2016

Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media

Hao Chen
Technological University Dublin, hao.chen@mydit.ie

Susan McKeever
Dublin Institute of Technology, susan.mckeever@dit.ie

Sarah Jane Delany
Dublin Institute of Technology

Recommended Citation

Chen, H., McKeever, S. & Delany (2016) Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media, *16th UKCI* September 2016, Lancaster University. doi:10.1007/978-3-319-46562-3_12

Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media

Hao Chen, Susan Mckeever, Sarah Jane Delany

Abstract The issues of cyberbullying and online harassment have gained considerable coverage in the last number of years. Social media providers need to be able to detect abusive content both accurately and efficiently in order to protect their users. Our aim is to investigate the application of core text mining techniques for the automatic detection of abusive content across a range of social media sources include blogs, forums, media-sharing, Q&A and chat - using datasets from Twitter, YouTube, MySpace, Kongregate, Formspring and Slashdot. Using supervised machine learning, we compare alternative text representations and dimension reduction approaches, including feature selection and feature enhancement, demonstrating the impact of these techniques on detection accuracies. In addition, we investigate the need for sampling on imbalanced datasets. Our conclusions are: (1) Dataset balancing boosts accuracies significantly for social media abusive content detection; (2) Feature reduction, important for large feature sets that are typical of social media datasets, improves efficiency whilst maintaining detection accuracies; (3) The use of generic structural features common across all our datasets proved to be of limited use in the automatic detection of abusive content. Our findings can support practitioners in selecting appropriate text mining strategies in this area.

1 Introduction

Bullying [22] is defined as repeated exposure to negative actions on the part of others. The emergence of bullying in the online world has been termed *cyberbully-*

Hao Chen
Dublin Institute of Technology, e-mail: hao.chen@mydit.ie

Susan Mckeever
Dublin Institute of Technology e-mail: susan.mckeever@dit.ie

Sarah Jane Delany
Dublin Institute of Technology e-mail: sarahjane.delany@dit.ie

ing, where bullying is carried out using electronic forms of contact repeatedly and over time [27]. Social media sites allow users to enter comments, post messages and publish micro-blogs, opening avenues for cyberbullying or harassment. It is of prime importance to be able to detect such content when it is posted by users, and to moderate the content prior to publishing it. *Harassment*, a less targeted version of bullying, is defined as communications in which the user intentionally annoys one or more others in a web community [32]. Cyberbullying and harassment comments share the same features of containing text that is designed to annoy, to hurt and to abuse. In this paper, we address the problem of how to detect such content automatically. We examine text mining strategies to enable us to automatically detect cyberbullying and harassment content for social media sites, using eight datasets. We use the general term *abusive* content to refer to user content that has been interpreted as either cyberbullying or harassment.

Published research related to abusive content on social media sites has initially concentrated on identifying the nature and extent of the problem as it has emerged over the past decade. More recently, research on the computational approaches to detecting such content has gained traction. Recent work has focussed on the use of natural language processing (NLP) and machine learning approaches. Existing work has examined the use of suitable features for classification, and the improvement of classification algorithms. We explore these in section 2.

The use of text mining techniques hold promise for this area, given the requirement for automated classification and the textual nature of user comments. There are a wide range of social media sites, each with their own information formats, purposes and target user bases. To support our work, we have selected eight labelled datasets across the following social media sources: Twitter, MySpace, Formspring, Kongregate, YouTube and SlashDot. Each of these datasets contain user comments that have been categorised into two classes: positive classes for instances that are labelled as containing abusive content and negative classes for instances that are deemed not abusive. We focus on a number of points: (1) How to address the inherent imbalance of social media datasets when used for automated classification. There is typically a far higher proportion of negative class instances than positive, thus lowering detection accuracies for abusive content. (2) Optimal text representations for the actual dataset instances. (3) Appropriate strategies to address the high feature dimensionality that is inherent in social media datasets due to the textual nature of the social media user content. Dimension reduction techniques including feature selection and extraction strategies are examined to enable faster, more accurate detection. (4) Analysing detection accuracies using generic structural features that are common to all our datasets. Our results demonstrate that the impact of these features is limited for improvement of accuracies in detection of abusive content.

A key contribution of this paper is the application of core text mining techniques for abusive content detection across a variety of social media datasets from leading social media sites. Existing works on text mining in the social media domain use a smaller number of datasets in comparison - whereas our approach using eight datasets allows us to compare text mining techniques for abusive content across a variety of social media forums. In particular, our work contains an empirical comparison of core text mining techniques in the social media domain, which will be useful in the selection of applicable approaches. As part of this, our findings highlight the occurrence and treatment of imbalanced classes in order to maximise abusive content detection accuracies. We also identify the importance of appropriate metrics in this domain, where the impact of a false negative result is arguably higher than a false positive.

