Online Public Shaming on Twitter Detection Analysis and Mitigation

ABSTRACT

Public shaming in online social networks and related online public forums like Twitter has been increasing in recent years. These events are known to have devastating impact on the victim’s social, political and financial life. Notwithstanding its known ill effects, little has been done in popular online social media to remedy this, often by the excuse of large volume and diversity of such comments and therefore unfeasible number of human moderators required to achieve the task. In this paper, we automate the task of public shaming detection in Twitter from the perspective of victims and explore primarily two aspects, namely, events and shamers. Shaming tweets are categorized into six types- abusive, comparison, passing judgment, religious/ethnic, sarcasm/joke and what aboutery and each tweet is

classified into one of these types or as non-shaming. It is observed that out of all the participating users who post comments in a particular shaming event, majority of them are likely to shame the victim. Interestingly, it is also the shamers whose follower counts increase faster than that of the non-shamers in Twitter. Finally, based on categorization and classification of shaming tweets, an web application called

BlockShame has been designed and deployed for on-the-fly muting/blocking of shamers attacking a victim on the Twitter.

**EXISTING SYSTEM**

* Sood et al. [3] examine the effectiveness of list based profanity detection for Yahoo! Buzz comments. Relatively low F1 score (harmonic mean of precision and recall) of this approach is attributed to distortion of profane words with special characters (e.g., @ss) or spelling mistakes and low coverage of list words. The first caveat was partly overcome by considering words as abusive whose edit distance from a known abusive word equals the number of “punctuation marks” present in the word. Rojas-
* Galeano [4] solves the problem of intentional distortion of abusive words in order to avoid censorship by allowing homo-glyph (characters which are similar in appearance, e.g., ‘a’ and ‘a’) substitution to bear zero penalty in calculating edit distance between an abusive word and a distorted word, thereby increasing recall rate substantially. Hate speech, though well defined as- “Abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation” [7], is often used in several other connotations (e.g., in [6]). Warner and Hirschberg [8] attempt to identify hate speech targeting Jews from a data set consisting of Yahoo!
* Comments and known anti-Semitic web page contents. A similar type of work has been done on anti-black hate speech on Twitter [9]. Burnap and Williams [10] collected tweets for two weeks after the Lee Rigby incident [11] and trained a classifier on typed dependency and hateful terms as features. Waseem and Hovy [12] released a public data set of sixteen thousand tweets labeled in one of the three categories- racist, sexist or none. They achieved an F1 score of 0.73 using character n-grams with logistic regression. Recently, Badjatiya et al. [13] reported F1 score of 0.93 using deep neural networks on the same data set.
* Dinakar et al. [19] employ Open Mind Common Sense (OMCS) [20], a common sense knowledge database, with custom built assertions related to specific domain of interests, e.g., LGBT cyberbullying, to detect comments which deviate from real world beliefs and is a good indicator of subtler forms of bullying. For instance, asking a male which beauty saloon he visits can be a case of bullying as OMCS tells that beauty saloons are more likely to be associated with females. Additionally, the authors propose several techniques to counter these incidents ranging from delaying posts, issuing explicit warnings, etc., to educating users about cyberbullying. Stressing the difference between cyberbullying and other forms of cyberaggression,
* Hosseinmardi et al. [21] consider instagram pictures with a minimum of fifteen comments of which more than 40% contain at least one profane word, to account for repetitiveness of bullying. Their best performing classifier uses uni-gram and tri-gram text features with image category (e.g., person, car, nature, etc.) and its meta data to achieve an F1 score of 0.87.

**Disadvantages**

* + There is no accurate analysis lack of Classification using Support Vector Machine
  + Only text classification and there is no sentiment analysis for different online public shaming.

**PROPOSED SYSTEM**

In the proposed system, the system proposes a methodology for the detection and mitigation of the ill effects of online public shaming. We make three main contributions in this work- (a) Categorization and automatic classification of shaming tweets

(b) Provide insights into shaming events and shamers

(c) Design and develop a novel application named Block Shame that can be used by a Twitter user for blocking shamers

**Advantages**

* The System is very effective due to AUTOMATED CLASSIFICATION OF SHAMING TWEETS.
* The System provides Analysis in the presence of Classification using Support Vector Machine.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL