2017

Presenting a Labelled Dataset for Real-Time Detection of Abusive User Posts

Hao Chen
Technological University Dublin

Susan McKeever
Technological University Dublin, susan.mckeever@tudublin.ie

Sarah Jane Delany
Technological University Dublin, sarahjane.delany@tudublin.ie

Recommended Citation

Chen, H., Mckeever, S. & Delany, S.J. (2017). Presenting a labelled dataset for real-time detection of abusive user posts. *WI'17:Proceedings of the International Conference on Web Intelligence*, Leipzig, Germany, August 23-26. doi:10.1145/3106426.3106456

Presenting a Labelled Dataset for Real-Time Detection of Abusive User Posts

Hao Chen
Applied Intelligence Research Centre
Dublin Institute of Technology
Dublin, Ireland
hao.chen@mydit.ie

Susan Mckeever
Applied Intelligence Research Centre
Dublin Institute of Technology
Dublin, Ireland
susan.mckeever@dit.ie

Sarah Jane Delany
Applied Intelligence Research Centre
Dublin Institute of Technology
Dublin, Ireland
sarahjane.delany@dit.ie

ABSTRACT

Social media sites facilitate users in posting their own personal comments online. Most support free format user posting, with close to real-time publishing speeds. However, online posts generated by a public user audience carry the risk of containing inappropriate, potentially abusive content. To detect such content, the straightforward approach is to filter against blacklists of profane terms. However, this lexicon filtering approach is prone to problems around word variations and lack of context. Although recent methods inspired by machine learning have boosted detection accuracies, the lack of gold standard labelled datasets limits the development of this approach. In this work, we present a dataset of user comments, using crowdsourcing for labelling. Since abusive content can be ambiguous and subjective to the individual reader, we propose an aggregated mechanism for assessing different opinions from different labellers. In addition, instead of the typical binary categories of abusive or not, we introduce a third class of 'undecided' to capture the real life scenario of instances that are neither blatantly abusive nor clearly harmless. We have performed preliminary experiments on this dataset using best practice techniques in text classification. Finally, we have evaluated the detection performance of various feature groups, namely syntactic, semantic and context-based features. Results show these features can increase our classifier performance by 18% in detection of abusive content.

KEYWORDS

Abusive Detection; Machine Learning; Labelling Strategy; Feature Selection

ACM Reference format:
Hao Chen, Susan Mckeever, and Sarah Jane Delany. 2017. Presenting a Labelled Dataset for Real-Time Detection of Abusive User Posts. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017,* 7 pages.
DOI: 10.1145/3106426.3106456

1 INTRODUCTION

The volume of user generated content (UGC) has increased rapidly with the growing usage of social media websites, resulting in a need to moderate inappropriate content. This content includes insulting or hurtful language that is offensive to either an individual, or a group of people sharing the same characteristics (e.g. a particular race, religion, nationality etc.). For the purposes of this work, we consider abusive content as content that has negative consequences, through the apparent deliberate targeting of an individual or group. Companies have a responsibility to ensure that the content on their websites is appropriate. We identify that the common moderation approaches used in most websites can be divided into two categories: pre-publication and post-publication. In pre-publication moderation, each comment is checked prior to publication (e.g. BBC online news); In post-publication moderation, comments are freely published, with the identification of abuse after publishing via user reports and/or site moderators determining that a comment has violated community rules (e.g. Yahoo News and Reddit). In general, abusive content forms a small proportion of total content in online environments (e.g. 0.2% in tweets, [22]). Detecting all abusive content pre-publication, akin to finding a needle in a haystack, may take an impractical amount of time and effort if done manually. Moreover, if abusive content is available to the public until a user reports it or community moderator finds it, as with post-publication moderation, the negative consequences may have already happened. These scenarios present cost versus risk trade-offs. Therefore, the development of real-time automatic systems to aid moderators in reviewing user-generated content is a priority. Determining whether content is abusive is a subjective exercise, so we are not attempting to design a tool to completely replace manual inspection by moderators. Rather our aim is to provide some level of automatic detection and risk flagging in order to improve the overall efficiency of abusive content moderation.