The paper is laid out as follows: Section 2 discusses related work in the area of text-mining and abusive content detection. Section 3 explains our approach to our work, the techniques to be applied and the datasets used. Our experimental work and results are presented in section 4, with conclusions in section 5.

2 Related Work

Text mining has been widely used for analysing user generated content in social media sites and online services, in the areas of sentiment analysis [24, 23], spam filtering [17, 20], political orientation [5] and recommendation systems [33]. The use of such techniques to detect abusive content has only emerged in recent years. A barrier for the use of text mining techniques for abusive content detection is the lack of labelled datasets in the field. At present, researchers collect data, and annotate by one of two approaches - their own labelling effort [30, 9, 1, 19] which is time-consuming or through the use of crowdsourcing services [25, 2] such as Amazon's Mechanical Turk which can be costly.

A key to successful classification is the generation of appropriate features and this is a focus of research effort in the field of abusive content detection. Reynolds et al. [25] developed a simple keyword matching approach that uses a list of profane words. In order to avoid misspelling and abbreviation, Sood et al. [28] improved on this static keyword-based approach by using the Levenshtein Distance. However, a high percentage of profanity words do not in fact constitute inappropriate content, so are not suitable for discrimination [16].

Much research in this area uses the standard feature representation of Bag-of-Words and N-grams [30, 19, 32] but have not included any feature selection techniques to improve effectiveness and efficiency. Other research takes extra information into account to enhance classification accuracy such as user profile features [9, 8] including age and gender; semantic features of the user comment [30, 9, 1, 32] such as parts of speech, number of pronouns; and features such as profanity word occurrences [1, 32, 4]. Text context features have also been analysed in recent years. Given the interactive nature of cyberbullying, Bayzick et al. [1] processed conversations by using a moving window of 10 posts to capture context features; Dadvar et al. [11] boosted cyberbullying detection performance by using history activities, including the frequency of profanity in the user's previous comments. Although additional features improve the accuracy results on given datasets, their limitation is that they may not generalise well across different sources as they are specific to the domain or source from which they are gathered [18]..

Other research has focused on investigating classification algorithms. Well-known classifiers have been used in this domain including Support Vector Machines (SVM) [9, 16], Näive Bayes [4, 13], logistic regression [29], and decision trees [25, 16]. Classifier ensemble solutions have also proven successful. Dinakar et al. [13] identified that automatically detecting cyberbullying by using binary classifiers for individual labels in a multiclass problem outperforms multiclass classifiers. Silva et al. [7] presented an ensemble classifier that predicts content by averaging a variety of classifiers results in order to avoid over-fitting.

In addition to detecting abusive text content, previous research has focussed on detecting other abuse related information through the use of machine learning. Dadvar et al. [10] detected cyberbullies instead of detecting text content; Nahar et al. [21] used social networks to present a graph model, identifying the most active cyberbullying predators and victims; Xu et al. [31] explored the detection of roles within cyberbullying, and identified those of bully, victim, accuser and reporter. In addition, research [30, 13, 29] used Latent Dirichlet Allocation (LDA) to extract the main topics for each text content. Using this, Xu et al. [30] identify what topic of cyberbullying merits focus;

In our work, we focus on improvements to abusive content detection performance using text mining techniques that can be used across any social media user comment text sources. We will use a proven supervised learning classifier for our algorithm, concentrating on the evaluation of core text mining approaches that enhance the feature representation and dimension reduction part of the classification process.

3 Approach

We have used eight published datasets from social media sources that have already been labelled for cyberbullying or harassment. The datasets have several common characteristics: (1) All datasets consist of user comments posted by users of the service. (2) The datasets are imbalanced, typically containing far fewer positive (abusive) instances than negative (non-abusive). (3) The datasets contain unstructured text, with a resultant high number of sparsely populated features when tokenised. We focus on identifying optimal text representations, dimension reduction techniques, the need and approach for class re-balancing, and finally, the impact of generic structural features of the data. Firstly, we explain each of the datasets used, followed by a description of text mining and classifier techniques we propose to use.

3.1 Datasets

Our eight datasets will be referred to as D1, D2 through to D8 for the reminder of the paper. They have been collected from publicly available social media message boards and user comment areas. The social site platforms are varied, including Question&Answer, forum, micro blogging, chat and multimedia sharing. All sources contain free-format user comments, so are prone to receiving abusive content. The following is an overview of each dataset, with Table 1 showing the source and summary statistics including the number of instances, the proportion of positive and negative instances and the average number of characters across instances.

