International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 12 (2019), pp. 2330-2335

Automatic Cyber Bullying Detection in Arabic Social Media

Bedoor Y. AlHarbi[1], Mashael S. AlHarbi[1], Nouf J. AlZahrani[1], Meshaiel M. Alsheail[1,3], Jowharah F. Alshobaili[1] and Dina M. Ibrahim[1,2]

[1]*Information Technology Dept., College of Computer, Qassim University, Qassim, Saudi Arabia.*

[2]*Computers and Control Engineering Dept., Faculty of Engineering, Tanta University, Egypt.*

(ORCID: 0000-0002-2991-7299[3], 0000-0002-7775-0577[2])

Abstract

Cyberbullying is considered to be one of the cybercrimes that take a form of bullying or harassment using electronic means, also known as online bullying. It has become increasingly common, especially among teenagers usually happened on social media sites. Negative comments can have a serious impact on teenagers. For that, successful prevention depends on the detection of harmful messages automatically. Automatic cyberbullying detection in social media text can analyse the text, twits, by using sentiment analysis. There are two techniques that can be used for performing sentiment analysis in an automated manner, these are Rule-based, also known as lexicon or sentiment lexicon, and Machine Learning based. There is a shortage in the lexicons that support Arabic language especially in cyberbullying. The target of this paper is to generate cyberbullying lexicon based on the PMI, Chi-square, and Entropy approaches then compare between them to conclude which one is better for detecting cyberbullying in Arabic. The results show that the PMI approach gives the best performance in detecting cyberbullying in comparison with Chi-square and Entropy approaches.

Keywords: Cyberbullying, Sentiment Analysis, Arabic Lexicon.

I. INTRODUCTION

Social media allows people to interact easily exchange ideas, pictures, videos, and help each other such as Facebook, YouTube, Instagram, and Twitter. Twitter one of the most popular social media web/app around the world, which allowed people to interact via twits to express their thoughts or feelings about different subjects. However, you cannot control social media content poses so it is considered serious challenges with a huge amount of information where users can view the opinion of other users. Some users may write offensive twits to others which known as cyberbullying [1]. Online bullying is defined as the use of the Internet, cell phones, video game systems, or any other techniques to send messages, publish texts or images intended to harm or embarrass another person or group of people, or other deliberate action by one person or group of persons. Through digital means such as sending messages or posting comments against the victim. In recent years, someone has been able to express and share his point of view through social media. However, given the current situation, some users are negative for social media.

We are truly living in the information age where the data is generated by both humans and machines at an unprecedented rate, therefore it's nearly impossible to gain insights into such data for making intelligent decisions, manually. One such insight is assessing/calculating the sentiment of a big dataset [2]. Sentiment analysis can help detect cyberbullying after processing a large amount of data by using sentiment classification. Sentiment classification aims to automatically classify the text, and techniques can be roughly divided into machine-learning, lexicon-based, and hybrid approaches [3].

Machine learning uses algorithms to analyse data, learn from those data, and make decisions based on their learning. Supervised Machine learning is the search for algorithms that cause external resource situations to provide general hypotheses [4].

II. BACKGROUND ON LEXICON-BASED APPROACHES:

These approaches adopt a lexicon to perform sentiment analysis by counting and weighing sentiment words that have been evaluated and tagged. The most common lexicon resources are SentiWordNet, WordNet, and ConceptNet, and among these resources, SentiWordNet is the most widely used [3]. There are three approaches to generating a sentiment lexicon:

1) Corpus-Based Approach:

The corpus-based approach utilizes a corpus and a set of Sentiment bearing words. Words are extracted from the corpus and compared to the set of sentiment words using different statistical methods that measure semantic similarity. Statistical approaches that are commonly used include PMI, and Chi-Square [5].

2) Dictionary-Based Approach:

The dictionary-based approach as the name implies a dictionary is used by utilizing the synonym and antonym lists that are associated with dictionary words.

3) Manual Approach:

The technique starts with a small set of sentiment words as seeds with known positive or negative orientations. The seed words are looked up in the dictionary then their synonyms and antonyms are added to the seed set and a new iteration starts. The process ends when no new words are found. A manual inspection is usually done after the process ends to correct errors [6].

