with the dependency between the phrases. The datasets are divided into smaller subset $L \subset D$. The aim is to detect the CB instances from the twitter data that may vary from long to short paragraphs

1. Pre-processing
2. Feature Selection
3. ANN (Artificial Neural Network)
4. DRL Algorithm for Reward-Penalty decision

Results: The study has selected 30,384 tweets collected from the twitter datasets [4]. The tweets contain both CB and non-CB tweets, where automated labelling or tagging is carried out using feature selection methods. Out of 30,384, more than 1252 tweets are classified as CB datasets; however, the labelled data are not used to train the classifier. These labelled data act as an input for the DRL method, which rewards or penalizes the ANN mechanism. The entire datasets have more imbalanced classes that penalize the unsupervised ANN with inaccurate results in identifying the relevant instances. The ANN, on other hand, with imbalanced classes, ignores minor classes, and it performs well with major classes. The weight adjustment approach helps to avoid oversampling of the minority class, i.e., abnormal class and under sampling the majority class, i.e., the normal class. The entire set of experiments is conducted with the topmost algorithms performed well in existing methods that include the ANN, SVM, RF, and LR. These existing methods are compared with ANN-DRL to find the classification accuracy. The performance is estimated against various metrics that include accuracy, F-measure, geometric mean (G- Mean), percentage error, precision, sensitivity, and specificity mean), percentage error, precision, sensitivity, and specificity.

Conclusion:

In this paper, an integrated model using an ANN and DRL is designed for the classification of CB from raw text datasets of a social media engine. The extraction of psychological features, user comments, and the context has enabled better classification performance, where an ANN at the initial stage performs with improved classification results. The addition of a reward-penalty system using DRL has enhanced the classification to a much greater level than the ANN model. The simulation results illustrate the improved average classification accuracy of 80.69% using ANN-DRL than existing three-layered ANN (77.40%), SVM (75.44%), RF (75.55%), LR (75.10%), and NB (75.19%). In future, the convolutional neural network can be applied on image datasets to extract the information to serve the purpose on reducing the cyberbullying.

2.4 Semantic Analysis Techniques using Twitter Datasets on Big Data: Comparative Analysis Study

Title: Semantic Analysis Techniques using Twitter Datasets on Big Data: Comparative Analysis Study.

Published by: Belal Abdullah Hezam Murshed, Hasib Daowd Esmail Al-ariki, Suresha Mallappa

Introduction:

Cyberbullying is a conscious and persistent act of violence that aims to threaten or harm individuals, deliberately and repeatedly using communication and information technologies. According to statistical data more than half of adolescents have been involved in or have witnessed cyberbullying, whilst 10% to 20% witness it every day. The emergence and increased use of the internet, especially Twitter and Facebook, have exacerbated this situation. The study presented in this article is the first, we believe, to incorporate the level of cyberbullying severity using multi-class classification into an automatic cyberbullying detection model. Based on the literature and empirical evidence, we hypothesize that the incorporation of multi-class classification results in a more effective cyberbullying detection model, in contrast to a binary classification. In order to perform our multi-class classifier study, we categorized the annotated cyberbullied tweets into four levels; low, medium, high, and non-cyberbullying. Based on the classification, sexual and appearance-related tweets were classified as high-level cyberbullying severity; political and racial tweets as medium level; intelligence tweets as low-level cyberbullying severity, and non-cyberbullying tweets.

Related Work:

Based on the findings of these studies, our approach integrates some of these features the baseline model for detecting cyberbullying severity on Twitter. 1. Network-Based Features2. Machine Learning, 3. Features Summary, 4. Master Feature (PMI-SO). The objective of feature selection is threefold: improving the performance of the data mining model, providing a faster more cost-effective learning process, and providing a better understanding of the underlying process that generates the data Many applications are characterized by various dimensional data, where not all the features are important. Therefore, three feature selection techniques were used in the algorithm training process, namely: Chi-square, Information gain. Choosing the best classifier is the most

significant phase of the text classification pipeline obtained from the tweets have been used to build a model to detect cyberbullying severity. In order to select the best classifier, we tested several machine learning algorithms namely: NB, SVM with RBF kernel, DT, RF, and KNN.

Conclusion:

The use of the internet and social media has significant benefits for society, but the excessive usage of the internet and social media has major detrimental effects too. This includes unwanted sexual exposure, cybercrime, and cyberbullying. Cyberbullying is a conscious and persistent act of violence that aims to threaten or harm individuals, deliberately and repeatedly using communication and information technologies. This situation has been worsened by the increased use of the Internet, especially on Twitter and Facebook. Studies have shown that knock-on effects of cyberbullying can potentially be harmful including learning disabilities, psychological distress, and depression, escalating physical confrontations, and suicide. The developed model is a feature-based model that uses features from contents of tweets to develop a machine learning classifier for classifying the tweets as non-cyberbullied, and low, medium, or high-level severity of cyberbullying. Experiments to test the efficiency of five well-known classifiers, namely, NB, SVM, KNN, DT, and RF. All five classifiers were tested on Bo W, Word2Vec, using a proposed manually engineering technique to see the significance of classifiers' performance when text details imbalanced. RF achieved the highest Kappa of 84%, F-Measure 92%, and Accuracy 93% when parameters set to Base Classifier SMOTE Cost Adjusted + Predicted Features PMI. Finally, we performed multinomial logistic regression to identify highly significant predictors for cyberbullying severity.

2.5 Careful what you share in six seconds: Detecting cyber bullying instances in Vine

Title: Careful what you share in six seconds: Detecting cyber bullying instances in Vine
Published by: Rahat Ibn Rafiq, Homa Hossein Mardi Richard Han, Qin Lv, Shivakant Mishra

Introduction:

Mobile social networks like Instagram, Vine and Snap chat are booming in popularity, spurred by the revolution in smart phones, and therefore represent a natural target for investigating cyberbullying. Vine (purchased by Twitter) is interesting because it offers the opportunity to explore cyberbullying in the context of video-based communication, which has been gaining popularity recently. Vine is a mobile application that allows users to record and edit six-second looping videos, which they can share on their profiles for others to see, like and comment upon. Cyberbullying can happen in Vine in many ways, including posting mean, aggressive and hurtful comments, recording video of others without their knowledge and then sharing the Vines to make fun of or mock them, and playing "the slap game" in which one person records video while another person slaps or hits a person in order to record a reaction. They later share the Vine for the world to see. There are even violent versions called "knock-out" where someone punches an unsuspecting person to knock them out. Provides an illustration where the profile owner is victimized by hurtful and aggressive comments posted by others. In the following research analysis, we make a distinction between cyber aggression and cyberbullying. Cyber aggression is defined as a type of behaviour in an electronic context that is meant to intentionally harm another person. Cyberbullying is defined in a stronger and more specific way as aggressive behaviour that is carried out repeatedly in OSNs against a person who cannot easily defend himself or herself, creating a power imbalance. Thus, in order to understand cyberbullying, the factors of repetition of aggression and imbalance of power must be taken into account.

