Non-linguistic Features for Cyberbullying Detection on a Social Media Platform Using Machine Learning

YuYi Liu[1], Pavol Zavarsky[2(✉)], and Yasir Malik[2]

[1] Edmonton Public Schools, Edmonton, Canada
mandy.liu@epsb.ca
[2] Concordia University of Edmonton, Edmonton, Canada
pavol.zavarsky@concordia.ab.ca

Abstract. Cyberbullying on social media platforms has been a severe problem with serious negative consequences. Therefore, a number of researches on automatic detection of cyberbullying using machine learning techniques have been conducted in recent years. While cyberbullying detection has traditionally utilized linguistic features, the cyberbullying on social media does not have only linguistic features. In this paper, a holistic multi-dimensional feature set is developed which takes into account individual-based, social network-based, episode-based and linguistic content-based cyberbullying features. To test performance of the proposed multi-dimensional feature set, we designed and built cyberbullying detection models on the KNIME machine learning platform. Six different machine learning algorithms - Naïve Bayes, Decision Tree, Random Forest, Tree Ensemble, Logistic Regression, and Support Vector Machines - were used in our cyberbullying detection models. Our experimental results demonstrate that applying the proposed multi-dimensional feature set (i.e. the set not limited to the linguistic features) results in an improved cyberbullying detection for all tested machine learning algorithms.

Keywords: Cyberbullying detection · Cyberbullying features · Machine learning · Cyber safety

1 Introduction

Over the past few years, researchers and national organizations have been conducting studies to protect children and youths from cybercrimes. Progresses have been made in developing models to detect cyberbullying using machine learning techniques. It has been recognized that feature extraction of cyberbullying acts is the core component for an effective detection of cyberbullying. The major limitation of the existing models is that the feature extraction focuses primarily on linguistic analysis of bullying comments. However, considering cyberbullying as a behavior, the features of the behavior are not limited to linguistic features. For the reason, a holistic multi-dimensional feature set is developed in this paper based on our study and analysis of cyberbullying activities on a social media platform. The proposed multi-dimensional feature set can be used for an automatic detection of bullying incidents using machine learning and

J. Vaidya et al. (Eds.): CSS 2019, LNCS 11982, pp. 391–406, 2019.
https://doi.org/10.1007/978-3-030-37337-5_31

natural language processing techniques. The proposed multi-dimensional feature set takes into account individual-based, social network-based, episode-based and linguistic content-based features to detect cyberbullying on social media. Our experimental results confirm improvement in cyberbullying detection by using the proposed multi-dimensional feature set.

The remainder of this paper is organized as follows. Section 2 reviews the previous work on cyberbullying detection. In Sect. 3, the multi-dimensional feature set engineering for cyberbullying detection is proposed and justified. Section 4 discusses the design and construction of the cyberbullying detection model. The results of cyberbullying detection for three- and four-dimensional data sets processed by six machine learning algorithms - Naïve Bayes, Decision Tree, Random Forest, Tree Ensemble, Logistic Regression, and Support Vector Machines - are presented in Sect. 5. The performance of machine learning algorithms in detection of cyberbullying for the three- and four-dimensional sets are evaluated by the Precision, Recall, F1-measure, Accuracy, and Area Under Curve (AUC) metrices. We conclude the paper with final remarks and directions for future research in Sect. 6.

2 Related Work

Cyberbullying phenomenon, forms and impacts have been studied extensively in the realm of sociology and psychology. The theoretical interactional-normative framework for recognizing hostile content has been proposed in [2]. Different types of cyberbullying have been discussed in [3, 4]. Price and Dalgleish [1], Cowie [5] and Smith et al. [6] demonstrated the severe consequences and impact on youngsters induced by cyberbullying. 'Snowball effect' described in [7] illustrates that one single post can cause continuous harm to the victim if the post is reposted or liked by others.

Research on detecting and preventing cyberbullying has also made important advances in the recent years. As cyberbullying detection requires to distinct bullying from non-bullying posts, the dominant approaches are based on supervised algorithms with binary classifiers in the machine learning domain. The general solution is that the positive class represents post units containing cyberbullying, while the negative class includes posts containing non-bullying text. It is important in apply natural language processing approaches in cyberbullying detection research as the study object is mainly text generated by individuals.

Among the studies on cyberbullying detection, the chi-square information gain and odd ratio mutual information algorithms to detect and document evidence of email-based cybercrimes are explored in [8]. N-grams, Linguistic Inquiry and Word Count (LIWC) [10], Term Frequency/Inverse Document Frequency (TF/IDF), Part-of-Speech (POS) information [12], and Bag-of-Words (BoW) [11] have been applied in the detection of cyberbullying. More recent studies have demonstrated the value of considering other features, such as geoposition, time of publication [9] and network-based features. Moreover, cyberbullying detection in other languages than English has been explored by researchers in [14] for Arabic and in [16] for Dutch. Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) deep learning approaches have been used in cyberbullying detection models in [15].