In this proposal, we will use Corpus-Based Approach because it looks at the context of the sentence so it can detect cyberbullying. Nowadays, researchers are also using combined approaches, in which two or more approaches are combined to achieve better accuracy. They combined lexicon-based and machine-learning methods by considering a lexicon as the source of features and using a classification model to evaluate the lexicon [3].

Sentiment lexicon is a database of lexical units for a language along with their sentiment orientations. Once such a lexicon is available, it can be used appropriately to perform sentiment analysis on a document, either alone or in combination with classifier methods [7].

- *English Sentiment Lexicons*

Sentiment analysis of English texts has become a large and active research area, with many commercial applications, but the barrier of language limits the ability to assess the sentiment of most of the world's population. Sentiment analysis in a multilingual world remains a challenging Problem because developing language-specific sentiment lexicons is an extremely resource-intensive process [8]. In [9] Liu introduces an efficient method, at the state of the art, for doing sentiment analysis and subjectivity in English. In [10] characterized a novel lexical resource that aids the process of cyberbullying detection. It follows from a linguistically motivated definition of textual cyberbullying that identifies its necessary and sufficient parameters.

- *Arabic Sentiment Lexicons*

The language on social media is known to contain slang, nonstandard spellings and evolves by time. As such sentiment lexicons that are built from standard dictionaries cannot adequately capture the informal language in social media text [6]. They relied on four resources to create ArSenL: English WordNet (EWN), Arabic WordNet (AWN), English SentiWordNet (ESWN), and SAMA (Standard Arabic Morpho- logical Analyzer). Two approaches were followed producing two different lexicons:

The first approach used AWN, by mapping AWN entries into ESWN using existing offsets thus producing ArSenL-AWN [6]. The second approach utilizes SAMA's English glosses by finding the highest overlapping synsets between these glosses and ESWN thus producing ArSenL-Eng. Hence ArSenL is the union of these two lexicons. Although this lexicon can be considered as the largest Arabic sentiment lexicon developed to date, it is unfortunate that it only has MSA entries and no dialect words and is not developed from a social media context which could affect the accuracy when applied on social media text.

The second approach is lexicon SLSA, (Sentiment Lexicon for Standard Arabic) (Eskander and Rambow, 2015) was constructed by linking the lexicon of an Arabic morphological analyser Aramorph with SentiWordNet. SLSA starts by linking every entry in Anamorph with SentiWordNet if the one-gloss word and POS match. Intrinsic and extrinsic evaluations were performed by comparing SLSA and ArSenL which demonstrated the superiority of SLSA. Nevertheless, SLSA like ArSenL does not include dialect words and cannot accurately analyse social Media text [6].

In [12], the authors used AraSenTi-Entropy, AraSenTi-ChiSq and PMI, the results showed that the performance of the lexicon that was generated using PMI outperforms other lexicons. In [6] using the AraSenTi-Trans lexicon, they used the simple method of counting the number of positive and negative words in the tweet and whichever is the greatest denotes the sentiment of the tweet. And the AraSenTi-PMI lexicon, the sentiment score of all words in the tweet were summed up. The natural threshold to classify the data into positive or negative would be zero, since positive scores denote positive sentiment and negative scores denote negative sentiment. The results showed the superiority of the AraSenti-PMI lexicon.

II. CYBERBULLYING DETECTION USING LEXICON SENTIMENT: PROPOSED WORK

With the development of modern technologies, it has become possible to discover cyberbullying. There is one research in Arabic to detect cyberbullying here we will mention some studies of lexicon sentiment in Arabic: The Author in [13] discussed how machines can detect cyberbullying. In this context, the text mining approach and lexicon-based approach to detect cyberbullying in Arabic text are discussed and the text mining approach, the machine will not use any dictionary of bad words. Through our search for all the researchers that applied the dictionary with electronic bullying in Arabic, we found only one search. In [13] presented how machines can detect cyberbullying. In this context, the text mining approach and lexicon-based approach to detect cyberbullying in Arabic text is discussed. There was no detail about the method used nor the apparent outcome.

In this section, we present our proposed framework as shown in Fig. 1. To generate sentiment lexicons to detection cyberbullying we propose two stages. The first stage defined as Dataset-based stage talks about data collection, including processing of data and classification to prepare it for the second stage defined as lexicon-based stage, which talks about the approaches to generate lexicon. Finally, evaluate three lexicons by using test dataset which achieved high accuracy to detection cyberbullying in Arabic text.