Related Work:

Previous research on "cyberbullying" is more accurately described as research that is focused on studying cyber aggression as this research did not take into account the repetitive nature nor the power imbalance of the cyberbullying definition. Also, they are primarily focused on analysing and labelling text-based comments. Some researchers have Tire to incorporate other information to detect bullying behaviour and victims, such as looking at the number of received and sent comments, or considering some graph properties besides just text features. While research investigating profanity in Ask.FM and Instagram provided some insights into cyber aggression, it did not label the data for either cyber aggression or cyberbullying. Suggested a framework for using images besides text for detecting cyberbullying, and recent work has studied cyberbullying in the Instagram mobile social network, where labelling of media sessions (shared image

associated comments) has correctly distinguished between cyber aggression and cyberbullying, and a classifier was developed based on the labelled data. To our knowledge, our paper is the first to study cyberbullying in the context of a video-based mobile social network, in particular Vine.

Proposed Work:

As online social networks have grown in popularity, teenage users have become increasingly exposed to the threats of cyberbullying. The primary goal of this research paper is to investigate cyberbullying behaviour in Vine, a mobile based video-sharing online social network, and design novel approaches to automatically detect instances of cyberbullying over Vine media sessions. We first collect a set of Vine video sessions and use Crowd Flower, a crowd-sourced website, to label the media sessions for cyberbullying and cyberaggression. We then perform a detailed analysis of cyberbullying behaviour in Vine. Based on the labelled data, we design a classifier to detect instances of cyberbullying and evaluate the performance of that classifier.

Labelling Methodology: While designing the survey, our first goal was to choose the appropriate definitions of cyberbullying and cyberaggression. Cyberaggression is a broader term that includes using digital media to intentionally harm another person, whereas cyberbullying is a more restrictive form of intentional cyberaggression that is carried out repeatedly in an electronic context where the victim cannot easily defend himself or herself because of a power imbalance.

Analysis of Cyber bullying Labelling: A judgment was considered trusted if the trust score was at least 0.8, which was computed by Crowd Flower by incorporating the contributor's performance in answering the test questions and his/her overall trust score in Crowd Flower, thus giving us in total 4795 trusted judgments for 959 media sessions with 10 test questions. Average test question accuracy for the trusted, untrusted and all contributors were 86%,44% and 69% respectively. The contributors showed 76.6% and 79.49% agreement for the two questions, namely whether the media session constituted cyber regression or not and whether the media session constituted cyberbullying or not.

Classifier Performance: By considering the aforementioned features, we employ four classifiers namely Naive Bayes, AdaBoost, Decision Tree and Random Forest with 10-fold cross-validation. It demonstrates the very best combination of features that achieves the highest accuracy for each classifier. AdaBoost achieved the highest accuracy of 76.39%, using a combination of profile owner, media session, comment features.

Results:

• We investigated cyberbullying behaviour in Vine, a video-sharing mobile social network by labelling the videos along with the comments associated with them according to the appropriate definition of cyberaggression and cyberbullying.

• We presented a thorough analysis of the labelled videos, the associated comments, different features and metadata of the media-sessions and the relationship between these features and both cyberaggression and cyberbullying.

•The design and evaluate classifiers to effectively identify instances of cyberbullying based on the labelled data and all the features associated with the videos and comments

Conclusion:

We found that the percentage of high profanity-containing media sessions in Vine is quite low. we discovered that a significant fraction of the high profanity-containing media sessions were not labelled as cyberbullying, though in general there was a trend towards increasing identifications of cyberbullying as the percentage of profanity increased. This suggested that the percentage of profanity in a media session should not be used as the sole indicator of cyberbullying, but should be supplemented by other input features to the classifier. we found that not all media sessions that exhibit cyberaggression are instances of cyberbullying, validating the need to apply a stricter definition of cyberbullying. Fourth, we demonstrated that AdaBoost achieved the highest accuracy of 76.39%, using a combination of profile owner, media session, comment features and unigrams.

Discussion and Future work:

We plan to consider more sophisticated algorithms like Gradient Boosting classifiers in the future. We plan to consider other features as well. For example, differentiating the activities in the videos shared in Vine may prove helpful, namely is the activity related to sports, dancing, walking, etc. We would like to build automated classifiers so that the video activity category can be automatically input to the cyber bullying detection classifier

2.6 Detection of hate speech in Arabic tweets using deep learning:

Title: Detection of hate speech in Arabic tweets using deep learning

Published by: Areej Al Hassan, Hmood Al Dossari

Introduction:

According to The Arab Social Media Report, “the penetration of social media in some countries of the Arab region reached 90% of the population”. It shows also that 58% of the Arabs are expressing both of their positive and negative thoughts through social networks. Twitter has shown a rapid growth in the recent years. Arab users generate 27.4 million tweets per day. From that big number, we can assume that hate speech can spread easily and quickly through these platforms.

Social networks provide a chance for radical groups to aggregate people with similar thoughts to create a solidarity for some ideology to follow together. Hate speech is a controversial issue that cannot be prevented unilaterally due to the massive scale of social networks. s. This will be achieved by harnessing the power of supervised deep learning techniques to automate the identification of Arabic hate speech in Twitter

Related works:

Starting with the World Wide Web, Warner and Hirschberg, are the first to investigate how to identify hate speech in the World Wide Web. Their work is targeted to specific type of hate which is anti-Semitic. For Twitter, Watanabe et al. proposed a supervised approach for hate-speech detection. Their approach proved that supervised classifier performs better in the binary classification when compared with ternary classification. Another binary classifier is developed by Burnap and Williams that detects hateful and non-hateful tweets from labelled dataset.

