**LITERATURE SURVEY**

**1) Ensemble fake profile detection using machine learning (ML),**

**AUTHORS:**  E. Karunakar, V. D. R. Pavani, T. N. I. Priya, M. V. Sri, and K.

Tiruvalluru

The social network, a crucial part of our life is plagued by online impersonation and fake accounts. According to the ‘Community Standards Enforcement Report’ published by Facebook on March 2018, about 583 million fake accounts were taken down just in quarter 1 of 2018 and as many as 3-4% of its active accounts during this time were still fake. In this project, we propose a model that could be used to classify an account as fake or genuine. This model uses Support Vector Machine (SVM) as a classification technique and can process a large dataset of accounts at once, eliminating the need to evaluate each account manually. The community of concern to us here is Fake Accounts and our problem can be said to be a classification or a clustering problem.

**2) Deep learning to filter SMS spam**

**AUTHORS:** P. K. Roy, J. P. Singh, and S. Banerjee

The popularity of short message service (SMS) has been growing over the last decade. For businesses, these text messages are more effective than even emails. This is because while 98% of mobile users read their SMS by the end of the day, about 80% of the emails remain unopened. The popularity of SMS has also given rise to SMS Spam, which refers to any irrelevant text messages delivered using mobile networks. They are severely annoying to users. Most existing research that has attempted to filter SMS Spam has relied on manually identified features. Extending the current literature, this paper uses deep learning to classify Spam and Not-Spam text messages. Specifically, Convolutional Neural Network and Long Short-Term Memory models were employed. The proposed models were based on text data only, and self-extracted the feature set. On a benchmark dataset consisting of 747 Spam and 4,827 Not-Spam text messages, a remarkable accuracy of 99.44% was achieved.

**3) Automatically dismantling online dating fraud**

**AUTHORS:** G. Suarez-Tangil, M. Edwards, C. Peersman, G. Stringhini, A.

Rashid, and M. Whitty

Online romance scams are a prevalent form of mass-marketing fraud in the West, and yet few studies have presented data-driven responses to this problem. In this type of scam, fraudsters craft fake profiles and manually interact with their victims. Because of the characteristics of this type of fraud and how dating sites operate, traditional detection methods (e.g., those used in spam filtering) are ineffective. In this paper, we investigate the archetype of online dating profiles used in this form of fraud, including their use of demographics, profile descriptions, and images, shedding light on both the strategies deployed by scammers to appeal to victims and the traits of victims themselves. Furthermore, in response to the severe financial and psychological harm caused by dating fraud, we develop a system to detect romance scammers on online dating platforms. This paper presents the first fully described system for automatically detecting this fraud. Our aim is to provide an early detection system to stop romance scammers as they create fraudulent profiles or before they engage with potential victims. Previous research has indicated that the victims of romance scams score highly on scales for idealized romantic beliefs. We combine a range of structured, unstructured, and deep-learned features that capture these beliefs in order to build a detection system. Our ensemble machine-learning approach is robust to the omission of profile details and performs at high accuracy (97%) in a hold-out validation set. The system enables development of automated tools for dating site providers and individual users.

**4) Segregating spammers and unsolicited bloggers from genuine experts on twitter**

**AUTHORS:** M. U. S. Khan, M. Ali, A. Abbas, S. U. Khan, and A. Y. Zomaya

Online Social Networks (OSNs) have not only significantly reformed the social interaction pattern but have also emerged as an effective platform for recommendation of services and products. The upswing in use of the OSNs has also witnessed growth in unwanted activities on social media. On the one hand, the spammers on social media can be a high risk towards the security of legitimate users and on the other hand some of the legitimate users, such as bloggers can pollute the results of recommendation systems that work alongside the OSNs. The polluted results of recommendation systems can be precarious to the masses that track recommendations. Therefore, it is necessary to segregate such type of users from the genuine experts. We propose a framework that separates the spammers and unsolicited bloggers from the genuine experts of a specific domain. The proposed approach employs modified Hyperlink Induced Topic Search (HITS) to separate the unsolicited bloggers from the experts on Twitter on the basis of tweets. The approach considers domain specific keywords in the tweets and several tweet characteristics to identify the unsolicited bloggers. Experimental results demonstrate the effectiveness of the proposed methodology as compared to several state-of-the-art approaches and classifiers.

**5) Improving cyberbullying detection using twitter users**

**AUTHORS:** V. Balakrishnan, S. Khan, and H. R. Arabnia

Empirical evidences linking users’ psychological features such as personality traits and cybercrimes such as cyberbullying are many. This study deals with automatic cyberbullying detection mechanism tapping into Twitter users’ psychological features including personalities, sentiment and emotion. User personalities were determined using Big Five and Dark Triad models, whereas machine learning classifiers namely, Naïve Bayes, Random Forest and J48 were used to classify the tweets into one of four categories: bully, aggressor, spammer and normal. The Twitter dataset contained 5453 tweets gathered using the hashtag #Gamergate, and manually annotated by human experts. Selected Twitter-based features namely text, user and network-based features were used as the baseline algorithm. Results show that cyberbullying detection improved when personalities and sentiments were used, however, a similar effect was not observed for emotion. A further analysis on the personalities revealed extraversion, agreeableness, neuroticism and psychopathy to have greater impacts in detecting online bullying compared to other traits. Key features were identified using the dimension reduction technique, and integrated into a single model, which produced the best detection accuracy. The paper describes suggestions and recommendations as to how the findings can be applied to mitigate cyberbullying.