Although the natural language processing techniques contributed to the development of cyberbullying detection, authors of [9, 13] found that language analysis of comments or postings is not enough to effectively detect cyberbullying. False positives are inevitable when only text features are fed to machine learning classifiers. The findings in [17] confirmed that a number of swear words were found in non-cyberbullying media conversation sessions. Authors of [7, 18, 19] proposed that cyberbullying feature study should be broadened to include both psychologic and behavioral analysis areas.

The key challenge in cyberbullying detection research is the feature set extraction which is essential for development of cyberbullying detection models. However, most of the cyberbullying detection methods are limited to studying linguistic characteristics of comments in cyberbullying activities. The holistic feature set for an effective detection of cyberbullying has not been developed. To address the gap, we present in this paper a multi-dimensional feature set engineering as an approach to improve the effectiveness of cyberbullying detection.

3 Multi-dimensional Feature Set for Cyberbullying Detection

Cyberbullying feature set development is the primary task and core component for the success of detection of cyberbullying on social media platforms by machine learning. The main idea underlying the feature set engineering proposed in this section is that a cyberbullying act on social media platforms can be detected by combining the natural language processing and machine learning techniques. Based on the definitional characteristics of cyberbullying, we propose the cyberbullying feature set with a structure of five dimensions and four layers as shown in Fig. 1.

The five-dimensional feature set shown in Fig. 1 has individual-based dimension, social network-based dimension, content-based dimension, episode-based dimension, and the "others" dimension. The details of the proposed five-dimensional feature set structure are provided in the following paragraphs.

3.1 Feature Set Layer Structure

The proposed cyberbullying feature set for machine learning based detection of cyberbullying has a layered structure. The first layer is formed by four main traits: (1) Participants trait, (2) Behaviour trait, (3) Technology trait, and (4) Sociology trait.

- *Participant's trait* reflects the power imbalance of the bully, victims and bystanders involved in one cyberbullying episode. In the trait, the post owner's age, gender, activeness, popularity, anonymity, and different roles in cyberbullying incident are considered.
- *Behavior trait* derives from the aggressiveness and repetition of cyberbullying. Attributes under this trait are the language linguistic characteristics, intention to spread the rumours, influence scope, episode duration, and inter-arrival time of negative comments.

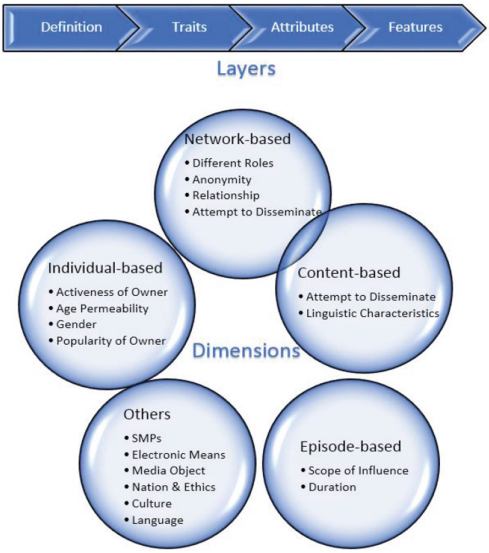

Fig. 1. Five-dimensional cyberbullying detection feature set structure

- *Technology trait* represents different online social media platforms' functions and regulations, diverse electronic means, the posting media types besides the text comments, such as pictures or video clips.
- *Sociology trait* considers national, ethnic and cultural differences. Besides cyberbullying posts in English, cyberbullying in other languages can be explored.

3.2 Feature Set Dimension Structure

Several datasets have become publicly available for cyberbullying research in the recent years. In our research, we adopted the labelled cyberbullying datasets on social media platform Instagram generously shared by the CU CyberSafety Research Center of the University of Colorado Boulder [20]. Using the dataset, we explored the Participant's and Behaviour Traits (see Fig. 2) that cover twelve Attributes with twenty-six Features to describe cyberbullying on the social media platforms. The twenty-six features were categorized into four dimensions: (1) individual-based, (2) social network-based, (3) content-based, and (4) episode-based dimension.

(1) ***Individual-Based Dimension***

In this dimension, see Figs. 1 and 2, we identified four Attributes with nine Features to differentiate cyberbullying postings by the participants.

(a) *Activeness of Owner*
We consider the online age as time since the user account was created in a given social media platform. The frequency of postings an account produced in the latest half-year can be used to estimate the activeness of the account user.