Table 1: Datasets Summary Statistics

	Data	**Dataset**	**Labelled**	**# of**	**Avg**	**Class Dist.**
	Source	**Style**	**for**	**Instances**	**Len**	**(Pos./Neg.)%**
D1	Twitter	Micro-Blog	Cyberbully	3110	72	42/58
D2	YouTube	Video-Sharing	Cyberbully	3466	887	12/88
D3	MySpace	Forum	Cyberbully	1710	1503	23/77
D4	Formspring	Q&A	Cyberbully	13153	101	6/94
D5	Kongregate	Chat	Harassment	4802	21	1/99
D6	SlashDot	Forum	Harassment	4303	429	1/99
D7	MySpace	Forum	Harassment	1946	251	3/97
D8	Twitter	Micro-Blog	Cyberbully	1340	67	13/87

D1-Twitter Xu et al. [30] collected a dataset from Twitter using the rule that each collected tweet contains at least one of the following keywords: bully, bullied and bullying. Each instance in this dataset was annotated by five non-author expert annotators. Twitter entries are short (140 characters maximum) and unthreaded, as retweets are not included.

D2-YouTube Dadvar et al. [9] created a corpus of comments from the video upload site, YouTube, by scraping comments from sensitive cyberbullying topics within the site. Each labelled instance consists of a single user's comment. The data was manually labelled for cyberbullying by non author researchers.

D3-MySpace Bayzick et al. [1] crawled posts from MySpace, which is thread-style forum social website. To allow for newer comments in a thread that may deviate from the initial topic, the authors grouped comments using a moving window of 10 posts to create a single post which then was labelled manually by three annotators for cyberbullying. So each instance consists of up to 10 different users' comments which is the longest average length (1503 characters) among the eight datasets. If any one individual comment contain abusive content, the overall instance containing the 10 posts is labelled as abusive.

D4-Formspring This popular question and answer style social site was crawled by Reynolds et al. [25] who used Amazon's Mechanical Turk, a crowd-sourcing service, to label the data for cyberbullying. Each instance contains a single user answer to one question.

D5-Kongregate, **D6-Slashdot**, **D7-MySpace** These datasets were selected by Yin et al. [32] from a corpus labelled by Fundacian Barcelona Media (FBM) for the CAW 2.0 workshop[1] for a task on harassment detection. D5 was collected from the chat style website, Kongregate, which provides an online space for users to discuss

[1] url: http://caw2.barcelonamedia.org

the game industry. The content is short length with few words (21 characters on average). Each instance is one user's comment but not always relevant to games. Slashdot, which was used for D6, is a web publisher of technology stories. It is a forum site of thread-posts style. Users can post comments on specific topic. Each post is a standalone post and belongs to one thread. D7, as the same source as D3, is gathered from Myspace and consists of standalone posts. However, unlike D3, posts in D7 are labelled for harassment individually.

D8-Twitter Mangaonkar et al. [19] scraped tweets from web pages that had reported cyberbullying instances. Non-bullying instances were gathered randomly from Twitter. The authors then manually labelled the data (single tweet) for cyberbullying for the purpose of validation.

For each dataset, we perform the following pre-processing operations: all letters are changed to lowercase; links started with 'http://' or 'https://' are replaced by the generic term 'url_links' ; names following the '@' symbol are replaced by the anonymous name '@username'. Given that social media site user comments are typically conversational in style and short, we have not implemented the removal of stop-words or stemming.

3.2 Methodology

3.2.1 Validation

To conduct our tests, we use 10-fold cross validation on each dataset. Each dataset is randomly partitioned into 10 equally sized and stratified groups or folds (the ratios of positive instances to negative instances are the same in each fold and in the original dataset). The model training is run 10 times, with each fold held back exactly one to be used as test data, with the remaining folds used for training. The results are then averaged across the folds, using suitable performance measures, to determine the accuracy of the positive and negative classes' detection.

3.2.2 Measures

The ultimate purpose in our work is to boost the ability to detect abusive content. In this context, the impact of a false negative (abusive content not detected and reaching the user audience) is arguably higher than a false positive whereby clean content is wrongly categorised as abusive and held back from the user base. Therefore, we need to focus in particular on the accuracy of the *positive class* (instances containing abusive content).