In addition, Badjatiya et al also used Waseem and Hovy’s corpus but for different deep learning scenarios. They compared different combinations of deep learning models and state-of-art classifiers. They concluded that combining deep neural network models with GBDT classifier will result in the best accuracy. Also, Zhang et al have conducted a comparative evaluation and examined combining both of convolutional neural networks and gated recurrent networks. Their work was performed on several public datasets. For Arabic hate-speech detection, one contribution is found which is specifically performs a binary classification of religious hate speech. For instance, Albadi et al have worked with the classification of religious hatred in Arabic tweets. They used both of supervised and unsupervised approaches.

Proposed Work:

This research aims to develop a model for detecting Arabic hate speech in Twitter platform, then classifying the Arabic tweets based on the type of hate used in each tweet. Hate classes-We choose to assign five distinct classes of hate (Religious, Racism, Sexism, General hate speech, Not hate speech. Prior work in Arabic hate-speech detection has targeted mainly binary classification of hate. In Table1, we summarized what constitute each one of the classes of hate by mixing the previous hate speech properties with local Arab culture. Model architecture- A high-level view of the system to visualize and summarize all the phases related to our Arabic hate-speech detection model

Results: SVM resulted in average accuracy of 75% of the classifier. But as we mentioned before, since we are working with imbalanced dataset problem, we will rely on Recall and precision (shown in Table 4) for evaluating the models. From the results, we can see that SVM is able to distinguish non-hate-speech tweets, this can be seen from the high recall resulted. On the other hand, SVM is not able to distinguish the hate-speech classes (Very poor recall in the other classes), which means that in our case, SVM works poorly for multi classification and imbalanced datasets, but it works fine in case of binary classification. We can justify the high precision of hate-speech classes by referring to its equation, the denominator of the precision equation describes the total number of tweets classified to a specific class by the model itself, so precision quantifies only the correctly classified tweets with respect to the total classified tweets by the model. Hence, we can imagine that the model was able to detect a very low number of hate speech tweets.

Conclusion:

The national need developing a model that automatically detects Arabic hate speech in twitter. After acquiring a sufficient knowledge in what constitutes Arabic hate speech, we used this knowledge to label a dataset of 11 K tweets. Then, we built our SVM baseline using TF-IDF words representation and proposed four deep learning architectures that can identify and classifying Arabic hate speech in twitter into 5 classes. We compared the proposed models with the SVM baseline. Comparison results show that our deep learning approaches outperformed the baseline in the multiclassification of hate classes. However, the ensemble model of CNN LTSM produced the best results. As future work we will consider expanding our data set and intensifying the training of our neural networks by including data from another platform “Facebook” as it is the most used platform in the Arab region.

2.7 Semantic analysis techniques using Twitter datasets on big data: Comparative analysis study

Title: Semantic analysis techniques using Twitter datasets on big data: Comparative analysis study

Published by: Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam.

Introduction:

The problem of calculating semantic similarity between two words/texts/ sentences/phrases is a long-standing issue in the field of Natural Language Processing (NLP). Generally, semantic similarity is a metric of the conceptual distance between two terms, based on the closeness of their meanings. Sentence similarity approaches play an increasingly significant role in studies and applications associated with text in several fields such as document clustering, classification of text, IR, topic tracking, topic detection, text summarization, machine translation, and so on. Semantic similarity among documents, sentences, phrases, texts, and words are extensively studied in different areas, encompassing NLP, semantic search engines, semantic web, and Artificial Intelligence (AI). This method is based on the shortest path similarity and the maximum depth of the taxonomy with log smoothing. This paper conducts a comprehensive review of various word and sentence semantic similarity techniques proposed in the literature. Corpus-based, Knowledge-based, and Feature-based are categorized under word semantic similarity techniques. String and set-based, Word Order-based Similarity, POS based, Syntactic dependency-based are categorized as sentence semantic similarity techniques. Using these techniques, we propose a model for computing the overall accuracy of the twitter dataset. The proposed model has been tested on the following four measures: Atish's measure, Li's measure, Mihalcea's measure with path similarity, and Mihalcea's measure with Wu and Palmer's (WuP) similarity. Finally, we evaluate the proposed method on three real-world twitter datasets. The proposed model based on Atish's measure seems to offer good results in all datasets when compared with the proposed model based on other sentence similarity measures. Conclusion Semantic similarity measures are widely used in many fields including Natural Language Processing, Web search, and so on. This paper investigated several techniques of computing semantic similarity measures, which measure both the word and sentence semantic similarity. Three categories introduced in word semantic similarities which are namely corpus-based, knowledge-based, and feature-based were described. The four categories presented in sentence semantic similarity techniques based on String and Set-

based, Word Order-based Similarity, POS-based, Syntactic dependency-based techniques were also described. The proposed model for calculating the overall accuracy of the twitter dataset based on the sentence semantic similarities presented has also been described. The experiments conducted on all three twitter datasets to evaluate the proposed model have also been covered in details. The experimental results seem to indicate that the model proposed based on Atish's measure is superior to the proposed model based on other similarity measures.

Related work:

The samples of the dataset consisting of 10 tweets derived from the "Ethiopian Airlines Plane Crash dataset" (EAPC_DS2019) were used as mentioned in Table 7. The experiments on this dataset were conducted and tested on three metrics, namely Mihalcea's algorithm with path similarity, Mihalcea's algorithm with Wup similarity, and Li's method. The various accuracy scores have also been computed and compared in Tables 8, 9, and 10. It can be observed that the semantic similarity between tweet (T1) and tweet (T2) using three methods are 0.41, 0.638, and 0.441 respectively. Further, the semantic similarity between tweet (T2) and tweet (T8) are 0.457,0.632, and 0.34 respectively. In addition, the semantic similarity between tweet (T9) and tweet (T10) using all three methods is 1 because their texts are the same. Two formulas namely ASS (1) as given in Equation 29 and ASS(2) as given in Equation 30 have been used. As pointed out in this paper, ASS (2) shows that the best accuracy score between all tweets in the samples.

Conclusion:

Semantic similarity measures are widely used in many fields including Natural Language Processing, Web search, and so on. This paper investigated several techniques of computing semantic similarity measures, which measure both the word and sentence semantic similarity. Three categories introduced in word semantic similarities which are namely corpus-based, knowledge-based, and feature-based were described. The four categories presented in sentence semantic similarity techniques based on String and Set-based, Word Order-based Similarity, POS-based, Syntactic dependency-based techniques were also described. The proposed model for calculating the overall accuracy of the twitter dataset based on the sentence semantic similarities presented has also been described. The experiments conducted on all three twitter datasets to evaluate the proposed model have also been covered in details. The experimental results seem to indicate that the model proposed based on Atish's measure is superior to the proposed model based on other similarity measures.