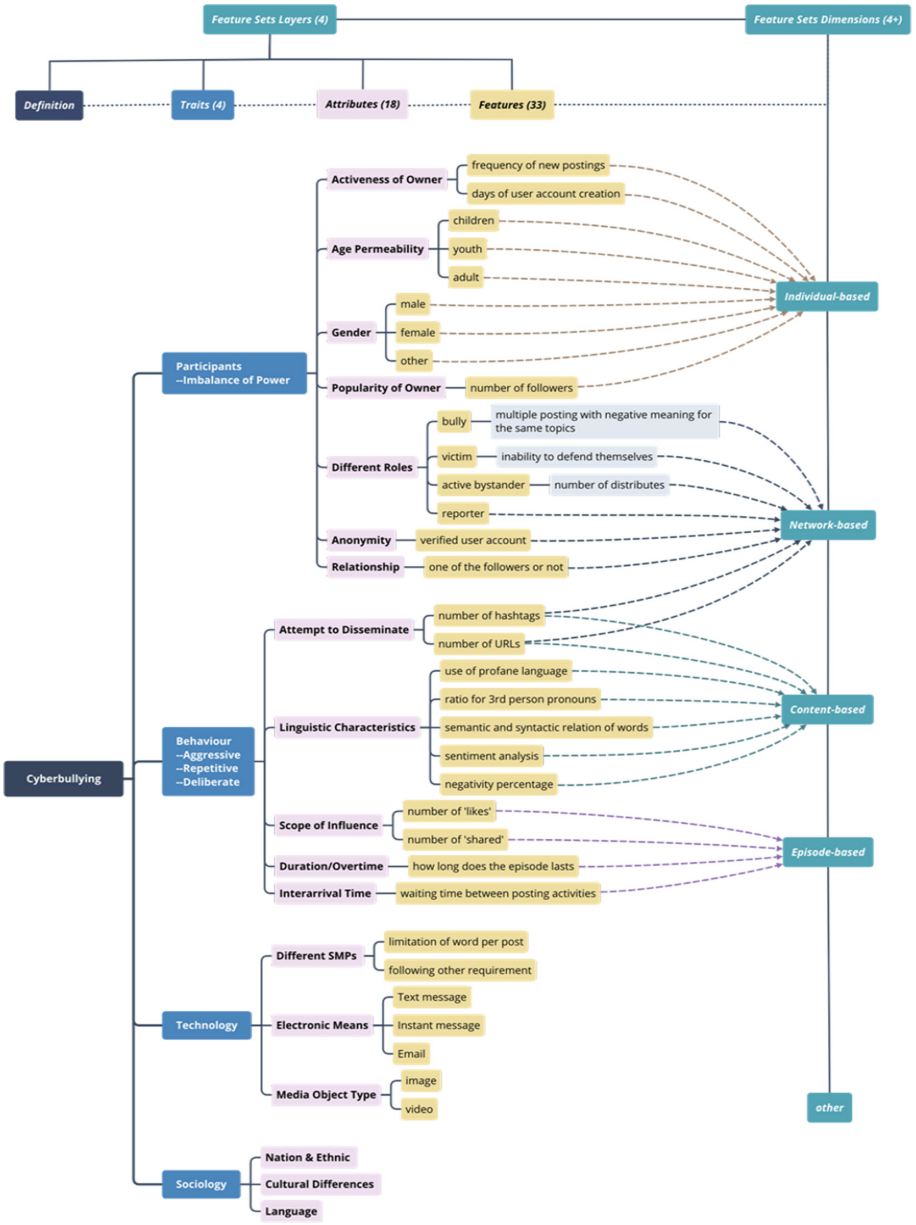

Fig. 2. Proposed multi-dimensional feature set for cyberbullying detection on social networks

(b) *Age Permeability*
We classify the participants of a cyberbullying episode into three categories based on the age range. Age under 14 are children, between 15 and 29 are youth, and 30 and over are classified into an adult group.

(c) *Gender*
We take into consideration genders (i.e., male, female, transgender, other) of the participants in the target episode involving a cyberbullying act.
(d) Popularity of Owner
The number of followers can quantify a user's popularity. Bullies are less popular in the perspective of fewer friends and followers than the typical users [13]. Another consideration is that the more followers a user has, the higher influence can be produced by the post owner.

(2) ***Network-Based Dimension***

In this dimension, we developed four Attributes with eight features of the cyberbullying social network characteristics.

(a) *Different Roles*
In the typical cyberbullying episode, the roles of the involvement could range from a bully, victim, bystanders, and reporters.
(b) *Anonymity*
On the Internet, people commonly use fake accounts to use negative, profane, or aggressive words in their posts.
(c) *Relationship*
Reciprocal follower means that the user and the follower follow each other on a social media platform. This metric quantifies the extent to which users interact with the follower connection they receive from other users [18]. The more interaction between them, the closer they are.
(d) *Attempt to Disseminate*
We consider the number of hashtags and Uniform Resource Locators (URLs) in the context of the posting. Typically, the bullies attempt to disseminate the cyberbullying behaviour by using more hashtags, URLs, and @s than the average users. More attention means more negative emotional experiences to victims [21]. We consider the number of hashtags and URLs in the comments or postings as one of the features of cyberbullying.

(3) ***Content-Based Dimension***

In the content-based dimension of cyberbullying features, the main focus is on linguistic characteristics.

(a) *Linguistic Characteristics*
The frequency of curse words, person pronouns, and positive words can be calculated by statistical algorithms. Furthermore, morphological, syntactic, and semantic analysis of dataset corpus can be performed. Natural language processing and data mining techniques, such as Bag-of-Words (BoW), latent semantic analysis (LSA), continuous Bag-of-Words (CBoW), skip-gram are practical word embedding methods that can be applied to represent words in vectors.

