International Journal of Scientific Research and Engineering Development-– Volume 5 Issue 3, May-June 2022

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Machine Learning-Based System for Text Filtering of Cyberbullying Contents on Twitter

Nwala, K. T.[1], Longe B.O[2], Adekunle, Y. A[3], Ogunlere, S. O[4].

[1]Department of Computer Science Babcock University, Nigeria, nwala1126@babcock.edu.ng
[2]Faculty of Computational Sciences and Informatics Academic City University College, Accra, Ghana, olumide.longe@city.edu.gh
[3]Department of Computer Science Babcock University, Nigeria, adekunlea@babcock.edu.ng
[4]Department of Computer Science Babcock University, Nigeria, ogunleres@babcock.edu.ng

Abstract

Cyberbullying is prevalent in most countries across the globe. The aim of this research is to develop a predictive model to identify the occurrence of cyberbullying tweets on Twitter. A decade ago, the internet of Things underwent a paradigm shift, resulting in a massive increase in the number of active users. This figure has now surpassed three billion. Websites for social networking are classic instances of Internet applications with a significant number of active users. With 330 million active users at any given time, Twitter is one of the most well-known social networking sites. Unfortunately, it is also a stage for users who engage in unethical internet behavior, and cyberbullying has become a worldwide phenomenon. It has a negative impact on the people who are affected by it. As a result of the shame and contempt connected with cyberbullying, some victims have committed suicide. With the use of psychological studies pertaining to individual behavior and the use of dialects relevant to verbal aggressiveness, a model will be constructed in this study, and each word in a tweet will be separately categorized based on the pragmatics of language. Linguistic Inquiry and Word Count (LIWC), a psychometric evaluation instrument that categorizes text based on Linguistic Processes, Psychological Processes, Personal Concerns, and Spoken Categories, was used to achieve this. They make up a total of 67 sub-word categories. The psychometric evaluation will proceed to the next step. The degree to which different word categories were utilized by people in cyberbullying was calculated by LIWC. As a result, psychometric evaluation aided in effective text categorization and quantifying the degree of word usage, which had previously been noted as a gap in previous research. As a consequence, tweets were transformed into a relational numeric dataset with multi-dimensional attributes. This dataset was extremely comprehensive in terms of the data it contained. This dataset was then utilized to train a machine learning classifier in order to construct a predictive model for cyberbullying detection.

Introduction

Amidst the rapid revolution of social media in the world, it has become paramount that social media applications such as Facebook, Twitter, Orkut, Myspace, and Skype are used extensively for the purpose of communication. Communication through Facebook, Twitter, Orkut, Myspace, and Skype can be with a person or a group of persons via the internet. Today, some people especially the young people are addicted to the different social media sites for keeping in contact with their peers (Shuen, 2008). There are merits and demerits to the use of social media. One important advantage is the online sharing of knowledge and information among the different groups of people using these social media sites (Shuen, 2008). This online sharing of information also promotes the

increase in the communication skills among the people,especially among students in educational institutions.

Furthermore, offensive language has risen to become a major issue for both online communities and their users' well-being. Viewing inappropriate language has a harmful impact on users' mental health, especially for children and teenagers.(Bauman & Walker, 2013). When offensive language is found in a user communication, the mechanism for removing the offending language becomes the primary problem.Cyberbully, a type of aggressive language, is a common social problem that has a harmful impact on people's life in countries all over the world. Cyberbullying, for example, is when someone is harassed on the internet. Bullies on the internet usually target a single person or a group of people as their victims (Paridhi& Ashish, 2013). Cyberbully is most commonly found on social networking sites, where victims are humiliated. Based on psychological surveys conducted in several nations, many government and non-government groups have highlighted the negative impacts of cyberbullying on victims (Bauman & Walker, 2013). Suicidal tendencies have been reported as a result of cyberbullying in some circumstances, and in a few severe cases, victims have committed suicide as a result of cyberbullying (Bauman & Walker, 2013).