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 12 (2019), pp. 2330-2335

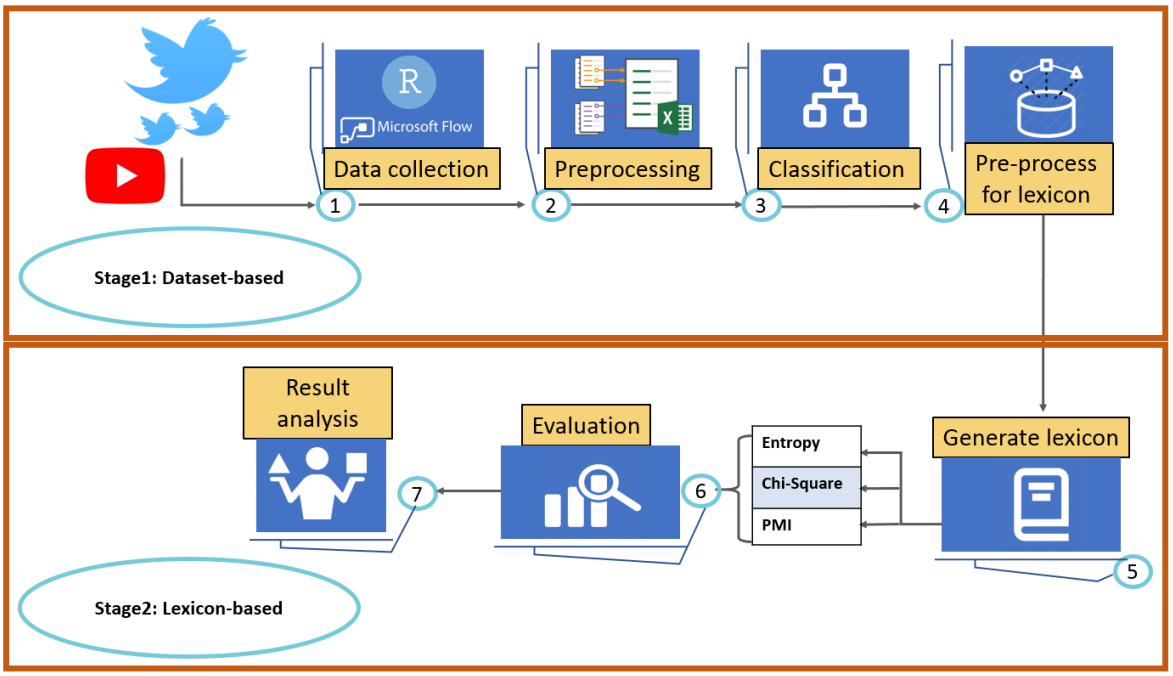

Fig. 1. Framework of the proposed work

II.I. DATASET-BASED STAGE

This is the first implementation step of any project. Acquiring data is an easy task since datasets are public available online, but acquiring an appropriate in Arabic and suitable data for specific domain can be very difficult task the proposed project is a classification, there was no dataset available online that meets the need of the project and is ready for classifying bullying or not, so raw data has been collected by using Twitter API, Microsoft-Flow and YouTube comments were then grouped into a single file containing about 100,327 tweets and comments. We used R language to extract the data. After collecting the data by divided it into three Excel files to facilitate the cleaning process: The first file contains 50,000 tweets. While the second file has 50,000 tweets and comments YouTube. And the third file contains 327 tweets from flow. The cleaning steps and data pre-processing are shown Fig. 2.

After the data cleaning step, we move on to the data classification step to bully or not. The data were classified as 1 containing bullying and 0 did not contain bullying. It was classified by three people and use an odd number of people to be the last classification after the majority opinion, the files were then separated into two files, a bullying file and a file other than bullying. After classification will do some work manually on the two files, bullying and non-bullying file to be prepared for lexicon.

Each tweet was transformed into separate words, formed in one column and each word in a cell with Excel, we count the number of iterations for each word as shown in Fig. 3. This step and the previous step was done on the bullying file and the non-bullying file.