2.8 Cyber bullying identification in twitter using support vector machine and information gain-based feature selection

Title: Cyber bullying identification in twitter using support vector machine and information gain-based feature selection.

Published by: Gita Dwi Purnamasari, M. Ali Fauzi, Indriati, Liana Shinta Dewi

Introduction:

The advance of information and communication technology has been brought many benefits to the society. One of the brief examples is social media in which people can find a lot of friends and extend their networks. However, the rise of this new technology tends to be double-edge knife as suggested by Segal because it also bring some damaging effects such as cyberbullying. Cyberbullying can have an impact on victim's mental, even there is many cases where the victims of bullying end up in suicide because they cannot stand with much pressure. A research on cyberbullying used SVM and Naïve Bayes (NB) method for the classification process. The data used were obtained from Kaggle. The dataset contained 1600 conversations Indonesian Language from Formspring.me website. The researchers also compared their results from previous studies by Reynolds that used Decision Tree and K-Nearest Neighbour (K-NN). From the research, it was found that SVM method is better in cyberbullying classification than K-NN and Decision Tree method with accuracy level for SVM, Decision Tree, and KNN were 99.41%, 78.28%, and 89.01% respectively. Several prior works suggested that the use of feature selection can make the text classification more effective and efficient. Information Gain (IG) is the widely used technique in the text classification task.

Related Work:

The first step of this work is text pre-processing and followed by feature selection using IG. Then, the classification task is conducted using SVM method to get the cyberbullying identification result. The pre-processing involves some processes such as tokenizing, filtering, stemming and then term weighting. The features used in this work are Bow with term frequency-inverse document frequency (TF-IDF) as the term weighting method. The features are then used as the input of SVM.

Proposed Work: The proposed system suggests the use of SVM (Support Vector Machine) algorithm. SVM is also capable of working on high dimensional datasets using Cyberbullying

 kernel trick. There are several functions of the SVM kernel, such as: Linear, Polynomial, Gaussian RBF, Sigmoid, Multi Quadratic Inverse, and Additive.

In this research kernel function used is SVM Polynomial. Linear SVM is used when data to be classified can be separated by a hyperplane, whereas a non-linear SVM is used when data can only be separated by curved lines. SVM Polynomial has a function definition.

Result:

In the testing process, Accuracy, Precision, Recall, and F-measure were used. The amount of data used is 300 tweets, of which 150 tweet contain bullying and 150 tweet are not contain bullying. The data were manually labelled by an expert. In the experiment, the data was spitted into 240 tweets as training data and 60 tweets as testing data.

•There were six parameters tested on sequential training SVM with ten different experimental value i.e., lambda, gamma, epsilon, maximum iteration, and complexity (C) values. The sequential training SVM parameter values used in the test are $\lambda = 0.5$, $\Upsilon = 0.001$, $\varepsilon = 0.0001$, $C = 1$, and maximum iteration = 100.

•After performing the classification using SVM, the best SVM parameters obtained are maximum iteration = 20, $\lambda = 0.5$, $\gamma = 0.001$, $\varepsilon = 0.000001$, dan $C = 1$.

Conclusion:

Based on the experiment result, it can be concluded that the cyberbullying tweet identification using SVM method and IG feature selection get a promising result. The most optimal SVM parameters obtained in this work are maximum iteration=20, λ=0.5, γ=0.001, ε=0.000001, and C=1. Meanwhile, the best threshold value of IG feature selection is 90% with accuracy of 76.66%, precision of 72.22%, recall of 86.66%, and f-measure of 78.78%.

2.9 Identification and characterization of cyberbullying dynamics in an online social network:

Title: Identification and characterization of cyberbullying dynamics in an online social network
Published by: A. Squicciarini, S. Rajtmajer, Y. Liu, C. Griffin .

Introduction:

Cyberbullying, defined as using information technology to wilfully and repeatedly hurt, insult or harass others has become a serious problem among children, adolescents and young adults. According to recent statistics 19% of teens engaged in online social networking activities reported being victims of some form of cyberbullying. Compared with traditional bullying, cyberbullying tends to be more sinister because cyberbullying is not restricted by time and space and can occur more frequently and intensely, making it more difficult to control. Studies in the fields of psychology and sociology have investigated the dynamics of cyberbullying, bullies' motives, and interactions. Researchers have focused on personality, social relationships and psychological factors involving both the bully and the victim. Many of these works note the relevance of peer pressure and the peer group in incidences of bullying and cyberbullying.

We validate our models on two distinct social network datasets, comprised of over 16, 000 posts and more than 1000 users in total, finding that social network features are central to classifying cyberbullying dynamics. In a large study of posts on the Myspace network, we achieve 81.5% accuracy in detecting the peer-to-peer transmission of bullying behaviour with the inclusion of graph metric features, as compared to 78.7% without. In a similar experiment on second dataset from the Form spring social network, we find that even social network features alone are extremely powerful, achieving 78.1% accuracy in the detection of pairwise interactions between users involving bullying. Classic content-based features, e.g., offensive words, sentiment, achieve just 61.7% accuracy, a result echoed in a prior study on this dataset. In this work, we expand upon this idea, leveraging the underlying graph structure of our dataset to determine peer pressure dynamics that lead to the spread of bullying. In capturing the features of pairwise social network ties that transmit bullying, we suggest a model for integrating textual, user/demographic and graph-theoretic features for targeted detection of cyberbullying and prediction of cyberbullying dynamics in social networks.

Related works:

Our work is most similar to that of Huang and colleagues, who recently proposed a simple study hinting at the importance of social network features for bully detection. Their study considers a corpus of Twitter messages and associated local ego-networks to formalize the local neighbourhood around each user. Their results suggest that the social signals are useful for detecting cyber bullying, and that using multiple channels of information result in higher detection performance. In this work, we expand upon this idea, leveraging the underlying graph structure of our dataset to determine peer pressure dynamics that lead to the spread of bullying. In capturing the features of pairwise social network ties that transmit bullying, we suggest a model for integrating textual, user/demographic and graph-theoretic features for targeted detection of cyberbullying and prediction of cyberbullying.

Proposed Work:

Detection of Cyber bullies-

• For the detection of cyber bullies, i.e., a general classification of users into bully, no bully, we develop three feature vectors for each user's personal descriptors, social network metrics and content-specific features, respectively.