- Use of profane language
  We examine the number and frequency of abusive words in the postings and comments. For this purpose, word lists from noswearing.com [22] and the hatebase database [23] can be employed to score the extent of swear and hateful words on [0,100] scale. Besides profane words, other topics, such as religion, death, body, and sexual hints commonly have highly-frequent occurrence in cyberbullying postings.
- Use of the third person pronoun
  The occurrence of the third-person pronouns (i.e., he, she, and they) is higher than the use of the first person singular pronoun (i.e., I) in cyberbullying involving comments [17]. Therefore, we consider the use of third-person pronouns as a feature of cyberbullying.
- Semantic and syntactic relation of words
  Word embedding, a class of techniques which allows words with similar meaning to have a similar representation as real-valued vectors, can find both semantic and syntactic relation of words. We applied TF/IDF (term frequency, inverse document frequency) scheme to calculate weight of the importance of words.
- Sentiment analysis
  We consider metrics across the user's posting and other users' comments, such as the number of uppercase text which could indicate an intense emotion. SentiStrength is a tool to estimate the strength of sentiment in short texts from extremely positive (+5) to extremely negative (−5).
- Negativity percentage
  An interesting and unexpected founding in [17] is that most cyberbullying have the percentage of negativity in the comments between 50%–60%, rather than the higher percentage such as more than 60%–70%. We consider this pattern as one of the features to detect cyberbullying incidents.

(4) ***Episode-Based Dimension***

In this dimension, we identified three attributes to describe the cyberbullying phenomenon based on the episode criteria. Since the cyberbullying happens under a context, we set the threshold of fifteen comments for each episode.

(a) *Scope of Influence*

- Number of 'likes'
  The average number of likes per posting for non-cyberbullying is four times the average number for cyberbullying episodes [17], which means the cyberbullying conversations have a lower number of likes than the regular posts.
- Number of 'shared'
  Count of 'shared' episodes on social media can give an aggregated numerical view of the spread of those shares and imply the impact of the cyberbullying episodes across social networks.

(b) *Duration/Overtime*
Although each social media platform has its average content lifespan, most social content peaks their impressions within a few hours after the posts publishing. 75%

of total comments in one post episode is received at 2.5-h mark on Facebook and 6-h on Instagram. Since the cyberbullying is a hostile act and derogatory message that bully tries to impose to victim repeatedly, the period of such posts and their comments can be longer than the average online conversations.

(c) *Interarrival Time*
Bullies and aggressors tend to be more impatient compared to the spam and normal users. According to results in [17], 40% of the cyberbullying comments were generated in less than one hour after the previous comments in one cyberbullying session.

4 Cyberbullying Detection Using the Proposed Multi-dimensional Feature Set

4.1 Machine Learning Models for Cyberbullying Detection

For cyberbullying detection experiments, we constructed six machine learning models for six algorithms. We employed the Konstanz Information Miner (KNIME) [24] to establish the experimental environment to build and test the machine learning models. The models are Naïve Bayes model, Decision Tree model, Random Forest model, Tree Ensemble model, Logistic Regression model, and Support Vector Machines model. The Random Forest model of the cyberbullying detection is shown in Fig. 3. Other five cyberbullying detection models have a similar work flow as the Random Forest model. In the cyberbullying detection model, Document Creation and Preprocessing are two meta nodes for processing text type data in the work flow. The process of creation of the two meta nodes is shown in Figs. 4 and 5.

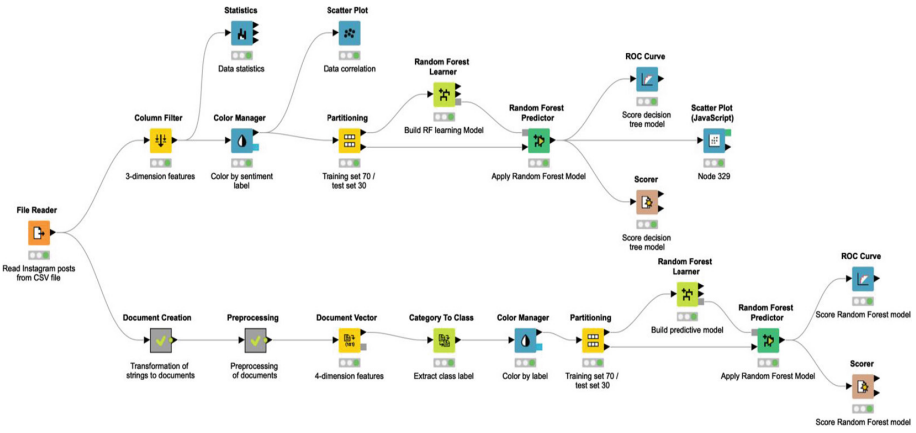

Fig. 3. Cyberbullying detection machine learning model with the Random Forest classifier