In addition, most of the datasets have an imbalanced class distribution, as shown in Table 1, with a far higher occurrence of negative instances. The Kongregate dataset for example, has just 1% of the dataset contents tagged as abusive content. In this case, viewing accuracy at dataset level, an accuracy of 99% can be achieved by blindly categorising every instance as negative. We will therefore use *average class accuracy*, also known as *average recall*, rather than overall dataset level accuracy in the rest of paper to avoid hiding underlying issues with the positive class. Recall for a class is calculated as total instances correctly identified for the class, over total number of instances of that class. The equation for *Recall* for the positive class is given in Equation 1.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{1}$$

3.2.3 Text Representation

Both bag-of-words (BoW) and N-grams are commonly used for text representation in text-mining. In both cases, the text has been split into a set of tokens or features. The BoW approach simply considers all text as an unordered set of word features, with each separate word becoming a single feature. N-grams, an improvement on BoW in the text-mining field, split the text into a set of features consisting of N continuous sequential character occurrences. These two approaches have been widely and successfully used in text classification in domains such as sentiment analysis and spam filtering. We use both tokenising approaches and use normalised term frequency tf_{ij} for the feature value of feature i in instance j as given by Equation 2 where the numerator is the number of feature i in instance j and the denominator is the euclidean distance of all features in instance j.

$$tf_{ij} = \frac{Num_{ij}}{\sqrt{\sum_{i=1}^{n} {Num_{ij}}^2}} \tag{2}$$

3.2.4 Classifier

For this work, we start with two commonly used classifiers from the supervised learning domain which perform well for text classification. The Näive Bayes (NB) classifier is based on applying Bayes' theorem, with independence assumptions between the features. It is one of the simplest and most efficient algorithms in classification, and is used extensively in text classification. Support vector machines (SVMs), a maximal margin classifier algorithm, which is very effective in high dimensional spaces is also used extensively for text classification.

3.2.5 Baseline Results

Using the original datasets, the performance of two classifiers, SVM and NB, each with the text representation of BoW and N-grams was measured. Figure 1 shows the results with the actual performance of each dataset for each classifier/representation denoted by a dot. This graph is a modified box-plot where the average performance of each classifier/representation choice is denoted by the longer horizontal line and the standard deviations shown by the shorter horizontal lines. Using the Friedman test, we determined that there is no statistically significant difference between the accuracy results for any individual dataset obtained using either of the two classifiers (SVM and NB) with either of the two text representations (BoW and N-grams), N-grams is character level, ranging from 2 to 4. For the remainder of this work, we will therefore use SVM as our baseline classifier and Bag-of-words as feature representation.

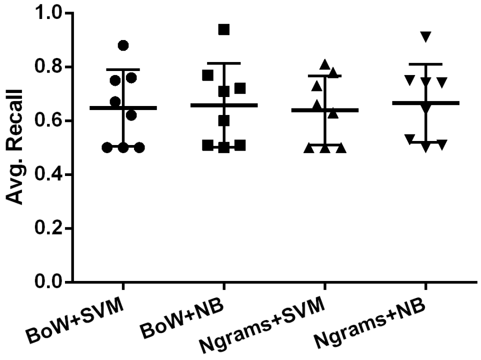

Fig. 1: Average recall using Naive Bayes & Support Vector Machine classifiers on all datasets, using BoW and N-Grams text representations

3.3 Dataset Balancing

All of our datasets show imbalance across the classes, with the majority of datasets showing a very small proportion of positive instances. D1 and D3 are less imbalanced due to the initial selection done when choosing the data. D1 from Twitter only includes tweets that contain the word 'bully' or a derivative, which is likely to have a higher concentration of bullying tweets than in a typical random Twitter dataset. D3 from MySpace was annotated using a moving window to group ten posts at a time, with the result that a single bullying post will appear in a multiple groups, boosting the occurrence of the positive instances by up to ten. Imbalanced datasets can cause serious issues in classification [15]. Without sufficient knowledge to learn from the minority classes, classifiers may over-assign instances to the majority classes. We tested the accuracy of detection on the raw imbalanced datasets. Figure 2 shows how bad the positive recall results are for the datasets when a large imbalance occurs, with no detection of abusive instances in datasets D2, D6 and D7.