On the one hand, there are those who do not match the "norm" and do not belong to the hegemonic dominant mainstream majority. As a result, the majority of cyberbullying victims are of non-dominant sex (in this example, non-male and non-cisgender), race, ethnicity, and nationality (Balakrishnan, Khan, &Arabnia, 2020). (Balakrishnan, Khan, &Arabnia, 2020). Twitter, for example, already has a tool that detects cyberbullying automatically. It uses machine learning to detect psychological characteristics in its users (Balakrishnan et al., 2020). Normal, spammer, bully, and aggressor are the four sorts of Twitter users recognized (Balakrishnan et al., 2020). Hate speech is defined as textual sharing on social media with the intent of inciting hatred toward an individual or group. (Charitidis, Doropoulos, Vologiannidis, Papastergiou, &Karakeva, 2020)

Cyberbullying has recently taken a toll on Twitter, one of the most prominent social networking platforms. Twitter is a microblogging and social networking service where users can send 140-character messages known as tweets. Twitter had over 336 million active users as of April 2019 (Shuen, 2008). A user can send direct messages to other users as well as broadcast messages via Twitter. Any tweet on Twitter creates a chain reaction. The followers of the user who posts a tweet can see it, as well as the followers of the user who receives the tweet. Furthermore, if one of the followers retweets the original tweet, his or her followers will be able to view it. The growing number of retweets spread the message like wildfire on Twitter, which has been termed “Going viral” in social media (Georgios, Pitsils,&Langseth, 2018).

Because of the large number of active users on Twitter, cyberbullying has become a worldwide phenomenon. Cyberbullying on Twitter appears to be on the rise on a daily basis, according to the trend. This increase could be linked to web 2.0, a fundamental paradigm shift in the Internet that occurred a decade ago (Shuen, 2008). It ushered in significant advancements in internet application usability, availability, flexibility, and portability. As a result, the number of people using the internet has increased dramatically. In 2018, 3.9 billion people used the internet, accounting for 51.2 percent of the world's population (ITU, Statistics, 2018). This figure was around 400 million in 2000. (ITU, Measuring the information Society Report, 2014). This increase in the number of users has led to a confounding-amount of information flow on the internet, generally referred to as “big data”. The flexibility of “Web 2.0” applications has bridged the communication gap between users across continents. Furthermore, with the advent wireless technology, people have

access to mobile internet devices that can be used to transmit information from anywhere. As a service, the internet is now considered a boom. However, as the number of internet users grows, it is becoming clear that many users are abusing the system by engaging in behaviors such as cyberbullying, cyber terrorism, E-Commerce fraud, misleading marketing and advertising, privacy breaches, unethical hacking, and identity theft, to name a few.In line of the above discussion, this study would employ predictive analytics to detect hostile content on Twitter. Effective text classification of the dataset is required for predictive analysis, which may subsequently be utilized to construct a model to detect cyberbullying tweets. To detect cyberbullying from tweets, it's first required to specify the data properties that could work as predictors. Previous research has revealed this as a significant flaw. However, the main focus of this paper will be on defining the data qualities that might be utilized to characterize bullying tweets and filter the message. Individuals' writing patterns can provide insight into their social behavior patterns, according to three decades of psychological research. (Miller &Rollnick, 2013). As a result, the text classification of the data in this study will be based on language pragmatics. Every person has his or her own distinct writing style, which means that the pragmatics of language utilized vary from person to person (Beck, 2011). The words employed in composing text, on the other hand, are divided into four categories: Linguistic Process, Psychological Processes, Personal Concerns, and Spoken Categories (Beck, 2011). There are 67 sub-categories of words based on these four psychometric qualities.

Psychometric Analysis

Psychometrics, generally, is viewed as the field of study that evaluates behavioral contrasts of people. It entails two major tasks: the creation of tools to evaluate psychological factors, as well as estimations based on the data collected from these measures. In a nutshell, the field of psychometrics is concerned with the development of a behavioral or psychological scale and the analysis of the data obtained from this scale.

Thus, an estimation study utilizing factual system, and inferring estimations by examining this measurable information, is characterized as the study of psychometrics (Brown, 2000). For any logical study to propel, that requires evaluation process, its strategy must be founded on a strong develop of instruments for estimation. These instruments are for the most part estimations and look at the general enormous of the research being led. For the most part, development of these tool needs exact definition and also brings about incorrect estimations. Also, the subsequent estimation is available to huge mistakes. Hence, the estimation tool utilized with the end goal of measurement includes rehashed attempts in different various ways. When creating relations between the wide scopes of factors, every factor is estimated more than once, storing up countless figures. In this manner, breaking down mental estimations dependent on a factual methodology is general multivariate (Browne, 2000).