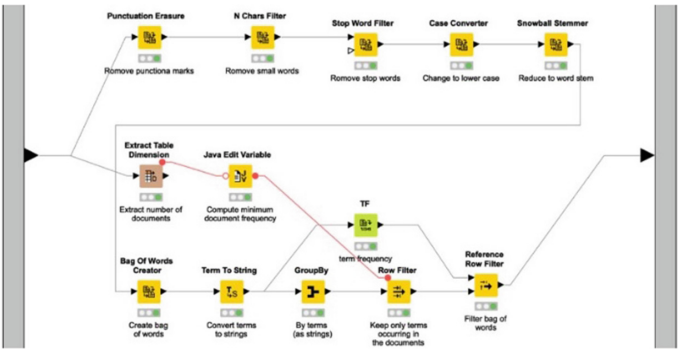

Fig. 4. The process of the Preprocessing meta node

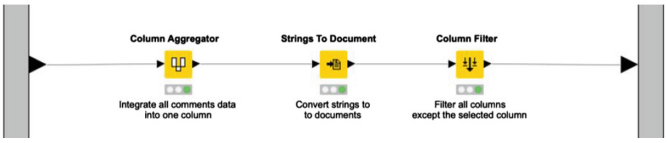

Fig. 5. The process of meta node document creation

4.2 Dataset Collection

We used nine datasets from different social media, such as Twitter, Facebook, and Formspring, available at the ChatCoder, Kaggle Dataset, University of Wisconsin-Madison, and CU CyberSafety Research Center of the University of Colorado Boulder sharing resources. The datasets used in our experiments are listed in Table 1.

Table 1. Dataset collection

Dataset	Social media platform	Format	Size	Label
Bayzick Bullying Data	Myspace	XML	17.8 MB	Yes
University of Wisconsin-Madison	Tweet (7321 tweets)	csv	53.5 MB	Yes
CU CyberSafety Research Center	Ask.fm	txt	2.94 GB	Yes
			4.56 GB	
	Instagram	csv, jpg, txt	186.9 MB	Yes
	Vine	csv, json, txt, mp4	18.06 GB	Yes
Unknown	Facebook	mat	209.8 MB	unknown
Kelly Raynolds	Formspring.me	XML, csv	15.5 MB	Yes
General Data	unspecified	XML, csv	12.7 MB	No
Text mining and cybercrime data	unspecified	txt, HTML	77.4 MB	No

4.3 Dataset Screening

The collected datasets were compared and screened based on criteria, such as whether the data had been labelled or not and whether the information in the dataset corpus included different feature set dimensions required for verification of our proposed multi-dimensional feature set based cyberbullying detection by machine learning. The experimental results described in the following sections are based on the Instagram API collection of the CU CyberSafety Research Center [17].

To be able to discern whether a user is behaving aggressively based on the contexts and scenario information, each cyberbullying episode in our dataset has the initial post and its following associated comments from other users. There are two types of users on Instagram, the ones with private profiles and the ones with public profiles. Our sample dataset comprised the information from public profiles. According to [19], there are approximately sixteen related comments following the original post in one conversation by the users other than friends on Instagram. In our study, the threshold for the lowest number of comments in each episode was set to 15. The basic requirement for episodes being chosen is that either in the posting or in related comments, the profane language or swear words [22] were found at least once in the context. The datasets then were labelled manually by five people. For an episode to be labelled as containing cyberbullying, at least 3 out of the five people had to label it as cyberbullying according to the same standards for the judgment. In the resulting sample dataset, 478 sessions are labelled as cyberbullying and 444 sessions are labelled as non-cyberbullying. The dataset is in csv format, with 922 rows and 215 columns. Each row represents one episode with 215 criteria to describe one-episode instance. In total, 59459 comments from 922 conversation episodes with different topics comprise the dataset used in our experiments.

4.4 Dataset Preparation

The procedure we used to prepare the dataset for further processing by the machine learning algorithms has the following three components.

- Criteria reduction
  Four statistical techniques, Missing Value, Low Standard Deviation, High Correlation, and Low Skewness are applied to eliminate unneeded data columns.
- Record cleaning
  Outliers, noisy and empty or sparsely populated records are removed.
- Transformation
  In this step, we transform raw data to the format that can be processed by the machine learning tool, e.g. by changing the ‘likes’ criteria format from ‘string’ to ‘number (integer)’. Each episode in the prepared dataset includes the criteria ‘episode_id’, ‘class’, ‘comments’, ‘likes’, ‘owner_id’, ‘shared_media’, ‘followed_by’ and ‘follows’.

The three-dimensional feature set for our cyberbullying detection experiments shown in Table 2 includes two features from the episode-based dimension, one feature from the individual-based dimension, and one feature from network-based dimension. The

Table 2. Three-dimensional and four-dimensional feature sets

a) 3-dimensional feature set

Feature	Dimension	Criteria	Type
Number of 'likes'	Episode-based	'likes'	number
Number of 'shared'	Episode-based	shared_media	number
Number of followers	Individual-based	followed_by	number
One of followers	Network-based	follows	number

b) 4-dimensional feature set

Feature	Dimension	Criteria	Type
Number of 'likes'	Episode-based	'likes'	number
Number of 'shared'	Episode-based	shared_media	number
Number of followers	Individual-based	followed_by	number
One of followers	Network-based	follows	number
Sentiment analysis	Content-based	comments	string

feature Sentiment analysis from the content-based dimension is added into the feature set to form the four-dimensional feature set. The features are explained in Sect. 3.