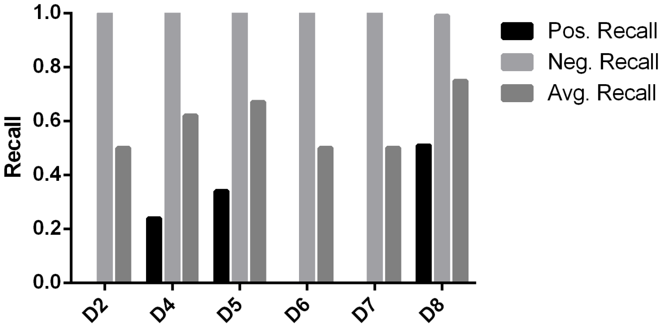

Fig. 2: Positive, negative and average recall for imbalanced datasets using an SVM classifier

In order to re-balance the datasets, we will use an approach of randomly oversampling [3] the positive class in datasets D2, D4, D5, D6, D7 and D8 to balance the two categories. To briefly explain sample size influence, we look at D2 and D8 as examples. Figure 3 shows the change in positive recall as the number of the positive class instances increases with resampling. The resample intervals on the x-axis are increasing at intervals of 0.5 (i.e. by 50%). Positive recall increases as the minority class is resampled. The impact of re-balancing across the datasets is examined in more detail in our experimental section.

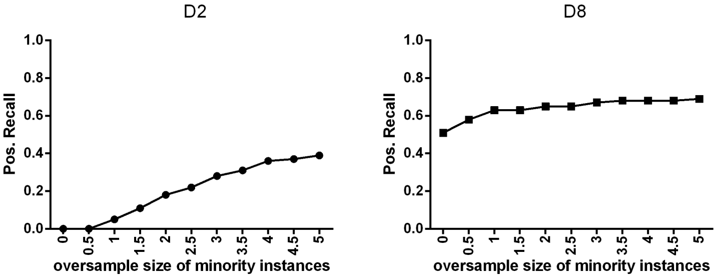

Fig. 3: Positive recall for D2 and D8 as oversample instances are increased at intervals of 50%

3.4 Dimension Reduction

The representation created from each dataset is both high dimensional, in terms of features, and sparse. For our datasets, the numbers of features using BoW or N-grams (character level, 2-4) are shown in Table 2.

Table 2: Number of features in each dataset

	D1	D2	D3	D4	D5	D6	D7	D8
BoW	6301	44202	11258	17475	3639	18417	10381	4085
N-grams	43431	161622	57932	89067	25181	84422	56273	29861

For many learning algorithms, training and classification time increases directly with the number of features and high numbers of features may even negatively impact on classifier accuracy [6]. An effective approach to reduce the number of features is the use of *document frequency (DF)* reduction [26]. DF uses the number of features that occur in a corpus, such that those features that occur most often and least often can be removed. Figure 4 shows the average reduction in the number of features across all datasets as we adjust the threshold for DF from 0.1% up to 1% on BoW.

At the 0.1% threshold, where the most and least frequent 0.1% of features are excluded, we can see for most of datasets, at least a 50% reduction in the volume of features. In our experimental section, we evaluate the impact on our detection accuracies of using DF.

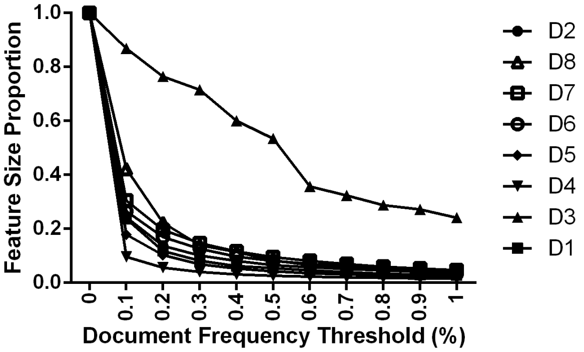

Fig. 4: Average feature size as a proportion of the original feature size across all datasets when applying DF reduction. The distinctive line is from D3, which grouped every 10 posts as one 'window' of content. In this case, every word occurrence is multiplied, causing the line of document frequency to decrease smoothly.

Even after the use of DF reduction to reduce the number of features, we still have high dimensionality for our datasets. To further enhance classification performance, we investigate two more advanced dimension reduction approaches, namely, *chi-square* and *singular value decomposition* (SVD), which are *feature selection* and *feature extraction* techniques respectively.

Feature selection involves techniques that choose the best subset of the existing features [6]. Typically they rank the features using algorithms that correlate the features to the target class label and choose the top ranked features. Widely used approaches include information gain, odds ratio and the chi-square measure. In this paper, we deploy chi-square, which is a statistical test for comparing proportions. It produces χ^2 scores of correspondence between features and categories in the dataset. Depending on the χ^2 scores ranking, we can eliminate noisy and less predictive features to significantly reduce the dimensionality without losing classification accuracy.