Given a lack of gold standard datasets in abusive detection [20], we have collected our own dataset from a news commenting website, and approached the labelling process using a crowdsourcing service. Labelling abusive comment is not an objective exercise: a comment can receive different judgements by different labellers. Therefore, we have used a specific consensus labelling methodology to aggregate multiple labellers' views, to counteract subjectivity. We required labellers to categorise our data into three classes (abusive, non-abusive, and undecided) and then to identify the severity of abusive posts on a scale. Next, a set of text classification techniques based on our previous work [3] was used to generate a benchmark for this dataset on detection accuracy. The techniques include feature representation, feature reduction and dataset balancing. In addition, we proposed two groups of extracted features, textual and context-based, to boost detection performance. Finally, we explored the prediction of severity of abuse for abusive comments.

WI '17, Leipzig, Germany

DOI: 10.1145/3106426.3106456

The rest of paper is organised as follows: The literature review is discussed in Section 2. In Section 3 we explain the intrinsic characteristics of our dataset, and our labelling methods. Feature representation is presented in Section 4, followed by describing the experimental methodology in Section 5. Our experimental work and results are presented in Section 6, with conclusions and future work in Section 7.

2 RELATED WORK

This paper has two purposes: (1) to present a labelled dataset for the detection of abuse in user generated comments and (2) to apply best practice text classification techniques and additional feature engineering on our dataset to maximise detection accuracies. In this section, we will focus the literature review on two research areas: labelling strategies and detection techniques.

2.1 Labelling Strategies

Although analyzing abusive text content on social media sites has emerged in the machine learning field as a research area, the lack of reliable labelled datasets presents a barrier. As one of the earliest public repositories in this domain, the labelled dataset for the CAW2.0 workshop (organised by Fundacian Barcelona Media) was popular with many researchers [5, 11, 23]. Subsequent work has involved the manual collection and labelling of data. A simple labelling mechanism is to use colleagues to complete annotations. Dadvar et al. [7] asked two graduate students to review YouTube comments; Nobata et al. [15] used trained employees in Yahoo! to create their corpus from the Yahoo! News forum; Bayzick et al. [1] used three undergraduates for labelling of a Myspace dataset; In [9] Dinakar et al. were assisted by two educators from middle school to manually label abusive posts from YouTube.

Rather than using internal colleagues, the use of crowdsourcing services to label data has becoming an increasingly popular approach in the machine learning labelling field [2, 13, 20]. Crowdsourcing services enable individuals or organisations to distribute their labelling tasks across a variety of online users. Two crowdsourcing platforms are widely used, namely Amazon Mechanical Turk (AMT)[1] and CrowdFlower[2]. Kontostathis et al. [13] and Sood et. al [20] used AMT to label their dataset for cyberbullying and insulting content respectively. In both cases, three labellers were used to review a single post and majority votes applied to decide the label; CrowdFlower was used by Burnap et al. [2] to label hate speech in a Twitter dataset.

2.2 Detection Techniques

The most basic approach to detect abusive text content is lexicon-based filtering. Reynolds et al. [17] built a detection model using a weighted profane words list. Sood et al. [19] applied Levenshtein Distance to alleviate against word abbreviations and misspelling, gaining an improvement in detection compared to static lexicon matching. Zhao et al. [24] introduced word2vec to dynamically expand the words vocabulary of insulting terms.

In addition to lexicon-based approaches, researchers in this area have applied supervised machine learning techniques for abusive content detection. Several have focused on feature engineering. Chen et al. [4] devised a Lexical Syntactic Feature (LSF) architecture that contains lexical features and syntactic features, which outperformed the traditional textual representations of Bag of Words (BoW) and N-grams. Dadvar et al. [6, 7] extracted user profile information such as age and gender to increase classifier performance. Well-known classifier algorithms including Naive Bayes [4], Decision Trees [17], Logistic Regression [21] and Support Vector Machines (SVMs) [2] are widely used in this domain. Additionally, Mangaonkar et al. [14] presented an ensemble classifier that predicts abuse by combining a variety of classifiers' results to avoid over-fitting. Instead of using classification, Bosque et al. [8] addressed abusive detection using a linear regression model which involved predicting an abusive score to each comment. Their results show that regression models are a solid candidate for scoring the abusive degree of the tweets.

3 DATASET

This section outlines the source and collection methods used for our corpus of user generated comments and describes how the actual dataset was extracted from this corpus and the labelling approach used.

3.1 Data Collection

For our data collection, we sought an online service which facilitates free, unhindered posts from a public base of users. With this requirement in mind, we collected data from a general news website that provides an open forum for discussion and debate, anchored by news stories. News stories are editorially divided into at least one of the following eight categories: Local, Politics, International, Opinion, Family, Culture, Technology and Business. News stories can be tagged into multiple categories. For example, a news story with title *'Digital can deliver vital votes - but how do politicians earn the 'like' love?'* can be found in both the 'Opinion' and 'Politics' categories.

The site claims that no moderation techniques are carried out before and after comments are published, which in theory gives us an opportunity to collect original comments without prior editing. However, from our own test posts and observation, the site is using a profane words filter for some level of pre-publication moderation. There is also a user option to flag abusive comments, as per post-publication moderation. Users of the site can only post comments if they have linked to their social media account, either Facebook or Twitter. This reduction in anonymity is likely to lower the risks of abusive posts [10].

In our corpus, we gathered user-generated comments from 3,765 news articles between August 2015 and September 2016. The distribution of comments in each news category is shown in Table 1. As an individual news story may be tagged in multiple categories, the sum of percentages is higher than 100%. We noted that the highest proportion of user comments originated from 'Politics' and 'Opinion', accounting for 28% and 16% respectively. In addition, the higher ratio of comments to corresponding news articles in these two categories indicate that 'Opinion' and 'Politics' as topics generate more user postings than the news articles in other categories.

[1] https://www.mturk.com/mturk/
[2] https://www.crowdflower.com

Table 1: Distribution of news articles and comments for each category in our corpus. The category name is abbreviated to the first three letters.

Total	Bus	Cul	Int	Loc	Fam	Pol	Tec	Opi
News	12%	12%	16%	15%	14%	20%	13%	11%
Comments	12%	10%	14%	13%	12%	28%	7%	16%

As the website displays news in order of publishing time, the latest news appears on the front page of the website. Therefore, recent news stories are highly visible to users and gain comments rapidly. As time goes by, new comments diminish as the news is no longer getting attention from users. We chose to collect comments when news articles have been published for a long time (more than 10 hours) to make sure that user reading of and commenting on that article has reached a saturation level.

In addition to the actual comment text, we collected other meta-data: (1) the number of 'likes' and 'dislikes' for the comment, (2) the username of the person who posted the comment, (3) whether the user is linked to Facebook or to Twitter (4) whether the comment is a reply to the previous comment or a direct response to a news story.

3.2 Analysing User Commenting Behaviour

The average length of comment in our corpus is 37 words. As shown in Figure 1, the majority of comments are under 50 words in length, indicating that users prefer to post their views using brief messages. However, the comments themselves vary greatly as the standard deviation of comments' length is approximately 51 words.

To show the wide spread of users of this website, we then analyzed the distribution of comments per user. After trimming some outliers, Figure 2 shows how frequently users use this site to post comments. Since almost half of users in our corpus posted just a single comment within our dataset, there is a wide spread of users captured in the corpus. In addition, 64% of comments are responses to other users' views as opposed to a direct news article comment, which suggests that users like discussion and debate. Furthermore, a narrow majority of users link their account to Twitter (57%) rather than Facebook.

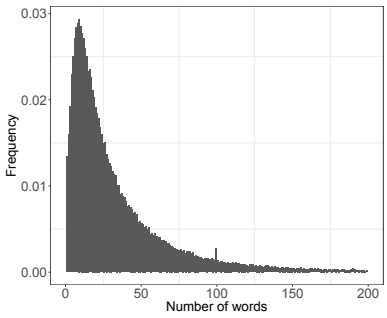

Figure 1: Distribution of the number of words per comment

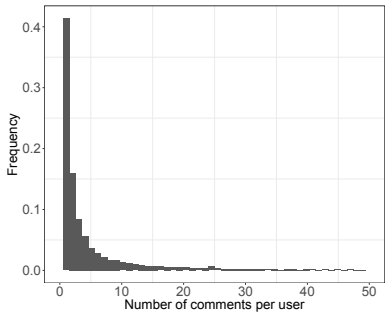

Figure 2: Distribution of the number of comments per user

3.3 Sampling Data

Due to the size of our corpus, manual labelling of all comments was impractical. We needed to select a sample for labelling, and to ensure that this sample contained an adequate level of abusive comments to make our dataset useful for model training. Rather than randomly choose data from the corpus where the majority are not abusive, we used lexicon-based filtering to increase the likelihood of abusive comments in our sample with the assumption that profane words are indicative of abusive content. We used three public profane word lists, namely, CMU[3], Github[4], and Noswearing[5]. They contain 1383, 376 and 349 abusive words respectively. We observed that, using CMU, the largest words list, nearly 24% of comments in corpus include at least one profane word. However, using Github and Noswearing, as smaller word lists, filtered just 2% of comments from corpus each. We wished to avoid over restriction of our sample on specific words, so we used the CMU list to complete the pre-filtering because it contains the widest list of potentially profane words.

Prior to sampling, we first eliminated approximately 10% of comments in the corpus based on their length. This was suggested by Sood et al.[20] in order to eliminate the comments that are either too short to meaningfully interpret or too long to digest quickly. We then filtered the remaining data using the CMU list. The distribution of comments according to the number of profane words they contain is shown in Figure 3. 76% of comments do not have any profane word, with decreasing percentages as the number of profane word matches increases, up to just 0.56% comments containing at least 4 profane words. We then sampled our dataset for labelling by randomly selecting 400 comments from each of these groupings of comments. In total, there are 2000 comments across 5 groups.

[3]https://www.cs.cmu.edu/ biglou/resources/bad-words.txt
[4]https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en
[5]http://www.noswearing.com/dictionary

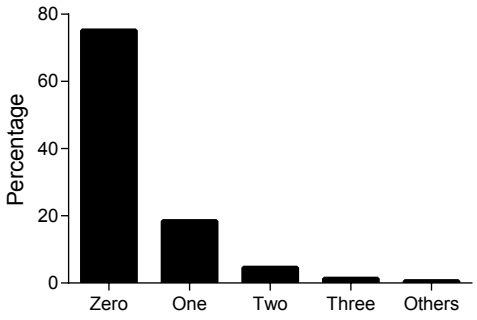

Figure 3: Distribution of the number of profane words

3.4 CrowdSource Labelling

Obtaining reliable labels plays an essential role in our research. In this paper, we used CrowdFlower[6] to conduct our labelling process. Every comment was individually displayed to the labeller with an explanation of what we considered abusive content (see Section 1). Each labeller was asked *'Is the comment abusive or not?'*. Instead of using binary labels - 'yes' or 'no' for abusive or non-abusive, we treated this labelling task as a ternary problem with the inclusion of a third label choice - 'undecided'. The decision to introduce an 'undecided' category was to reflect the fact that the assessment of a comment as abusive or not can be ambiguous, and labellers can disagree. We suggest that flagging a post as 'undecided' can be considered as flagging risk. For comments that the labeller considered abusive, a second question captured the severity of the abuse. We required labellers to apply a scale to the level of abuse from 1 to 4 in which 1 is very slightly abusive, 2 is slightly abusive, 3 is strongly abusive and 4 is very strongly abusive. Each comment was labelled by six labellers and not all labellers labelled every comment.

When we launched our labelling task in CrowdFlower, the quality control rules of this platform had to be observed with regard to labellers' consistency. However, we recognised that quality control using pre-defined test questions would affect the labellers' subjective views. Therefore, to minimise the potential bias of the test questions, we carefully selected twenty comments that were obviously distinguishable. These test questions were randomly interspersed in the real labelling tasks. In addition, we restricted our labellers to those located in the same region as the news website, to allow for knowledge of local language expressions. A minimum judgement time per comment was set at 3 seconds.

3.5 Labelling Results

In total, 89 participants worked on our labelling task. 9 were eliminated due to their low labelling accuracy, which is defined as less than 60% accuracy on our pre-defined testing questions. As each comment was labelled by six different labellers, it was not appropriate to carry out a standard measure (such as Fleiss Kappa) for examining agreement. We defined the following rules in order to determine the decision of the final label for each comment:

- If the comment receives at least 3 votes for 'undecided', or the comment has the same votes for 'abusive' and 'non-abusive', the final label is 'undecided';
- Otherwise:
  - If the comment receives at least 3 votes for 'abusive', the final label is 'abusive';
  - If the comment receives at least 3 votes for 'non-abusive', the final label is 'non-abusive';

The labelling outcomes on the dataset is presented in Table 2 and some examples are provided in Table 3. The majority of comments are tagged as 'non-abusive', while 15% of comments are labelled as 'abusive', and 'undecided' comments is at just 6%. However, the proportion of unanimous (all labellers agree) labels accounts for less than 40% in all labelled data, indicating the subjectivity of manual labelling in this domain. Furthermore, non-abusive comments have a higher proportion of unanimous labels (41%) than abusive comments (27%). So, subjectivity is more evident in the abusive content.