Before the three- and four-dimensional feature set data is fed into the machine learning models, certain preprocessing steps are required to process the textual data. We applied natural language processing, text mining, and information retrieval techniques to enable the cyberbullying detection model to read, process, mine and visualize textual data in KNIME. The preprocessing includes (a) cleaning of the columns without any comments; (b) integration of all the initial comments and following posts into one column for each episode; (c) removing punctuation marks; (d) filtering of small words and stop words [33]; and (e) conversion of the terms to the lower case formatting. Then the word stem is extracted using 'snowball stemming' technique to make sure the words referring to the same lexical concept reflect the same information in our cyberbullying detection models.

Word embedding requires the conversion of text to word vectors for the latent language sentiment analysis. We extract the terms, create the Bag-of-Words (BoW) data table which can be used as the input to generate document vector. After the BoW table has been created, we filter out all terms that occur in less than nine documents. We set the minimum number of documents to 9 since we assume that a term has to occur in at least 1% of all documents (9 out of 922) to represent useful information for classification. Based on these extracted words, the document vectors are numerical representations of the text and can be used for classification by a binary classifier.

4.5 Classification

For cyberbullying detection, we performed binary classification experiments using KNIME [24] with six different classifiers: Naïve Bayes, Decision Tree, Random Forests, Tree Ensemble, Logistic Regression and Support Vector Machines (SVM). The 3-dimensional and 4-dimensional feature set data corpora were independently used by all six machine learning algorithms. In supervised learning, a machine learning algorithm takes a set of training instances of which the label is known, and seeks to build a pattern that generates a desired prediction for the unseen instances. In our cyberbullying detection models, the portion of the training set and test set was set to 70 to 30, which means that of all the 922 instances (conversation episodes) in the dataset, 645 episodes were in the training set and 277 episodes in the test set.

5 Experimental Results

In this section, performance of the six machine learning algorithms is compared for the 3-dimensional and 4-dimensional feature sets. The Instagram posting dataset with 59459 comments in 922 conversation episodes used in the experiments is described in Sect. 4. Precision, Recall, F1-measure, and Accuracy performance metrics are calculated on the cyberbullying positive class. We also report Area Under Curve (AUC) scores, a performance metric that is considered to be more robust to data imbalance than Precision, Recall and F1-measure [25]. The results are shown in Table 3.

The metrics used to evaluate the performance of machine learning algorithms in cyberbullying detection using the feature sets of different dimensions are as follows.

(1) *Confusion Matrix:* The confusion matrix is a summary of prediction results of an algorithm or classifiers and provides an assessment of the selected algorithm by the values of true positive (TP), false positive (FP), false negative (FN), and true negative (TN).

(2) *Precision:* Precision is the value of instances that are genuine of a class divided by the total instances classified as that class (also called Positive Predictive Value).

$$Precision = \frac{TP}{TP + FP}$$

(3) *Recall:* Recall is the proportion value of instances classified as a given class divided by the actual total in that class (equivalent to TP rate, also called Sensitivity). Recall means what proportion of actual positives has been identified correctly.

$$Recall = TP\ Rate = \frac{TP}{TP + FN}$$

(4) *F-Measure:* F-Measure is a weighted harmonic mean of Precision and Recall.

$$F - measure = \frac{\left(\beta^2 + 1\right) \times Precision \times Recall}{\beta^2\ Precision + Recall}$$

The weight $\beta \in [0, \infty]$. $\beta = 1$ is for equal weight on Precision and Recall. This situation is referred as F1-measure. We used the F1-measure in our experiments.

(5) *Accuracy:*

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

(6) *Area Under Curve (AUC):* AUC represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one [25]. The curve is the Receiver Operating Characteristic (ROC) curve that is a function of TPR against FPR at various threshold settings. The TPR is defined as TP/((TP + FN)) and FPR is defined as FP/((FP + TN)). The AUC is scale-invariant and can measure how well predictions are ranked.

Table 3. Assessment of performance of cyberbullying detection by machine learning for the three-dimensional and four-dimensional feature sets

Algorithm	Three-dimensional feature set					Four-dimensional feature set				
	Precision	RecaU	Fl-measure	Accuracy	AUC	Precision	RecaU	Fl-measure	Accuracy	AUC
Naive Bayes	80.36	29.03	42.65	56.32	59.97	68.25	**89.58**	77.48	72.92	73.46
Decision Tree	71.21	60.65	65.51	64.26	66.08	75.33	78.47	76.87	75.45	77.50
Tree Ensemble	73.13	63.23	67.82	66.43	71.18	78.21	84.72	81.33	79.78	85.87
Random Forest	74.63	64.52	69.20	67.87	71.54	78.15	81.94	80	78.7	84.99
Logistic Regression	59.43	43.75	50.40	55.23	56.85	**86.15**	77.78	**81.75**	**81.95**	**90.39**
Support Vector Machines	74.55	26.45	39.05	53.79	58.89	82.96	77.78	80.29	80.14	87.36

The results in Table 3 are shown in their graphical forms in Figs. 6 and 7. The results demonstrate better performance of the cyberbullying detection models with the four-dimensional feature set than with the three-dimensional for all algorithms except the Precision metric results of the Naïve Bayes probabilistic classifier.