In contrast to feature selection, *feature extraction* is a dimension reduction approach that transforms the existing features to a set of alternative, more compact, features while retaining as much information as possible. Common methods include the unsupervised principle component analysis (PCA) approach which performs a singular value decomposition (SVD) to transform the data into a reduced feature space that captures most of the variance in the data. We will apply SVD for dimension reduction.

3.5 Additional Structural Features

A key step to successful use of supervised learning is the choosing of appropriate features [14].We propose a list of additional text structural features which can be extracted from the dataset.

We specifically wish to focus on generating features that are suitable across all of our eight datasets. However, generic features are challenging to find, due to the different characteristics of the various datasets. For example, Twitter data contains information such as number of followers, number of friends and the hashtag topic, but these cannot be retrieved from other datasets. For YouTube data, we can obtain information such as the number of thumbs up, number of thumbs down, number of replies and the comments date. User profile information such as username, gender and age is not uniformly available across each of the datasets. Such features are likely to be rich sources of information, and will be the focus for future, domain specific work. Focussing on generic features, we extracted the following information for each dataset:

- The average word length and average character length;
- The count of punctuations, uppercase and URL respectively;
- The ratio of punctuations and uppercase to characters;
- The count of sentences, separated by '.!?';
- The ratio of word length to sentences and ratio of character length to sentences

4 Experimentation and Results

Our experimental work investigates the following techniques on abusive content detection in our eight datasets D1 to D8, as discussed in our approach: dataset re-balancing, dimension reduction using document frequency, feature selection and extraction, and the addition of generic structural features.

4.1 Dataset Resampling

We demonstrated that oversampling techniques can increase the positive recall, as shown previously in Figure3. However, it is evident that excessive oversampling can over-fit the minority class instances and damage the majority class accuracy. In Figure 5, we explore D2 as an example to briefly explain the impact of different resampling proportion. Positive recall which indicates accuracy of correctly detecting abusive content is boosted as oversampling increases. On the contrary, negative recall is falling off. To balance the average recall, we therefore considered the problem of choosing oversample size as a trade-off and applied two target rules:

1. *Obtaining the best results on the minority class*
2. *Obtaining the least damaging results on the majority class*

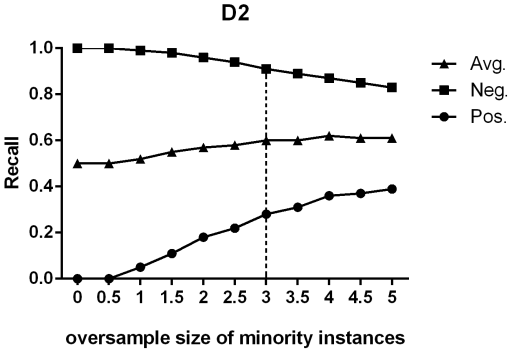

Fig. 5: Optimal point of resampling size for D2

Table 3 shows the results of applying the same heuristics across the remaining imbalanced datasets, where optimal rebalancing of the positive class has been determined. As D1 and D3 were not seriously imbalanced, we excluded them from this resampling. Using the ratio of re-balancing in the Table 3, we examined the effects of resampling by comparing the classification results of the original datasets and the resampled datasets. In order to reduce the impact of random selection, we supplemented our cross-validation approach by resampling for each of our training folds three times, and recording average results across the three resampled configurations for each fold.

Table 3: Ratio of the positive instances (+) to negative instances (-) before and after resampling

	D2	**D4**	**D5**	**D6**	**D7**	**D8**
	(+/-)	**(+/-)**	**(+/-)**	**(+/-)**	**(+/-)**	**(+/-)**
Before	12/88	6/94	1/99	1/99	3/97	13/87
After	35/65	20/80	9/91	10/90	20/80	38/62

Figure 6 shows the results on two metrics, positive recall and average recall. It is clear from the both recall graphs that randomly resampling imbalanced datasets increases the accuracy of abusive content detection in all cases. It is worth noting that for the original datasets D2, D6 and D7 where no abusive content was detected using unbalanced data, oversampling detection increases significantly.

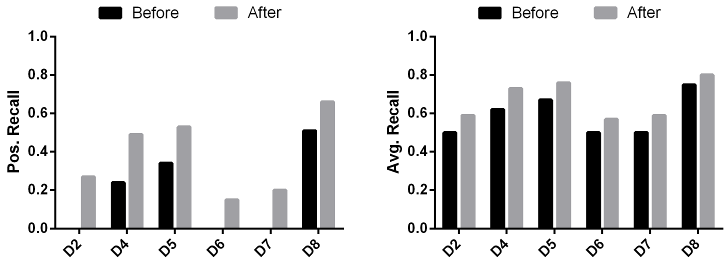

Fig. 6: Positive recall (left) and average recall (right) on datasets before and after resampling

To sum up, resampling techniques are an important consideration if the data has a large imbalanced class distribution. This is particularly relevant in the detection of abusive content on social media, where abusive comments will typically form a small fraction of all content.