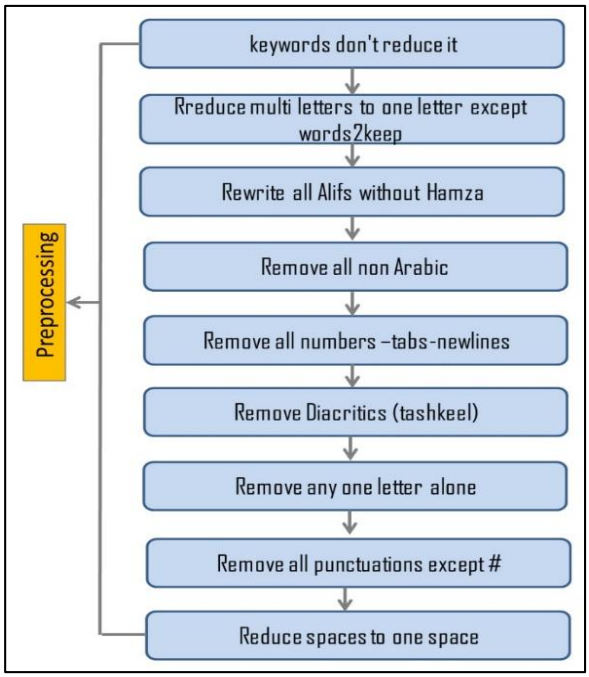

Fig. 2. Data pre-processing steps

	A	B	C	D	E	F	G	H	I
1	من	12275							
2	في	8442							
3	الله	6541							
4	على	5600							
5	لا	4771							
6	ان	4375							
7	ما	4373							
8	كل	3662							
9	يا	3339							
10	الي	3241							

Fig. 3. Sample of data after prepared for lexicon

II.I. LEXICON-BASED STAGE

In the lexicon generation phase, PMI, Chi-square, and Entropy approaches will be used. Each of the three approaches has a different code, but they all use the two files (Bullying and Non-Bullying files). Below is an explanation of the work of each approach to get the lexicon from them.

1) Entropy approach:

Entropy approach is often used in Information Theory to measure expected information content; in the case of two labels, entropy is highest when the data is evenly distributed, and lowest when all of the data is under one label [12]. Figure 4 illustrates a sample of Entropy lexicon output.

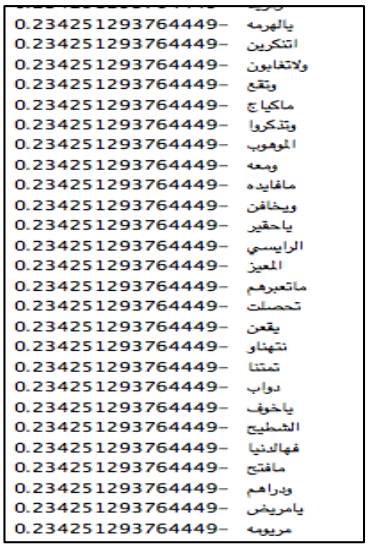

Fig. 4. Entropy lexicon output sample

2) A Chi-square approach:

This test is used to check the validity of some null-hypothesis by evaluating the statistical significance of the difference between observed and expected values [12]. In Fig. 5 a sample of Chi-square lexicon output is illustrated.

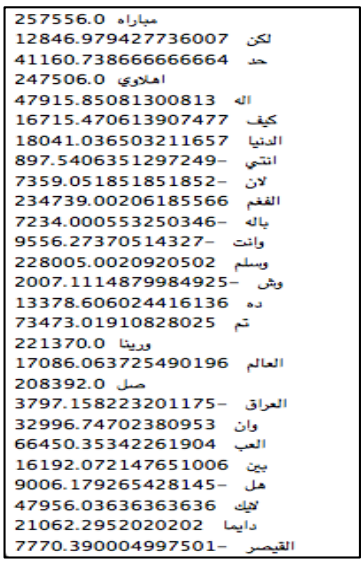

Fig. 5. Chi-square lexicon output sample

3) PMI approach:

Whether it has a positive or negative connotation can be measured by looking at whether that word co-occurs more with clearly positive words or clearly negative words here in Fig. 6 an output sample of PMI lexicon is appeared.