Psychometric assessment has valuable application. For example, it may be utilized to decide a person's character (Chamorro-Premuzic&Furnham, 2014). This can be utilized to assess one's qualities and shortcoming that by and large outcome in careful heading of their subjective capacities and general social conduct style (Kline, 2013). Because of this, numerous organizations while procuring perform psychometric assessment on contender to distinguish potential counterpart for explicit employment job. It additionally helps in distinguishing aptitudes of people wherein; they can determine certain vocation areas explicit to singular test. Subsequently, Psychometric assessment can recognize different mental ramifications behind each individual's style of writing (Chamorro-Premuzic&Furnham, 2014). Further, psychometric assessment can be inferred on writings, aggregately produced by different individuals on a comparative point. Twitter gives its users the capacity to utilize 'hashtags' that by

and large diverts to a particular inclining point. Various individuals may have various suppositions relating to that particular pattern. NLP is utilized to separate between those assessments of various people (Cambria, Schuller, Xia, &Havasi, 2013). Distinctive NLP modules can separate, arrange and help recognize the degree with which these individual sentiments contrast (Zhang, 2014). This depends on separating utilization of words by different individuals wherein, each word falls in a word-class relating to a particular psychometric property. The four essential psychometric properties involve Linguistic procedure, psychological procedure, individual concerns and Spoken classes (Tausczik&Pennebaker, 2010). Henceforth, psychometric assessment can recognize psychometric qualities of various individuals that triggers a particular sentiment in them about a particular thing. Hence, it tends to be utilized to gauge the degree with which these psychometric properties vary for various individuals. Returning to a similar model from presentation; "This cake is so yummy", and "This is a delicious cake". These statements have diverse grammatical structure however both show a specific 'preferring' towards a cake. Psychometric assessment reveals to us that these comparable feelings trigger perceptual process which is additionally positive. Additionally, in Cyberbullying, various individuals utilize various styles of writing. To put it plainly, the word utilization varies yet they inevitably demonstrate Cyberbullying. In this research, the fundamental center is to distinguish explicit psychometric properties that are utilized to pass on Cyberbullying. Understanding tormenting conduct or mindset is an extent of concentrate because of incorporation of different factors identified with pragmatics of language. As a result, psychometric testing contributed in the development of a scale that established a pattern that machine learning may use to develop a model to detect cyberbullying. The chance of occurrences of a Cyberbullying tweet on Twitter was recognized using this approach, which was based on the psychometric traits it possessed. This is in contrast to previous methods for detecting cyberbullying that rely on raw text features like filthy or swear words.

Psychometric evaluation using LIWC

Linguistic Inquiry and Word Count (LIWC) is a metric that measures how much people use different types of words in a variety of contexts, such as messages, discourse, sonnets, and everyday conversation. It gives you the option to focus on how much any exam content uses beneficial or unfavorable assumptions, as well as how much up close and personal notices, causal terms, and various other vernacular assessments are used. (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007)

LIWC can distinguish between a variety of typical ASCII content records and Microsoft Word documents in terms of semantic and behavior measures. Pronouns, prepositions, and articles, for example, are linguistic qualities of text. In addition, categories such as positive and negative emotions, anger, and sadness are psychological dimensions of text.It moreover allows you to manufacture your very own pledge classification references that can be based upon these semantic and conduct measurements to separate and examine the content especially critical to your study (Pennrbaker et al., 2007).

LIWC is capable of predicting the many structural, emotional, perceptual, and cognitive components present in people's spoken written discourses (Pennebaker et al., 2007)

LIWC applications are designed to look into and analyze written content on a word-by-word basis, compute the rate at which the words appear in the content by matching them to the 67-word categories provided in the LIWC default dictionary, and save the results as a tab-delimited document that can be read into applications like Microsoft Excel (Pennebaker et al., 2007).

The LIWC dictionary contains lexicon of words and word classifications, recently referenced as semantic measurements, which classified words ought to be counted in the objective content document. Words read and analyzed by LIWC are target. Words in the LIWC inbuilt default lexicon are word reference words. Gatherings of dictionary words relating to a specific dimension, for example positive feeling words, are characterized as word categories (Pennebaker et al., 2007)

LIWC analyzes the text document word by word from beginning to end for classification purposes. It then looks in its lexicon for a match with a dictionary word and assigns it to the appropriate category. When a target word matches a dictionary word, the scale for that category is increased (Pennebaker et al., 2007).