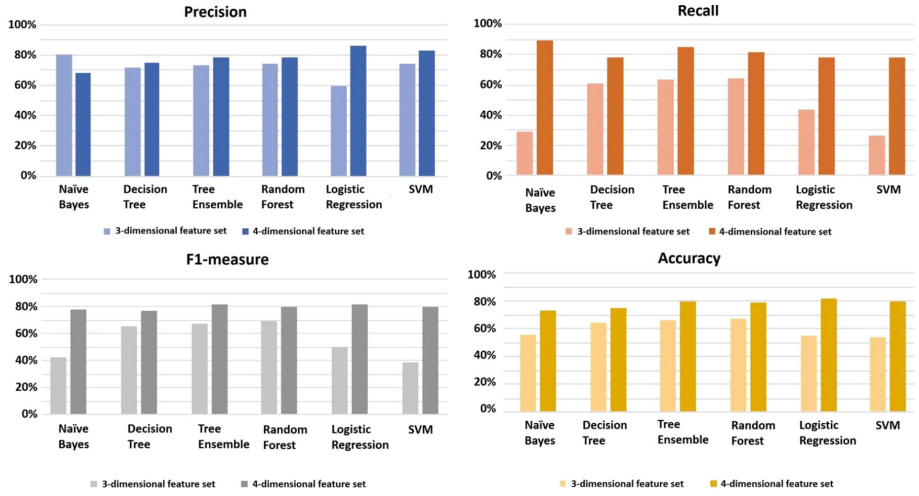

Fig. 6. Precision, Recall, F1-measure and Accuracy metric of cyberbullying detection performance using 3-dimensionional and 4-dimensional feature sets

Figure 7 shows the AUC metric of cyberbullying detection performance of the selected six machine learning algorithms. Similar to the Recall and Accuracy results, the four-dimensional feature set outperforms the three-dimensional feature set in all machine learning algorithms for the cyberbullying detection. The ROC curves for the Logistic Regression model are also shown in Fig. 7.

Although the Naïve Bayes has the highest Recall score, the Precision value makes its overall performance lower compared to the other machine learning classifiers. The Tree Ensemble and Random Forest have a very similar performance as they both belong to the ensemble classification algorithm category. In our experiments, both the Tree Ensemble and Random Forest outperformed the Decision Tree in the cyberbullying detection. Support Vector Machines performed better than Tree Ensemble and Random Forest algorithms, with the respective AUC scores of 87.36%, 85.87% and 84.99% respectively. The Logistic Regression provided the best results regarding Precision, F1-Measure, Accuracy and AUC. The Area Under Curve for the Logistic Regression in Fig. 7 illustrates the improvement of in the cyberbullying detection task with the AUC score being increased from 56.85% to 90.39%.

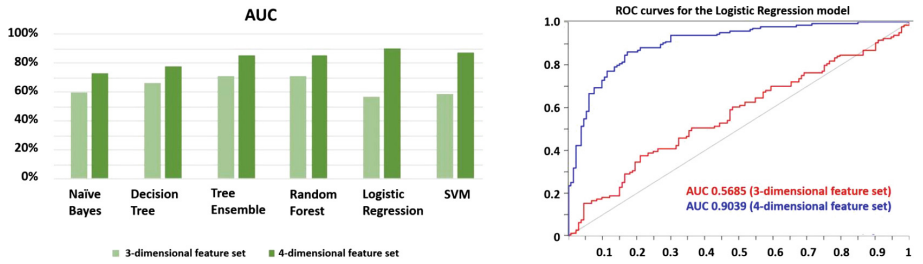

Fig. 7. AUC metric and ROC curves of the Logistic Regression for 3-dimensional (red) and 4-dimensional (blue) cyberbullying feature sets. (Color figure online)

6 Conclusions and Future Work

In this paper, we investigated the feasibility of improving cyberbullying detection on a social network by machine learning by expanding the feature set of cyberbullying behavior to dimensions with features not limited to the linguistic features of cyberbullying acts. The multi-dimensional feature set proposed in this paper expands the traditional linguistic content feature set by taking into consideration non-linguistic features of a cyberbullying behavior on a social network. In total, eighteen attributes were developed to describe and differentiate the Participants, Behavior, Technology, and Sociology traits. Under the eighteen attributes, thirty-three features were identified to facilitate a more accurate detection of cyberbullying incidents and to distinguish cyberbullying from another behavior, such as cyber harassment and cyber stalking. We applied the multi-dimensional feature set in the cyberbullying detection data pipeline built on KNIME machine learning platform. In our experiments, we tested 922 episodes with 59459 comments from Instagram. The experimental results demonstrate that cyberbullying incidents on social media platforms can be more effectively detected by using cyberbullying feature sets that are not limited to the linguistic content dimension. The improved detection of cyberbullying was achieved for all six machine learning algorithms used in our experiments - Naïve Bayes, Decision Tree, Random Forest, Tree Ensemble, Logistic Regression, Support Vector Machines - by using 5-set evaluation

metrics. Our experimental results and evaluation show that Logistic Regression and Support Vector Machines outperform the Naïve Bayes, Decision Tree, and Ensembles classification on the cyberbullying detection task.