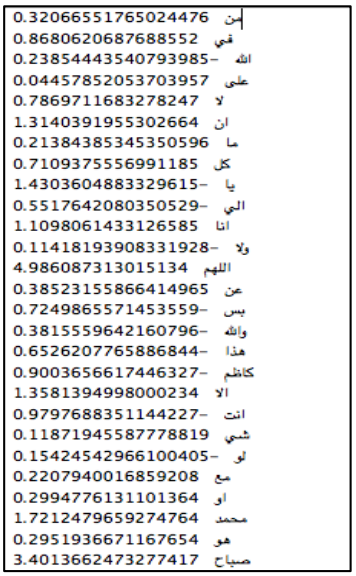

Fig. 6. PMI lexicon output sample

IV. RESULTS ANALYSIS AND EVALUATION

The test was performed on each tweet so that it takes a word by word and calculates the weight of it based on the lexicon that you summoned and then calculates the full weight of the tweet collecting the weight of each word in it as shown in Fig. 7. We took the weight of each word from the lexicon, if the value is greater than or equal to zero it is a non-bullying tweet and if negative, tweet bullying. Wherever the words have no weights; the lexicon calculated it as 0 value.

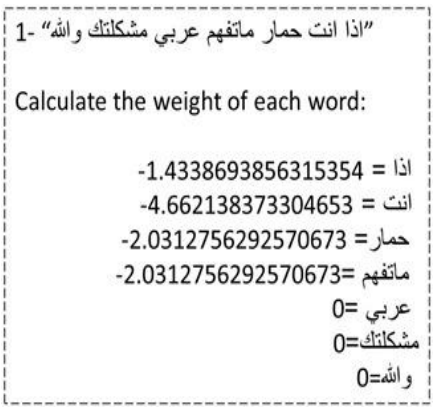

Fig. 7. Example of weight per word in PMI Lexicon

Meanwhile the datasets are unbalanced, we will measure the lexicon performance of the bullying categories by compute the precision (P_{bull}), the recall (R_{bull}), and the F-score (F_{bull}) as in the following formulas:

$$P_{bull} = \frac{TN}{TN+FN} \quad (1)$$

$$R_{bull} = \frac{TN}{TN+FP} \quad (2)$$

$$F_{bull} = \frac{2*R_{bull}*P_{bull}}{P_{bull}+R_{bull}} \quad (3)$$

We will repeat the same measurements of lexicon performance of the nonbullying categories by compute the precision ($P_{nonbull}$), the recall ($R_{nonbull}$), and the F-score ($F_{nonbull}$) as in the following formulas:

$$P_{nonbull} = \frac{TN}{TN+FN} \quad (4)$$

$$R_{nonbull} = \frac{TN}{TN+FP} \quad (5)$$

$$F_{nonbull} = \frac{2*R_{nonbull}*P_{nonbull}}{P_{nonbull}+R_{nonbull}} \quad (6)$$

Where *TP* is the number of true positives, *FP* is the number of false positives, *TN* is the number of true negatives and *FN* is the number of false negatives. Then we calculated the F-score (F_{avg}) for both values (Bullying and Non-bullying) to each lexicon as follows:

$$F_{\text{avg}} = \frac{F_{bull}+F_{nonbull}}{2} \quad (7)$$

The results of the three approaches PMI, Entropy, and Chi-square were calculated using the previous formulas, as in Table 1. A number of 100327 tweets were used to train three lexicons and 2020 tweets for testing. With this dataset, the results showed that the PMI outperformed 81% compared with Chi-square and Entropy, which give 62.11% and 39.14%, respectively. The results show that the PMI approach gives the best performance in detecting cyberbullying in comparison with Chi-square and Entropy approaches, as illustrated in Fig. 8.

Table 1. Result analysis of the three lexicon approaches for Bullying and Non-Bullying categories.

Lexicon approach	Bullying			Non-bullying			F_{avg}
	P_{bull}	R_{bull}	F_{bull}	$P_{nonbull}$	$R_{nonbull}$	$F_{nonbull}$	
PMI	57.8	75.7	65.5	78.9	62.2	69.5	81
Entropy	87.5	1.8	3.5	59.7	99.8	74.8	39.14
Chi-square	65.7	37.6	47.9	67	88.6	75.3	62.11