4 FEATURE SPACE

Supervised machine learning requires a representation for the text content and an associated label. The most common representation is the vector-space model [18] where the text content is represented as a vector of features characteristic of the content and the feature value represents the frequency of occurrence of the feature in the text. Bag of Words (BoW), where the features are the words, is a common representation for text. However, BoW disregards word order which can contain syntactic and semantic information, so in addition to content features we include syntactic, semantic, and context-based features.

Table 2: Distribution of labelling results

	Unanimous	Not Unanimous	Total
Abusive	4%	11%	15%
Non-Abusive	32%	47%	79%
Undecided	0%	6%	6%
Total	36%	64%	100%

Table 3: Examples of labelled data, Y/N/U stand for abusive, non-abusive and undecided respectively, the number is in front of Y/N/U is the number of votes

Comment	Labels	Final Label
Stop shaming obese people ..the fat b*ds will probably gang up on you!	6Y	Abusive
Dot, it's there in black and white for you. Are you blind or unable to read?	5Y, 1N	Abusive
You haven't heard of Gary's Mattress then, hes the king of crazy mattress sales man. https://www.youtube.com/watch?v=MArsmQ9tGN0	6N	Non-Abusive
@XXX, you're in quite the mood today, aren't you? But nevertheless you and David are absolutely correct, it's past time that this stupid ban is lifted	5N, 1U	Non-Abusive
Anything touched by Kelly should be abolished!	3Y, 3N	Undecided
If they were Muslim it would be called a terrorist attack and they'd be dead by now	2Y, 2N, 2U	Undecided

[6]https://www.crowdflower.com

4.1 Content-Based Features

To represent the content we used n-grams which captures some syntactic and semantic information by splitting text as subsequences of n continuous terms. Text pre-processing was performed on the content prior to tokenisation into n-grams. All text was changed to lowercase, links were replaced by the generic term 'url_links', and mentions (user names preceded by the '@' symbol) were replaced by the anonymous term '@username'. Given that comments in the dataset are typically conversational in style and short, we did not apply stemming or stop word removal. Following results from our previous work [3], Document Frequency (DF) feature reduction was used to cut down on high or low frequency occurrence terms without jeopardising model performance. Term Frequency (TF) was used to normalise the feature values.

4.2 Textual Features

To produce a more sophisticated feature representation and to boost model detection performance, we also compiled some additional features which we term 'textual features', derived from the text of the comment itself. These features include both syntactic and semantic features:

- *Syntactic Features* - The set consists of basic syntactic information (e.g. number of words for each comment, number of characters for each comment) and advanced syntactic information where we can extract personal writing style such as 'average word length' and 'average sentence length'.
- *Semantic Feature* - Table 4 shows features which were extracted for their semantic information, which we consider as potentially valuable indicators of abuse. For example, amount of punctuation or uppercase characters can be used to indicate users' emotion. (e.g. 'You are a loser' can be considered less strong than 'You are a LOSER!').

Table 4: Feature Representation

Content-Based Features (Ngrams)	
Textual Features	
	Syntactic
A1	Number of words in the comment
A2	Number of characters in the comment
A3	Number of sentences in the comment
A4	Average word length (#characters divided by #words)
A5	Average sentence length (#words divided by #sentences)
	Semantic
B1	Profane words usage level (#profane words divided by #words)
B2	Personal pronouns usage level (#personal pronouns divided by #words)
B3	Uppercase letter usage (#uppercase letters divided by #sentences)
B4	Punctuations usage (#punctuations divided by #sentences)
B5	URL usage level (#URL divided by #words)
B6	@Username usage level (#@Username divided by #words)
Context-Based Features	
C1	Whether the comment is a reply to a previous comment or a new comment
C2	Comment's news categories
C3	User identified by Facebook versus Twitter
C4	#Comments up to this point.