Let us scale back to the going with model to comprehend the LIWC content taking care of module. Consider a ten-word statement (w1, w2,..., w10) that should be structured into five word characterizations (wc1, wc2,..., wc5) to be explicit. As shown in Figure, the substance handling module will assign each target word to a vocabulary word. As previously stated, the suitable word scale is incremented after each match. (w4), on the other hand, does not correspond to any word arrangement. Figure 1 shows that wc4 does not rise as a result of the fact that none of the target words arrange that particular class.

Since w1, w2, and w3 were all classified as wc1, this simply means that three out of ten words have expanded the definition of that word. As found in table 3, LIWC creates yield for each word class dependent on the accompanying recipe in Equation 1.

$$V(wc_n) = \frac{N}{TWC} \%$$

Equation 1 –LIWC Word Category Output
Where,

"V (wc_n)" is the output value generated by "LIWC" for a particular-word-category,
"N" isthe number ofword-categorized in that particular-word-category, and
"TWC"stands for total word count.
By implementing this "formula" for each word-classification, the out result-document of LIWC looks like that shown in Table 2

Table 1 –LIWC Example Output

Wc1	Wc2	Wc3	Wc4	Wc5
30%	20%	20%	0%	20%

So far, we have seen how LIWC helps in effective word categorization by using its inside default dictionary to analyzetext and categoryevery-target word to its different word-order. What's more, LIWC processes complete level of words in some random content that have a place with a particular word-classification. Hence, it empowers the user to comprehend the style of composing winning in textual records. It is these styles of composing instead of explicit word use that people use in Cyberbullying, on which the prescient model is based upon in this research.

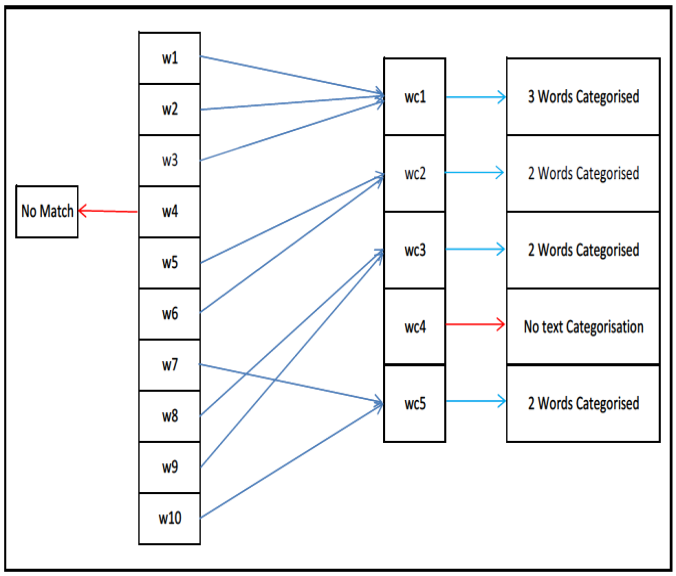

Figure 1 – Text Categorization in LIWC(Tausczik&Pebbbaker 2010)

The previously mentioned model is downsized altogether to comprehend the operational

utilization of LIWC content arrangement. In any case, (Tausczik&Pebbbaker 2010) provides comprehensive large-scale models in which they have utilized LIWC for successful content order of content reports containing more than a huge number of words. Multiple judges with relevant knowledge made a choice based on relevant facts in pragmatics of language along with numerous meetings to produce new ideas, ultimately deciding which words must be remembered for explicit word-classifications while building up the LIWC lexicon. The decisions were based on: the majority of judges' understanding of 'votes,' and the distinction between unwavering quality of words falling into similar classifications using Cronbach's alpha. Finally, they used individual correlational tests to approve LIWC content arrangement on a few hundred million words, and they discovered a very strong link between LIWC scales and judges' ratings (Tausczik and Pennebaker, 2010).

LIWC

The utilization of LIWC is a strong way tocategoriesany composed content in 67 language measurements altogether. These language measurements are known as the psychometric properties of text. The yield of LIWC was a tab-delimited document thatcategorizedtext in 67 psychometric classifications of words forthe existing system. These words will be stored in the database, categorically in a specified table. Each individual tweet is alluded to as an occasion. In the following stage, LIWC was utilized to break down the recently saved categorized tweets exclusively and afterward two separate yield records were produced. These table will independently put something aside for additional adjustment.