• Classification was done with all three feature sets (user, content, social network) as well as different combinations of these sets in order to determine the relative contribution of the various features to classification accuracy. Results of these experiments are reported. Because this task canters on flagging users rather than individual posts, we take the average over all posts by the user to determine the values for content-specific features and consider an aggregate representation of network structure, so that two users are linked in the social network graph if they are linked at any point in time throughout the lifespan of the dataset.

• Bullies are identified by hand-labellers if they are responsible for at least one post which falls within the bounds of what the labellers consider cyberbullying, in particular if that content is offensive, rude or could be considered harassment.

Classification of Pairwise Influence-

- While what we seek is ultimately to predict the influence of one user on another in the system, the claim that we can truly observe influence is bold, so we use subsequent occurrences of bullying under conditions as a stand-in for the notion of influence as follows.
- The initial behaviour of user prior to any exposure to bullying as a proxy for baseline Pi0, making the assumption that the behaviours of a given user outside of any contact with a bully is inherent. Specifically, a positive instance of influence over edge eij in our dataset is one in which the following conditions all hold to be true:

  - User i posts a bullying comment in the thread
  - User j has observed history of no bullying (baseline)
  - User j posts a bullying comment in the thread, following the comment of user i

Results: Summarizes the results of our classification of influence over edges of the user interaction graph. We run classification with and without social network features, as well as with and without textual/content features.

Conclusion:

In this work, we characterize the influencer/influenced relationship by which a user observes a peer engaging in bullying and follows suit. In pinpointing the nature of this transmission at the local level, we introduced a framework for large-scale analysis of bullying dynamics over the social network.

Our work indicates promise for the inclusion of contextual features beyond just a bag-of-words text analysis for the detection of cyber bullies, interactions involving cyber bullies, and ultimately the characterization of cyberbullying dynamics in a social graph as well as empirical evaluation of the proposed global aggregate model for cyberbullying dynamics and possible cascade effects. Future Scope: Follow-up work should include an in-depth comparative empirical analysis with other datasets and additional features, as well as empirical evaluation of the proposed global aggregate model for cyberbullying dynamics and possible cascade effects.

2.10Semi-supervised Learning for Cyber-bullying Detection in Social Networks:

Title: Semi-supervised Learning for Cyber bullying Detection in Social Networks

Published by: Vinita Nahar, Sanad Al-Maskari, Xue Li, and Chaoyi Pang

Introduction:

Current studies on cyber-bullying detection are mainly focused on Supervised learning approaches, which rely on a human-intensive labeling process of data. Feature space is uniformly applied to a learner. Whereas, streaming text generated by Social Networks (SNs) is highly uncertain, noisy, and imbalanced. In such a changing environment, different training data samples may have varying levels of importance. Therefore, with the rapid growth of user-generated content in SNs, existing supervised approaches become unaffordable and impractical for automatic detection of cyber-bullying instances. In this paper, we focus on the detection of cyber-bullying in streaming text generated by SNs. For such detection the following challenges are identified.

1. Insufficient training instances
2. Uncertain and imbalance feature distribution

Related works:

Recently, Xu et al. explored regret behaviors in bullying messages assuming that people who posted bullying tweets may later want to delete those posts. They reported cross validation accuracy up to 60.7%. Dadar et al. used content-based, cyber-bullying, and user-based feature sets. The best recall obtained (55.0% recall, 77.0% precision, and 64.0% F1 measure) with user-based and pronoun-profanity window feature sets. Dinakar et al. deconstructed cyber-bullying detection into sensitive-topic detection, which is likely to result in bullying discussions, including sexuality, race, intelligence, and profanity. Using SVM, the accuracy archived is 79% under the topic sexuality. Nahar et al. utilized probabilistic features and user ranking, and achieved 99% accuracy. Yin et al. utilized various features including content, sentiment, and contextual features, showing 59.5% recall, 35.2% precision, and 44.4% F1 measure. However, these methods are conducted under supervised learning by directly applying the whole input feature space to a learner These techniques are unable to handle the imbalanced and noisy data, where some features are either irrelevant or less important for the decision function. In this paper, we introduce semi-supervised learning for cyber-bullying detection in streaming text.

Proposed Work:

1. Feature Space Modeling
2. Cyber-bullying Detection
3. By Using Fuzzy Approach
4. Clustering Process
5. Fuzzy Classifier

Results:

If we try to reduce the false positive, then the false negative increases. This is because discrimination of the positive features and the negative features is very vague. In the training data, we observe that many likely cyberbullying words are quite frequent in both cyberbullying and non-cyberbullying categories. Ignoring those words on one hand reduces the false positives, while on the other hand it increases the false negatives. Our objective is to reduce the false negatives; therefore, our system tolerates the false positives but maintains low false negatives. Nevertheless, in this experiment the objective was to achieve high Recall, which is achieved up to 79.3% Overall K-FSVM achieved the best results in both experiments. Moreover, in Scenario 4, when the positive to negative ratio is 1.5%, Random Forest maintains a very high precision at 93%, whereas, Naïve Bayes achieved the highest recall 92%. Such observation opens a future direction to combine both classifiers to improve the systems performance significantly. In scenario 5, K-FSVM outperformed all other methods in terms of precision (55%) and F1 measure (47%), whereas, Naive Bayes achieved 97% recall and Logistic regression achieved poor results.

Conclusion:

(i) We devised a new framework for automatic detection of cyberbullying for the streaming data with insufficient labels. The framework extracts reliable positive and negative instances by augmented training methods based on the confidence voting function.

(ii) The enriched feature sets were generated based on user context, linguistic knowledge, and baseline keywords were also incorporated during feature space design in the proposed method.

(iii) We also proposed a fuzzy SVM algorithm for the effective cyberbullying detection. The proposed method effectively tackles the dynamic and complex nature of the streaming data.

(iv) The experiments conducted under the different scenarios demonstrate that the proposed technique outperformed the traditional methods use for cyberbullying detection.