4.2 Dimension Reduction

With dimension reduction, we wish to determine whether we can reduce the high volume of features without damaging detection accuracies. We examined the performance of the three dimension reduction techniques discussed in our approach, in section 3.4; one is the simple dimension reduction technique using document frequency (DF), the other two being more advanced approaches, namely, chi-square and SVD. We applied DF first on our resampled datasets. Using our DF results as a baseline, we then compared the performance of feature selection using chi-square and SVD.

Whilst we wish to reduce the high number of features to increase training and classification performance, we need to ensure that we do not damage detection accuracies. To verify this, we compared average recall for each of our datasets, using the original datasets and the datasets reduced using the 0.1% DF threshold identified in Section 3.4. Figure 7 shows that average recall for each dataset is approximately the same with and without using 0.1% as threshold of DF. We applied the paired-t-test (p=0.38) to confirm that there is no statistical difference. This confirms that we can significantly improve the efficiency of classification through dimension reduction, without impacting class accuracy.

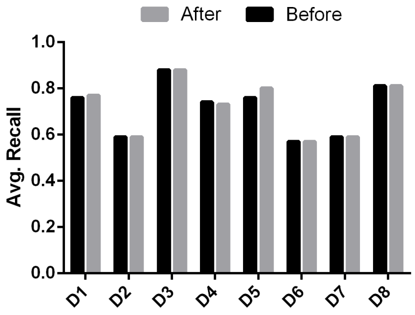

Fig. 7: Average recall of each dataset before and after DF

In the following experiment, we apply chi-square feature selection to examine the impact of further reduction, using the reduced datasets post Document Frequency reduction to train our model. Chi-square feature selection uses the χ^2 score to select a subset of the features with highest scores. In order to see the impact of selecting different numbers of features, we examined average recall for all datasets as we decrease the number of features selected in steps of 10%. Figure 8 shows that the average recall performance is maintained at a steady level across all datasets with a slight drop off in accuracy starting to appear when the number of features has been reduced by approximately half.

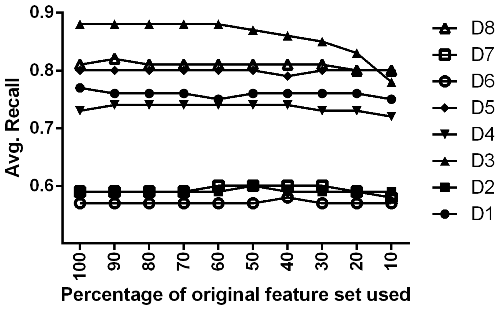

Fig. 8: Average recall using chi-square feature selection (with DF 0.1% as baseline) as the percentage of features selected reduces.

Unlike chi-square which allocates a score for all features, for the SVD algorithm we are required to specify a number of features. To compare both feature selection and feature extraction, Figure 9 shows the average recall of all datasets reduced to 10% of DF baseline feature size (i.e. after using Document Frequency (0.1%)), using both chi-square and SVD dimension reduction. Using the Friedman test, we determine that there are no significant statistical difference between these three feature dimension approaches.

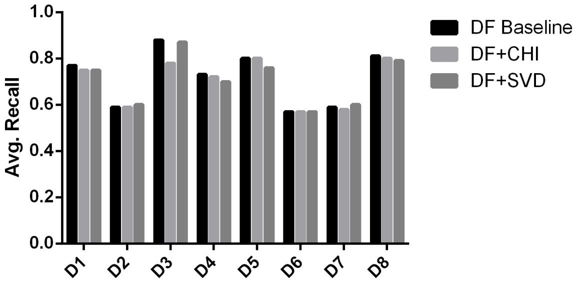

Fig. 9: Average recall of comparing CHI and SVD by reducing features to 10%.

We have shown that feature selection (e.g. using chi-square) or feature extraction techniques (e.g. SVD) can be applied for reducing the feature dimensionality without significantly reducing accuracy. The choice of dimension reduction technique is normally made with some preliminary experimentation for the domain in question. As feature selection techniques retain the original features they can be more directly interpretable as opposed to feature extraction where the transformed features have no meaning. In addition feature extraction takes typically significantly longer to execute than feature selection [2] but as training a classifier model is not normally required in real-time and can occur offline, this may not necessarily be a disadvantage.