4.3 Context-Based Features

This set of features aims at exploiting context that is information external to the content in the post itself. We identified 4 types of context features (Table 4): (1) C1 which captures whether the commenter is responding to the news story or responding to other comment contains information that probably indicates the abuse target (the news author or other news readers). 2) C2 captures the profane words usage across the different categories. 3) C3 captures whether the user is identified by a Twitter or Facebook login. Since Twitter allows anonymity, this may be linked more to abusive posts. And 4) C4 is the number of comments the article has received at the time of collection. The controversial news stories may be likely to have more comments.

There are other useful information sources which can be considered as features. For example, the website provides a function that allows users to express sentiment or opinion by clicking 'thumbs up' or 'thumbs down'. However, this information is only available after the comment has been available to users and so requires a long waiting time before it can be collected. As our aim is to detect abusive comment in a real-time model, we do not use this information even though we have gathered it in our corpus.

4.4 Feature Representation

The above features are concatenated to form the final feature representation. We applied normalisation technique (MinMax [16]) to transform these values to the same range level. Once a final feature representation for each comment has been obtained, we used Singular Value Decomposition (SVD) [3] to reduce the feature space before feeding them into the supervised machine learning algorithm.

5 METHODOLOGY

We use a Support Vector Machine (SVM) with a linear kernel, as one of the most efficient classification algorithms for text classification, as our classification algorithm [12].

In addition, our dataset is unbalanced, with the 'abusive' comments and 'undecided' comments occurring far less than 'non-abusive' comments (see Table 2). Therefore, we applied resampling [3] to randomly oversample the minority instances before training the model. To minimise the impact of conducting random oversampling, we resampled the training data 3 times, and used the average performance across the 3 result sets. Prior to resampling, stratified 10-fold cross validation was used where 9 folds are used for training and the remaining fold used for testing. Resampling was applied to training folds.

5.1 Performance Measure

The results of the experiments are reported using a standard text classification measure *Recall* which is the ability of the classifier to find all instances of a specific class. It is calculated as the proportion of instances for a class that are correctly identified. We assume that the consequence of failing to detect abusive content (False Negative) is arguably higher than a non-abusive comment incorrectly identified as abusive (False Positive). Therefore, we focus in particular on the recall for the abusive class, as shown in Equation 1 (where TruePositive is the proportion of abusive comments that were predicted as abusive and FalseNegative is the proportion of abusive comments that were predicted as non-abusive). Non-abusive recall and average recall (across all classes) are also reported.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (1)$$

6 EXPERIMENTS & RESULTS

We ran a set of experiments to establish a baseline, followed by a set of experiments using feature-based tactics to improve our detection performance: First, we applied basic lexicon-based filtering to detect abusive comments. We then used best practice text mining techniques to build a supervised machine learning model. In subsequent experiments, the performance of our model was boosted by textual features and context-based features. As part of this, we compared the difference between binary classification and multiple classification using a third class 'undecided', aiming to increase our ability to highlight potentially risky content. Finally, we experimented on classifying abusive severity.

6.1 Experiment 1 - Lexicon-Based

Three previously mentioned profane word lists, CMU, Github and Noswearing, were compared in this experiment. The principle is straightforward: if a comment contains any word from the profane word list, it is flagged as abusive. Otherwise, it is deemed non-abusive. As this is a binary classification, 'undecided' instances were removed for this experiment. *Recall* performance is shown in Table 5. *Abusive recall* in CMU is high at 0.91; this is due to the fact that our original dataset sample was selected by filtering words based on CMU (see Section. 3.3), which means 1600 comments in our sample (2000) include at least one profane word. However, the low *Non-abusive recall* performance of 0.23 indicates that using a large profane word list results in False Negatives, where a majority of comments including profane words are actually not abusive. Whilst decreasing the size of the profane word list can alleviate this issue, it adversely affects abusive comment detection. *Abusive recall* slumped to 0.2 after selecting smaller lists (Github or Noswearing) and down to 0.08 when we used overlapping words (i.e. contained in all three lists).

Table 5: Lexicon-based recall, size is the number of words in list.