2.11Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis:

Title: Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis

Published by: Junyi Chen, Shankai Yan, Ka-Chun Wong

Introduction:

Sentiment analysis and opinion mining task is one of the well-studied fields in text mining and natural language processing. It aims at detecting and analyzing human opinions, attitudes, and emotions. Application scenarios of sentiment analysis can stem from product reviews, advertisement distribution, stock market, social networks or even government intelligence. Many research and famous datasets of sentiment analysis such as IMDB and Yelp acquaint positive or negative opinion as known as PNO about a certain object from user comments. The dataset included manually collected paragraphs and paragraphs from 'Sentimen140' with labels renovated. Despite those two methods achieve good results on our verbal offense dataset; we were looking for models that can outperform our previous ones for verbal aggression detection. Convolutional neural networks (CNN) are originally designed to process and learn information from image features by applying convolution kernels and pooling techniques which are widely adopted for extracting stationary features; for instance, CNN has shown its adaptability in the field of text mining and NLP tasks. Kim et al. reported series of experiments with CNNs that achieve good results on sentence classification and sentiment analysis tasks. Lee et al. propose a weakly supervised CNN architecture to identify discriminating keywords in PNO tasks. Inspired by the successful examples of CNN applications in the field of text classification, we introduce a CNN model to detect verbal offenses from the aggression dataset we collected in the previous research to look for performance enhancement. The contribution of this work is to further improve the performance of the sentiment analysis task we previously proposed by introducing an efficient CNN-based deep learning model. In addition, by testing different kinds of models and methods, we discovered some interesting CNN architectures which can outperform others.

Related works:

To tackle this highly concerned problem, we propose a text classification model based on convolutional neural networks for the de facto verbal aggression dataset built in our previous work and observe significant improvement, thanks to the proposed 2D TF-IDF features instead of pretrained methods. Experiments are conducted to demonstrate that the proposed system outperforms our previous methods and other existing methods. A case study of word vectors is

carried out to address the difficulty in using pre-trained word vectors for our short-text classification task, demonstrating the necessities of introducing 2D TF-IDF features. Furthermore, we also conduct visual analysis on the convolutional and pooling layers of the convolutional neural networks trained.

Proposed Work:

System modeling-

1. Model architecture
2. Word features

Experiment settings-

1. Dataset and preprocessing
2. Learning algorithms and models
3. Performance benchmarking

Results:

We can be informed that, although embedding method reserves word level and ordinal information, it compromises the prediction performance. Both CNN and LSTM with embedding layer output worse performance compared with TF-IDF feature at the document level.

Conclusion:

In this paper, we present a new solution to the verbal aggression detection task we aroused in the prerequisite research based on convolutional neural networks (CNN) using 2-dimensional TF-IDF features and observe significant improvement. Firstly, experimental results indicate that CNN model achieves significant improvement compared with the baseline SVM and logistic regression methods in the previous study as well as the newly tested LSTM model in the problem. Moreover, we carried out experiments on the dataset to explain the selection of word lemmatizing. Finally, the problem that pre-trained word.

2.12Cyber bullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder

Title: Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder

Published by: Rui Zhao and Kezhi Mao

Introduction:

Social Media, as defined in is "a group of Internet based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content." Via social media, people can enjoy enormous information, convenient communication experience and so on. However, social media may have some side effects such as cyber bullying, which may have negative impacts on the life of people, especially children and teenagers. In this paper, we investigate one deep learning method named stacked denoising autoencoder (SDA). SDA stacks several denoising auto encoders and concatenates the output of each layer as the learned representation. Each denoising autoencoder in SDA is trained to recover the input data from a corrupted version of it. The input is corrupted by randomly setting some of the input to zero, which is called dropout noise. This denoising process helps the autoencoders to learn robust representation.

Related Work:

Text Representation Learning In text mining, information retrieval and natural language processing, effective numerical representation of linguistic units is a key issue.

Cyber bullying Detection With the increasing popularity of social media in recent years, cyber bullying has emerged as a serious problem afflicting children and young adults.

Proposed Work:

As a side effect of increasingly popular social media, cyber bullying has emerged as a serious problem afflicting children, adolescents, and young adults. Machine learning techniques make automatic detection of bullying messages in social media possible, and this could help to construct a healthy and safe social media environment. In this meaningful research area, one critical issue is robust and discriminative numerical representation learning of text messages.

In this paper, we propose a new representation learning method to tackle this problem. Our method named semantic-enhanced marginalized denoising auto-encoder (sm SDA) is developed via semantic extension of the popular deep learning model stacked denoising autoencoder (SDA). The semantic extension consists of semantic dropout noise and sparsity constraints, where the semantic dropout noise is designed based on domain knowledge and the word embedding technique. Our proposed method can exploit the hidden feature structure of bullying information and learn a robust and discriminative representation of text. Comprehensive experiments on two public cyberbullying corpora (Twitter and My Space) are conducted, and the results show that our proposed approaches outperform other baseline text representation learning methods.

Result:

The results are obtained using one layer architecture without non-linear activation considering the raw terms directly correspond to each output dimension under such a setting. It is shown that these reconstructed words discovered by smSDA are more correlated to bullying words than those by mSDA. For example, fucking is reconstructed by because, friend, off, gets in mSDA. Except off, the other three words seem to be unreasonable. However, in smSDA, fucking is reconstructed by off, pissed, shit and of. The occurrence of the term of may be due to the frequent misspelling in Internet writing. It is obvious that the correlation discovered by SMS DA is more meaningful. This indicates that SMS DA can learn the words' correlations which may be the signs of bullying semantics, and therefore the learned robust features boost the performance on cyber bullying detection.

Conclusion:

This paper addresses the text-based cyber bullying 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 smSDA as a specialized representation learning model for cyber bullying detection. In addition, word embedding's 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 cyber bullying 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.

2.13 Cyber BERT: BERT for cyber bullying identification

Title: Cyber BERT: BERT for cyberbullying identification

Published by: Sayanta Paul, Sriparna Saha

Introduction:

Online social media platforms allow people to share and express their thoughts and feelings freely and publicly with others. This can appear as an assortment of tech-empowered exercises, e.g., photo sharing, blogging, social gaming, social video sharing, business networks, comments & reviews, and many others. The information available over these social media is a rich resource for sentiment analysis or inferring other increasing uses and abuses. This increasing growth of social networking introduces continuous harassment and stalking which is commonly referred to as cyber bullying.

Classifying texts into specific categories is an ideal problem in Natural Language Processing (NLP). The important intermediate steps involve neural architecture design and data representation using word embeddings. This deep language representation has always been a crucial factor for efficient text categorization. Although BERT has achieved amazing results in many natural language understanding tasks, e.g., next sentence prediction, language inference, and so on, here, in this work, we have shown how BERT can be deployed to accomplish cyber bully detection task. The organization of this paper is as follows: a brief survey of previous works solving the cyber bullying detection task has been exhibited in the following section.