Regarding the future research, an interesting direction for future work would be the use of advanced Deep Learning techniques, such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) algorithms, in constructing cyberbullying detection models, given that a large amount of cyberbullying related dataset is available. State-of-the-art natural language processing techniques, such as Continuous Bag-of-Words, skip-gram, N-gram, dictionary tagger, Node2vec could be integrated into the Deep Learning model for a better performance in cyberbullying detection.

References

1. Price, M., Dalgleish, J.: Cyberbullying: experiences, impacts and coping strategies as described by Australian young people. Youth Stud. Aust. **29**, 51 (2010)
2. O'Sullivan, P.B.: Reconceptualizing 'flaming' and other problematic messages. New Media Soc. **5**(1), 69–94 (2003)
3. Vandebosch, H., van Cleemput, K.: Cyberbullying among youngsters: profiles of bullies and victims. New Media Soc. **11**(8), 1349–1371 (2009)
4. Willard, N.E.: Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Research Publishers LLC, Champaign (2007)
5. Cowie, H.: Cyberbullying and its impact on young people's emotional health and well-being. Psychiatrist **37**(5), 167–170 (2013)
6. Smith, P.K., et al.: Cyberbullying: its nature and impact in secondary school pupils. J. Child Psychol. Psychiatry Allied Discip. **49**(4), 376–385 (2008)
7. Slonje, R., Smith, P.K., Frisén, A.: The nature of cyberbullying, and strategies for prevention. Comput. Hum. Behav. **29**(1), 26–32 (2013)
8. Ghasem, Z., Frommholz, I., Maple, C.: Machine learning solutions for controlling cyberbullying and cyberstalking. J. Inf. Secur. Res. **6**(2), 55–64 (2015)
9. Galán-García, P., et al.: Supervised machine learning for detection of troll profiles in twitter social network: application to real case of cyberbullying. Log. J. IGPL **24**(1), 42–53 (2015)
10. Kasture, A.S., Nand, P., Tegginmath, S.: A predictive model to detect online cyberbullying (2015)
11. Zhao, R., Zhou, A., Mao, K.: Automatic detection of cyberbullying on social networks based on bullying features. In: 17th International Conference on Computer Networks - ICDCN 2016 (2016)
12. Engman, L., Janlert, L.E., Bjorklund, H.: Automatic detection of cyberbullying on social media. In: Proceedings of 16th International Multidisciplinary Scientific Conference SGEM 2016, pp. 505–512 (2016)
13. Chatzakou, D., et al.: Mean birds: detecting aggression and bullying on Twitter (2017)
14. Haidar, B., Chamoun, M., Serhrouchni, A.: A multilingual system for cyberbullying detection: Arabic content detection using machine learning. ASTES J. **2**(6), 275–284 (2017)
15. Agrawal, S., Awekar, A.: Deep learning for detecting cyberbullying across multiple social media platforms. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 141–153. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_11

16. Van Hee, C., et al.: Automatic detection of cyberbullying in social media text. Plos One 1–21 (2018)
17. Hosseinmardi, H., Mattson, S.A., Ibn Rafiq, R., Han, R., Lv, Q., Mishra, S.: Analyzing labeled cyberbullying incidents on the instagram social network. Social Informatics. LNCS, vol. 9471, pp. 49–66. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27433-1_4
18. Hosseinmardi, H., et al.: Towards understanding cyberbullying behavior in a semi-anonymous social network. In: Proceedings of 2014 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2014, pp. 244–252 (2014)
19. Hosseinmardi, H., et al.: A comparison of common users across Instagram and Ask.fm to better understand cyberbullying. In: Proceedings of 4th IEEE International Conference on Big Data and Cloud Computing, pp. 355–362 (2014)
20. Hosseinmardi, H.: Dataset - CU Cyber Safety Research Center, Univ. Colorado at Boulder. https://sites.google.com/site/cucybersafety/home/cyberbullying-detection-project/dataset
21. Pieschl, S., et al.: Relevant dimensions of cyberbullying - results from two experimental studies. J. Appl. Dev. Psychol. **34**(5), 241–252 (2013)
22. NoSwearing.com: Swear word list, dictionary, filter, and API. https://www.noswearing.com
23. Hatebase. https://www.hatebase.org/
24. Berthold, M.R., et al.: KNIME-the Konstanz information miner: ver. 2.0 and beyond. ACM SIGKDD Explor. Newsl. **11**(1), 26–31 (2009)
25. Fawcett, T.: An introduction to ROC analysis. Pattern Rec. Lett. **27**(8), 861–874 (2006)
26. Textfixer.com English stop words list. https://www.textfixer.com/tutorials/common-english-words.txt