Dictionary	Size	Abusive	Non-Abusive	Average
CMU	1383	0.91	0.23	0.57
Github	376	0.2	0.92	0.56
Noswearing	349	0.2	0.94	0.57
Overlap	56	0.08	0.99	0.54

6.2 Experiment 2 - Abuse Classification

We next conducted experiments using a supervised machine learning approach. In order to compare with the results from the lexicon comparison, we first used a binary classification model by temporarily moving 'undecided' instances into the 'abusive' class. We assumed that 'undecided' comments may contain potentially harmful content and have a potential risk of abuse. The following text mining techniques were used: character level n-grams, ranging from 2 to 4, document frequency reduction is at 1% threshold, where the most and least frequent 1% of terms are excluded (reducing the feature size by 81%), and SVD to reduce the feature space to 1000 (an optimal value based on a series of preliminary experiments). The results are shown in Table 6. Our classifier baseline result used only content-based features (i.e. n-grams), had an *abusive recall* performance of 0.34. It shows that the supervised machine learning approach is more effective than the previous lexicon approach. In addition, the two extra types of features (textual and context-based) can enhance classifier performance by a combined total 18% on abusive detection even though the dimension of these two features is considerably smaller than that of n-grams.

Table 6: Classification results (recall) with different feature sets

	Abusive	Non-Abusive	Average
Ngrams	0.34	0.89	0.62
Ngrams+Textual	0.39	0.88	0.64
Ngrams+Context	0.37	0.88	0.62
Ngrams+Both	0.4	0.88	0.64

Using the previous techniques, we also carried out experiments on multi-class classification where we include the instances that had been labelled as 'undecided', to supplement our positive and negative instances. It is not straightforward to compare ternary classification results to binary classification results. From our perspective, the issue of undetected abusive comments is more serious than the issue of neutral comments mis-labelled as abusive. We therefore focus on the *recall* of those classes containing risky instances. Since undecided comments have the potential to be abusive comments, risky instances in ternary classification is interpreted as the number of abusive and undecided instances that were predicted as either abusive or undecided. Accordingly, *risk recall* is the ratio of the number of *risk* instances to the total number of true labelled instances of abusive and undecided. As shown in the confusion matrix in Table 7, $risk\ recall = (132+32+19+9)/(311+113) \approx 0.45$, which is higher than the *risk recall* in binary classification (0.4), a gain of 12.5%. Consequently, our results indicate that in this case it is useful to keep the undecided category and treat abusive detection as a multiple classification, rather than just a binary classification problem.

Table 7: Confusion matrix of multiple classification

		Predict			
		Abusive	**Non-Abusive**	**Undecided**	**Total**
Label	**Abusive**	132 (6.6%)	147 (7.4%)	32 (1.6%)	311
	Non-Abusive	150 (7.5%)	1351 (67.6%)	75 (3.8%)	1576
	Undecided	19 (1%)	85 (4.3%)	9 (0.5%)	113
	Total	301	1583	116	2000

6.3 Experiment 3 - Severity Classification

In our labelling process (Section 3.4), after a comment was labelled as abusive, labellers were required to identify the level of severity of abuse on a scale. In this experiment, we tried to predict the severity level of the abuse a model built in a similar way (including resampling) as before using the abusive comments as input. We considered abusive comments with high severity as likely to have

a more negative impact in the online environment than abusive comments with low severity. Therefore, the aim of this experiment was to help moderators to prioritise when facing abusive comments. We simply grouped the severity degree into *high* and *low* classes. According to our experiments, the performance of classifying severity levels is not outstanding, achieving just 0.2 in *high recall* and 0.89 in *low recall*. We attribute these rudimentary results to a lack of data, with just 311 abusive comments to use.

7 CONCLUSION & FUTURE WORK

The purposes of this paper were to present a reliable labelled dataset in this area, and to explore automatic abusive detection using supervised machine learning techniques. We highlight the following from our work: Firstly, to address the issue of subjectivity in abusive detection, we used multiple labellers to label our dataset and defined a robust strategy for consensus and labelling. Secondly, we created a baseline for this labelled dataset using supervised machine learning techniques: N-grams to represent text, document frequency to reduce feature size, oversampling to balance the training data, SVD to further reduce feature dimensions and SVM for classification. Thirdly, we extracted two domain specific feature sets: textual and context-based. These boosted our model performance by 15% and 9% respectively. We gained a further 12.5% improvement when exploiting this as a multi-class problem and treating 'undecided' content as potentially abusive. Finally, we attempted to classify the abusive severity in 'high' and 'low' categories. However, the results show limited performance, hampered by a lack of abusive data.