Related works:

Identifying cyber bullying over social media has been an increasingly trending issue over the past few years. A great number of research activities have been published, trying to address this problem in social networks, and in various forms. In the literature, several works can be found towards providing an effective solution for identifying cyber bullying via text-based, image-based, or video-based, sometimes incorporating multimodality, as well. Natural Language processing (NLP) and other language technologies have shown their potential performance for solving problems like detection of hate-speech, fake news, harassment, cyber bullying, and abusive language. y. This system achieved 78.5% accuracy, came up with paragraph-to-vector-based distributed representations of comments over Yahoo social media that can easily detect hate-speech achieving a sound accuracy of 80%. Introducing precise cyber bullying identification, i.e., Racism, Sexism, and others using, CNN and LSTM architecture over Twitter data which achieved 93% F1-score. Recently, in 2020 the developed automatic cyber bullying detection mechanism using Twitter users' psychological features which include personalities,

sentiment, and emotion. Further, expanded the objective task by introducing rapidly changing vocabulary of social interactions.

Proposed Work:

The internal architecture of BERT is multi-layer bidirectional Transformer encoder, which uses bidirectional language models to learn general language representations. The input representation of BERT can represent both a single sentence and a pair of sentences in one token sequence. Here, a "sequence" is referred to the input token sequence to BERT, which can be a single sentence or multiple sentences packed together. BERT comprises of hundreds of millions of parameters, while preceding baseline models use much less parameters and perform much faster. Therefore, BERT comes with a huge computational cost. To minimize the cost of the network, we have used a simpler version of BERT, called knowledge distillation method. This method compresses information learned by a large model to a comparably small model. We have compared our fine-tuned BERT model against CNN, LSTM, Bi LSTM with attention layer, and also with two popular traditional machine learning-based text classification models. Machine learning models are trained on TF-IDF vectors of the document and those are implemented using Scikit-Learn 0.22.2. Py Torch 1.4.0 is used as the backend framework. Randomly sampled 80% of the data have been used for training and 10% each has been used for validation and testing, respectively.

Results:

All the reported results above are statistically significant as we have performed statistical t test at 5% significance level. Thus, to ensure that no ambiguity was introduced during training, the experiments were conducted for five times. In the data description section, we have seen that, for each of the corpora, the instances of bully class are very less in comparison to total number of non-bully instances.

Conclusion:

This work forges ahead the state-of-the-art in cyber bullying identification in more fine-grained way over various social media platforms. In this paper, we have conducted extensive experiments to investigate the fine-tuning approach of BERT for the cyber bullying detection task. The experimental results show that the performance of proposed BERT model is reasonably good. As the further extension of the present work, we are planning to expand the area of social networking sites by providing with our framework as a text-based automatic cyber bullying detection tool along with exploring the usage of combining extrinsic knowledge with BERT model. The future scope of this work also includes incorporating different modality of information, e.g., images, videos, and audio data.

2.14 ALBERT based fen tuning model for cyber bullying analysis

Title: ALBERT based fen tuning model for cyberbullying analysis

Published by: Jatin Karthik Tripathy, S. Sibi Chakkaravarthy, Suresh Chandra Satapathy Madhulika Sahoo, V. Vaidehi

Introduction:

The world is evolving as a result of technological progress. The present pandemic situation has put the entire human race into digital beings such as virtual and embodied agents, although not a part of the natural human habitat, but has become essential elements of human life for survival. Furthermore, with the surge of online media networking through many diverse forms such as social networks and blogs, the capability of understanding of context behind languages has also risen. With multiple uses wherein information acquired being used for sentiment analysis or extracting named entities, the misuse of online platform has increased rapidly, and it is evident that cyberbullying is very commonly found in social media, instant messaging service, SMS and email. Studies, have evident that cyberbullying is quite prevalent form of interpersonal aggression in today's modern society and therefore is an important topic for intervention and prevention.

Related works:

• In this paper, we introduced a new fne-tuning model based on ALBERT for cyberbullying detection in the benchmarked datasets.

• In this paper, we explore the effectiveness of downstream fne-tuning for the specific application of cyber bullying classification. We found experimentally that using ALBERT-large gave us the best possible returns when compared to the train times. While theoretically using a larger model, such as ALBERT-XXL or GPT-3, should increase the results, the results obtained were largely unimpressive when comparing with the ALBERT-large implementation.

Proposed Work:

ALBERT is a transformer-based architecture and thus even at its untrained form provides better contextual understanding than other recurrent units. This coupled with the fact that ALBERT is pre-trained on a large corpus allowing the flexibility to use a smaller dataset for fne-tuning as the pre-trained model already has deep understanding of the complexities of the human language. ALBERT showcases high scores in multiple benchmarks such as the GLUE and Squad showing that high levels of contextual understanding are inherently present and thus fne-tuning for the specific case of cyber bullying allows to use this to our advantage. With this approach, we have achieved an F1 score of 95% which surpasses current approaches such as the CNN+wordVec, CNN+GRU and BERT implementations. Keywords: ALBERT, Fine tuning, Deep learning,

worded, Gated recurrent unit, GRU •, CNN.

Results:

While the CNN + GRU implementation has the added benefit of being able to combine both word and character features, the heavy pre-training of ALBERT puts our implementation above the rest in terms of contextual understanding. Even while considering the BERT-Base+CNN and GPT-2 approaches, ALBERT-large gives better results because parameter sharing and increased initial pre-trained dataset allow for much more contextual understanding that of other pre-trained models. Hence taking into account the different implementations and their scores, we show that our fne-tuning-based approach for contextual understanding is more efficient thus giving better classification results and will be able to beat all other implementations.

Conclusion:

This paper also does comparative analysis on some of the different strategies used in the past for the same issue and hopes to shed insight on the motives behind each iteration. With that, we show ALBERT-based fne-tuning performs better and that inefficiencies of using older architectures such as the GRU and BERT are due to lack of higher levels of contextual understanding and being pre-trained on a smaller corpus, respectively. We back this claim by using the dataset provided, where we have achieved an F1 score of 95% in an effective manner in terms of computational time and hardware requirements

2.15 Effective hate-speech detection in Twitter data using recurrent neural networks

Title: Effective hate-speech detection in Twitter data using recurrent neural networks

Published by: Georgios K. Pitsilis1 • Heri Ramampiaro1 • Helge Langseth

Introduction:

Social media is a very popular way for people to express their opinions publicly and to interact with others online. In aggregation, social media can provide a reflection of public sentiment on various events. Unfortunately, any user engaging online, either on social media, forums or blogs, will always have the risk of being targeted or harassed via abusive language, expressing hate in the form of racism or sexism, with possible impact on his/her on-line experience, and the community in general. The existence of social networking services creates the need for detecting user-generated hateful messages prior to publication. Any published text that is used to express hatred towards groups with the intention to humiliate its members is considered a hateful message.