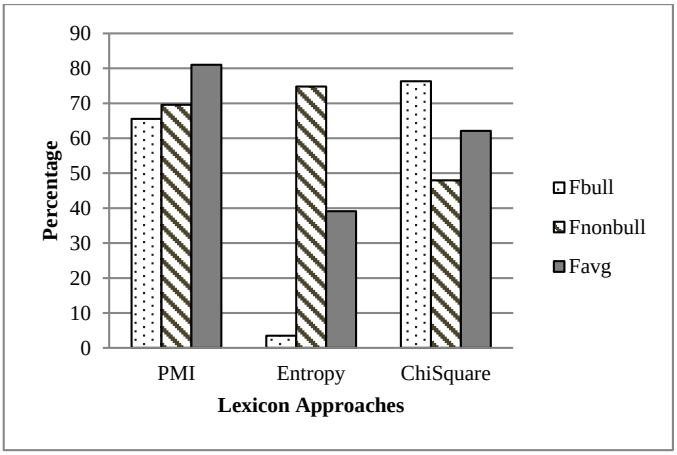

Fig. 8. The percentage of the F_{avg} in the three lexicons

V. CONCLUSIONS

This project proposed an automatic detection of cyberbullying by using sentiment analysis and lexicon approaches. This experimental work has been implemented using Java programming language and the dataset has been prepared for the experiment. Datasets have been collected from Twitter API, Microsoft-Flow, and YouTube comments. Then, they were grouped into a single file containing about 100,327 tweets and comments. After performing the data cleaning and pre-processing step, the data were classified to bullying non-bullying. It was classified by three people and use an odd number of people to be the last classification after the majority opinion. After the data is completed and configured for use in lexicon generation, we used PMI, Chi-square, and Entropy. The results show that the PMI approach gives the best performance in detecting cyberbullying in comparison with Chi-square and Entropy approaches.

REFERENCES

[1] Ting I. H., Liou W. S., Liberona D., Wang S. L., and Bermudez G. M. T., "Towards the detection of cyberbullying based on social network mining techniques," in 2017 International Conference on Behavioural, Economic, Socio-cultural Computing (BESC). IEEE, 2017, pp. 2-1.

[2] Janakiev N. Practical text classification with python and keras. [Online]. Available: https://realpython.com/python-keras-text-classification. [Accessed: 15 November 2019].

[3] Feng J., Gong C., Li X., and Lau R. Y., "Automatic approach of sentiment lexicon generation for mobile shopping reviews," Wireless Communications and Mobile Computing, vol. 2018,2018.

[4] Kotsiantis S. B., Zaharakis I., and Pintelas P., "Supervised machine learning: A review of classification techniques," Emerging artificial

intelligence applications in computer engineering, vol. 160, pp. 24-3, 2007.

[5] Li Q. and Shah S., "Learning stock market sentiment lexicon and sentiment-oriented word vector from stock twits," in Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL'17), 2017, pp. 310-301.

[6] Al-Twairesh N., Al-Khalifa H., and AlSalman A., "Arasenti: large-scale twitter- specific Arabic sentiment lexicons," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume :1 Long Papers), vol. 1, 2016, pp. 705--697

[7] Tang H., Tan S., and Cheng X., "A survey on sentiment detection of reviews," Expert Systems with Applications, vol. 36, no. 7, pp. 760-773, 2009**.**

[8] Chen Y. and Skiena S., "Building sentiment lexicons for all major languages," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume :2 Short Papers), 2014, pp. 389-383.

[9] Liu B. *et al*., "Sentiment analysis and subjectivity." Handbook of natural language processing, vol. 2, pp. 666-627, 2010.

[10] Power A., Keane A., Nolan B., and O'Neill B., "A lexical database for public textual cyberbullying detection," Revista de Lenguas para Fines Específicos, vol. 23, no. 2, pp. 186-157, 2017.

[11] Mahyoub F. H., Siddiqui M. A., and Dahab M. Y., "Building an Arabic sentiment lexicon using semi-supervised learning,',' Journal of King Saud University-Computer and Information Sciences, vol. 26, no. 4, pp. 424-417, 2014.

[12] AlNegheimish H., Alshobaili J., AlMansour N., Shiha R. B., AlTwairesh N., and Alhumoud S., "Arasenti-lexicon: A different approach," in International Conference on Social Computing and Social Media. Springer, 2017 pp. 235-226.

[13] Alduailej A. H. and Khan M. B., "The challenge of cyberbullying and its automatic detection in Arabic text," in 2017 International Conference on Computer and Applications (ICCA). IEEE, 2017, pp. 394-389.