In future work, we will focus our research in two main directions. Firstly, with the importance of quality labelled data for building detection models, we will continue to focus on the data labelling process by exploiting human-in-the-loop learning techniques such as active learning. Secondly, we noticed that most work to date in this area is based on feature engineering. The feature representation for the text content is usually extracted manually. Given the power of automatic feature selection by using deep learning, as well as designing domain-specific features, we would like to explore deep learning for feature representation to improve the abusive detection accuracies.

REFERENCES

[1] Jennifer Bayzick, April Kontostathis, and Lynne Edwards. 2011. Detecting the presence of cyberbullying using computer software. *WebSci Conferemce* (2011).

[2] Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7, 2 (2015), 223–242.

[3] Hao Chen, Susan Mckeever, and Sarah Jane Delany. 2016. Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media. In *Advances in Computational Intelligence Systems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, 2016*, Vol. 513. Springer, Springer, Lancaster, UK, 187.

[4] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust (SOCIALCOM-PASSAT '12)*. IEEE Computer Society, Washington, DC, USA, 71–80. DOI: https://doi.org/10.1109/SocialCom-PASSAT.2012.55

[5] Maral Dadvar and Franciska de Jong. 2012. Cyberbullying Detection: A Step Toward a Safer Internet Yard. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion)*. ACM, New York, NY, USA, 121–126. DOI: https://doi.org/10.1145/2187980.2187995

[6] M Dadvar, FMG de Jong, RJF Ordelman, and RB Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent, St. Pietersnieuwstraat 33, 9000 Gent, Belgium, 23–26.

[7] Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. 2014. Experts and machines against bullies: a hybrid approach to detect cyberbullies. In *Canadian Conference on Artificial Intelligence*. Springer, 275–281.

[8] Laura P Del Bosque and Sara Elena Garza. 2014. Aggressive text detection for cyberbullying. In *Mexican International Conference on Artificial Intelligence*. Springer, 221–232.

[9] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the Detection of Textual Cyberbullying. In *The Social Mobile Web, Papers from the 2011 ICWSM Workshop (AAAI Workshops)*. AAAI, Barcelona, Catalonia, Spain. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3841

[10] Homa Hosseinmardi, Rahat Ibn Rafiq, Shaosong Li, Zhili Yang, Richard Han, Shivakant Mishra, and Qin Lv. 2014. A comparison of common users across instagram and ask. fm to better understand cyberbullying. *arXiv preprint arXiv:1408.4882* (2014).

[11] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber Bullying Detection Using Social and Textual Analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia (SAM '14)*. ACM, New York, NY, USA, 3–6. DOI: https://doi.org/10.1145/2661126.2661133

[12] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98* (1998), 137–142.

[13] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th annual acm web science conference*. ACM, 195–204.

[14] A. Mangaonkar, A. Hayrapetian, and R. Raje. 2015. Collaborative detection of cyberbullying behavior in Twitter data. In *2015 IEEE International Conference on Electro/Information Technology (EIT)*. IEEE, Northern Illinois University Dekalb, IL, USA, 611–616. DOI: https://doi.org/10.1109/EIT.2015.7293405

[15] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 145–153.

[16] S Patro and Kishore Kumar Sahu. 2015. Normalization: A Preprocessing Stage. *arXiv preprint arXiv:1503.06462* (2015).

[17] Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*, Vol. 2. IEEE, IEEE, Hilton Hawaiian Village, Honolulu Hawaii USA, 241–244.

[18] Fabrizio Sebastiani. 2002. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* 34, 1 (March 2002), 1–47. DOI: https://doi.org/10.1145/505282.505283

[19] Sara Sood, Judd Antin, and Elizabeth Churchill. 2012. Profanity Use in Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1481–1490. DOI: https://doi.org/10.1145/2207676.2208610

[20] Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology* 63, 2 (2012), 270–285.

[21] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 1980–1984. DOI: https://doi.org/10.1145/2396761.2398556

[22] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, 656–666.

[23] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2 (2009), 1–7.

[24] Rui Zhao, Anna Zhou, and Kezhi Mao. 2016. Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features. In *Proceedings of the 17th International Conference on Distributed Computing and Networking (ICDCN '16)*. ACM, New York, NY, USA, Article 43, 6 pages. DOI: https://doi.org/10.1145/2833312.2849567