Related works:

The supervised learning models also include the Deep Neural Networks (DNNs). Their power comes from their ability to find data representations that are useful for classification and they are widely explored to handle NLP tasks. Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN) are the two main architectures of DNNs, which NLP has benefited from. CNNs are suited for multi-dimension input data sampled periodically, in which a number of adjacent inputs are convoluted into the next layer in the network.

Proposed Work:

i) A deep learning architecture for text classification in terms of hateful content, which incorporates features derived from the users' behavioral data,

ii) A language agnostic solution, due to no-use of pre-trained word embeddings, for detecting hate-speech,

iii) An experimental evaluation of the model on a Twitter dataset, demonstrating the top performance achieved on the classification task. We put special focus on investigating how the additional features concerning the users' tendency to utter hate-speech, as expressed by their previous history could leverage the performance. To the best of our knowledge, there has not been any previous study on exploring features related to the user's tendency in hatred content that has used a deep learning model.

Results:

The most interesting results from our experiments. We used standard metrics for classification accuracy, suitable for studying problems such as sentiment analysis. We used Precision and Recall, with the former being calculated as the ratio of the number of tweets correctly classified to a given class over the total number of tweets classified to that class, while the latter measuring the ratio of messages correctly classified to a given class over the number of messages from that class.

Conclusion:

This paper has made several main contributions in order to advance the state-of-the-art. First, we have developed a deep learning architecture that uses word frequency vectorization for implementing the above features. Second, we have proposed a method that, due to no-use of pre-trained word embeddings, is language independent. Third, we have done thorough evaluation of our model using a public dataset of labeled tweets, an open-sourced implementation built on top of Kera's. This evaluation also includes an analysis of the performance of the proposed scheme for various classes of users. The experimental results have shown that our approach outperforms the current state-of-the-art approaches, and to the best of our knowledge, no other model has achieved better performance in classifying short messages. Also, the results have confirmed the original hypothesis of improving the classifier’s performance by employing additional user-based features into the prediction mechanism.

REFERENCES:

1.A. S. Srinath, H. Johnson, G. G. Dagher, and M. Long, ‘‘BullyNet: Unmasking cyberbullies on social networks,’’ IEEE Trans. Computat. Social Syst., vol. 8, no. 2, pp. 332–344, Apr. 2021, doi:10.1109/TCSS.2021.3049232.

2. Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski,‘‘Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection, ’’Inf. Process. Manage., vol. 58, no. 4, Jul. 2021, Art. no. 102600, doi:10.1016/j.ipm.2021.102600.

3. N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A. Sharma, S. Rajeskannan, M. Soni, G. S. Gaba, M. A. AlZain, and M. Masud, ‘‘Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking,’’ Math. Problems Eng., vol. 2021, pp. 1–12, Feb. 2021, doi: 10.1155/2021/6644652.

4. B. A. Talpur and D. O’Sullivan, ‘‘Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in Twitter,’’ Informatics, vol. 7, no. 4, p. 52, Nov. 2020, doi: 10.3390/informatics7040052.

5. R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson,‘‘Careful what you share in six seconds: Detecting cyberbullying instances in vine,’’ in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2015, pp. 617–622, doi: 10.1145/2808797.2809381.

6. A. Al-Hassan and H. Al-Dossari, ‘‘Detection of hate speech in Arabic tweets using deep learning,’’ Multimedia Syst., Jan. 2021, doi:10.1007/s00530-020-00742-w.

7. B. A. H. Murshed, H. D. E. Al-ariki, and S. Mallappa, ‘‘Semantic analysis techniques using Twitter datasets on big data?: Comparative analysis study,’’ Comput. Syst. Sci. Eng., vol. 35, no. 6, pp. 495–512, 2020, doi:10.32604/csse.2020.35.495.

8. N. M. G. D. Purnamasari, M. A. Fauzi, Indriati, and L. S. Dewi, ‘‘Cyberbullying identification in Twitter using support vector machine and information gain-based feature selection,’’ Indones. J. Electr. Eng. Comput. Sci.,vol. 18, no. 3, pp. 1494–1500, 2020, doi: 10.11591/ijeecs.v18.i3.pp1494-1500.

9. A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, ‘‘Identification and characterization of cyberbullying dynamics in an online social network,’’in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, Aug. 2015, pp. 280–285, doi: 10.1145/2808797.2809398.

10. V. Nahar, S. Al-Maskari, X. Li, and C. Pang, ‘‘Semi-supervised learning for cyberbullying detection in social networks,’’ in Databases Theory and Application (Lecture Notes in Computer Science), vol. 8506. Cham, Switzerland: Springer, 2014, pp. 160–171, doi: 10.1007/978-3-319-08608-8_14.

11. J. Chen, S. Yan, and K.-C. Wong, ‘‘Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis,’’ Neural Comput. Appl., vol. 32, no. 15, pp. 10809–10818, Aug. 2020, doi: 10.1007/s00521-018-3442-0.

12. R. Zhao and K. Mao, ‘‘Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder,’’ IEEE Trans.Affect. Comput., vol. 8, no. 3, pp. 328–339, Jul. 2017, doi:10.1109/TAFFC.2016.2531682.

13. S. Paul and S. Saha, ‘‘CyberBERT: BERT for cyberbullying identification,’’ Multimedia Syst., no. 0123456789, Nov. 2020, doi: 10.1007/s00530-020-00710-4.

14. J. K. Tripathy, S. S. Chakkaravarthy, S. C. Satapathy, M. Sahoo, and V. Vaidehi, ‘‘ALBERT-based fine-tuning model for cyberbullying analysis,’’ Multimedia Syst., Sep. 2020, doi: 10.1007/s00530-020-00690-5.

15. G. K. Pitsilis, H. Ramampiaro, and H. Langseth, ‘‘Effective hate-speech detection in Twitter data using recurrent neural networks,’’Appl. Intell., vol. 48, no. 12, pp. 4730–4742, Dec. 2018, doi: 10.1007/s10489-018-1242-y.