Lexicon Building

The dictionary is initially filled in a dictionary table in the database in a number of the recommended techniques. The inability to find opinion words with domain and context particular orientations is a common problem with dictionary-based techniques. Syntactic, structural, and sentence level features are used in determining the semantic categories of words and phrases to be included in an opinion lexicon using Linguistic Integration word count processing rule-based techniques. By integrating contextual variables that could potentially modify the semantic orientation of an opinion term, a lexicon is stocked with words and phrases that are more attuned to the domain. Intensifiers can increase or decrease the intensity of a surrounding lexicon item, while negations like no and never can modify the directionality of a lexical item.

Lexicon for Hostile words

Following is a description of the method of extracting and building a semantic dictionary for the hostile domain, based on the definition of hostile given above. The semantics of hostile words not only include typical opinion words with negative and positive polarities, but also employs rich linguistic stylistic devices. Haters have no shortage of vocabulary to express their nihilistic intentions, from similes to metaphors to juxtapositions. While some publications employ clear provocation and aggressive language, others utilize less explicit phrasing that may be confusing if just opinion-oriented words are used. Three separate sets of characteristics are used in the rule-based hostile words classifier that we eventually develop.

Hostile word filtering algorithm

Input: slist: An initial seed list of hate verbs, dv: A set of all verbs in the hostile words

Output: hlex: A set of lexicon of hostile words.

//initialize h_lex with s_list

Hostile_word_lexicon←slist

//create a set s and initialize

Set s←{}

For each word □slist

s=Getsynset() and Gethypernyms()

For each si□s

If(si appears in dv)

Add si to hlex
End if
End for
End for

For mining association rules, the performance of this approach is comparable to that of the classic Apriori algorithm. For each loop, the important steps include obtaining a seed list, obtaining a list of semantically related terms using synsets and hypernym associations, and comparing them to the entire lexicon dictionary using type threads or user comments. The number of terms in the hatred lexicon and the size of the seed list generated at each iteration affect how long these processes take.

Using Machine Learning

In context of the data prediction step, it is important to choose the correct approach for tackling the task appropriately. This is often done using machine learning methods. A major difference between humans and computers has been for a long time that a human beings tend to automatically improve their way of tackling a problem. Humans learn from previous mistakes and try to solve them by correcting them or looking for new approaches to address the problem. Traditional computer programs do not look at the outcome of their tasks and are therefore unable to improve their behavior. The field of machine learning addresses this exact problem and involves the creation of computer programs that are able to learn and therefore improve their performances by gathering more data and experience.

We apply machine learning classification methods in this study. The task of approximating the mapping function from input variables to discrete output variables is classified predictive modeling. The basic goal is to figure out which category or class the new data belongs to. For the experiments, we adopted several different classification approaches that were selected due to their extensive use, well-understood behavior, and promising results in a range of categorization tasks. Our goal was to examine classifiers that differ in terms of the functional forms of classification boundaries they may learn, as well as classifiers that are based on distinct assumptions about the relationship between distinct features. We studied the performance of a set of classification models including decision tree, random forest, support vector machine (SVM), feedforward neural networks (MLP), logistics regression, Bagging, AdaBoost, XGBoost.

Predict subjUser algorithm
Input: d: Text Document, sl: SUBJ lexicon
Output: count: count of subjective words
//initialize list and count
Subjsentlist: list of subjective sentences
count←0
sentence ← “”
Begin
While (d! = null)
sentence ← split (d)
word← “”
lexicon_Dictionary ← “”
For each sentence ϵ d
For each word ϵ sentence
word←CRFtagger(word)
If word matches lexicon_Dictionary then
Count++
If count >=2 then
Output count
End if


```
                                addSubjectSentlist(s
                                entence)
                            End if
                        End for
                    End for
End while
```

Figure 3. Subjective hostile word Prediction

Results and Discussion

The final censored sentence was produced as a result of series of test. When a user is about to post or comment on a particular thread; the algorithm of the developed system will loop through the words in sentence that is about to be posted or commented to check if there is a hostile word(s) by looking up to a predefined hostile word in a dictionary. If any of the words in the sentence contained a hostile word; it will temporary keep track of all the hostile words, with their category and the user information. Storing the hostile words if anyone was discovered in a sentence to help the system study individual writing pattern. All the hostile words discovered will be counted in other to know the number of asterisk (*) that will be used to replace the hostile words which shows it has been censored. The replacement is done using a regular expression; thereafter, the post or comment will be stored and immediately returned to the page using Ajax (Asynchronous Javascript).

Conclusion

This research examined the effectiveness of using a multi-dimensional training dataset on machine learning classifiers to predict cyberbullying tweets on Twitter. The multi-dimensional dataset was created based on the pragmatics of language. LIWC analyzed tweets individually in segments and categorized words from each tweet into 67 psychometric categories. The predictive models generated show that binary classifiers outperformed multiclass. The predictive model trained using the Random Forest classifier yielded 98.5% accuracy with a precision rate of 0.893 and a recall rate of 0.93. It is seen that the predictive ability of the classifiers deteriorated slightly when implemented on the testing dataset. However, the true positive rate could be improved by applying cost sensitive analysis to the classifiers. This indicates that the classifier algorithm had a high precision rate. The dictionary containing 67 psychometric categories and weighting based on Category Frequency Inverse Word Count (CF-IWC) formed the baseline of this study.The performance of SVM in training the predictive model made it a good choice to develop the predictive model.

Reference

1. Alessandro, M., Serena, P., Simonetta, V., & Pierluigi, V. (2017). Mining Offensive Language on Social Media. *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017* (pp. 252-256). Rome: Academia university press.
2. Anna, S., & Michael, W. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1-10). Valencia, Spain: Association for Computational Linguistics.
3. Barli´nska J, S. A., & M, W. (2018). Role of Affective Versus Cognitive Empathyin Increasing Prosocial Cyberbystander Behavior. *frontiers in Psychology*, 1-13.
4. Bauman, S. T., & Walker, J. L. (2013). Associations among bullying cyberbullying, and suicidal in high school students. *Journal of adolescence*, 341 – 350.
5. Beck, J. S. (2011). *Cognitive behavior therapy: Basics and beyond.* Texas, TX: Guilford press. .
6. Blackburn, J. K. (2014). *Predicting crowdsourced decisions on toxic behavior in online games.* STFU Noob.

7. Cynthia, V. H., Els, L., Ben, V., Julie, M., & Bart, D. (2015). Automatic Detection and Prevention of Cyberbullying. *ResearchGate*, 1-6.
8. Donald, F., & David, M. (1999). Ontological Problems of Pluralist Research Methodologies. *Association for Information Systems* (pp. 20-28). AIS Electronic Library.
9. Ehab A. Abozinadah, A. V., & Jr., J. H. (2015). Detection of Abusive Accounts with Arabic Tweets. *International Journal of Knowledge Engineering,*, 1-7.
10. Elizabeth, C., & Cherie, E. (2019). Embedding the dialogic in mixed method approaches to theory development. *International Journal of Research & Method in Education*, 239-251. Retrieved August 2, 2019, from https://orsociety.tandfonline.com/doi/abs/10.1057/ejis.2014.23#.XURquUco_IU
11. Fabio, P., Marco, S., & Manuela, S. (2017). Hate Speech Annotation: Analysis of an Italian Twitter Corpus. *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017* (pp. 263-268). Rome: Academia university press.
12. Georgios K. Pitsilis, H. R., & Langseth, H. (2018). Detecting Ofensive Language in Tweets using deep learning. *ArXiv*, 1-17.
13. Hossein, H., Sreeram, K., Baosen, Z., & Radha, P. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. *arXiv*, 30-45.
14. Isobelle, C., & Jack, G. (2017). Dimensions of Abusive Language on Twitter. *Proceedings of the First Workshop on Abusive Language Online* (pp. 1-10). Vancouver, Canada,: Association for Computational Linguistics.
15. ITU. (2014). Measuring the information Society Report. *International Telecommunication Union* (pp. 1-20). International Telecommunication Union. Retrieved July 4, 2019, from https://www.itu.int/en/Pages/default.aspx
16. ITU. (2018). *Statistics.* International Telecommunication Union. Retrieved July 4, 2019, from https://www.itu.int/en/Pages/default.aspx
17. J, W., Orlikowski, & Baroudi, J. J. (1991). Studying Information Technology in Organizations: Research Approaches and Assumptions. *Information System Research*, 1-28. Retrieved August 2, 2019, from https://doi.org/10.1287/isre.2.1.1