4.3 Impact of Structural Features

We then examined whether we can improve our detection accuracies by introducing the dataset structural features identified in Section 3.5. As the value of bag-of-words term frequency feature is from 0 to 1, for each dataset, we used min-max normalization technique to transform structure features value to 0-1 range and then added them to the DF baseline features. The impact of the new features on class accuracies for the eight datasets is shown in Figure 10. Average recall (right), as well as positive recall (left) for D1,D2,D7,D8 datasets has increased, particularly for D2 where positive recall performance has increased by almost 20%. However, for the remaining datasets, the results are approximately the same. We looked through the general feature distribution of each dataset, finding no pattern of the new features for all datasets. For example, in D2, the average length of positive instances is bigger than negative however, but in D6, it is opposite. Therefore, since each dataset has unique characteristics, the addition of general features cannot guarantee a significant beneficial impact on classification results for all datasets. Domain features extracted from specific datasets need to be explored for further research.

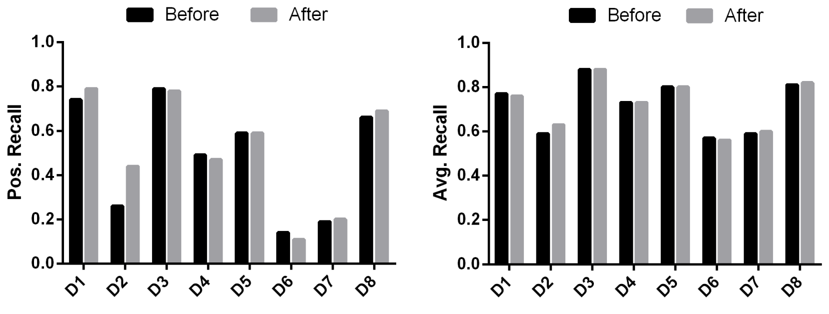

Fig. 10: Positive (left) and average recall (right) across the datasets with and without the additional structural features

[2] e.g. On D3, SVD reduction to 10% features took 20.75 secs while chi-square reduction took 0.03 secs

Table 4: Avg. and Pos. recall of all experiments using BoW txt representation and SVM classifier. The threshold of DF is 0.1%, chi-square and SVD are 10% respectively

		Original	**Re-Balanced**	**DF**	**DF+CHI**	**DF+SVD**	**Structure Feat.**
D1	**Avg.**	0.76	0.76	0.77	0.75	0.75	0.76
	Pos.	0.75	0.75	0.74	0.72	0.72	0.79
D2	**Avg.**	0.5	0.59	0.59	0.59	0.6	0.63
	Pos.	0	0.27	0.26	0.24	0.28	0.44
D3	**Avg.**	0.88	0.88	0.88	0.78	0.87	0.88
	Pos.	0.78	0.78	0.79	0.57	0.77	0.78
D4	**Avg.**	0.62	0.73	0.73	0.72	0.7	0.73
	Pos.	0.24	0.49	0.49	0.47	0.42	0.47
D5	**Avg.**	0.67	0.76	0.8	0.8	0.76	0.8
	Pos.	0.34	0.53	0.59	0.61	0.52	0.59
D6	**Avg.**	0.5	0.57	0.57	0.57	0.57	0.56
	Pos.	0	0.15	0.14	0.14	0.16	0.11
D7	**Avg.**	0.5	0.59	0.59	0.58	0.6	0.6
	Pos.	0	0.2	0.19	0.18	0.21	0.2
D8	**Avg.**	0.75	0.8	0.81	0.8	0.79	0.82
	Pos.	0.51	0.66	0.66	0.67	0.65	0.69

5 Conclusion

The purpose of this paper was to investigate the power of core text mining techniques to detect abusive content across a range of social media sources. For our work, we selected eight social media datasets, from Twitter(2), YouTube, Kongregate, Formspring, MySpace (2) and SlashDot. Using these, we conducted an empirical study of the following and their impacts on content detection accuracy: class re-balancing; optimal text representations; dimension reduction using document frequency, feature selection (chi-square) and feature extraction (singular value decomposition) techniques; and the use of generic structural features. The purpose of rebalancing and generic structural features usage was to *improve* detection accuracies, whereas DF, DF+chi-square and DF+SVD were applied to obtain efficiency gains whilst *maintaining* accuracy. Table 4 shows a summary of the experimental results.

We highlight the following observations: Firstly, data balancing is a key consideration in social media due to the highly imbalanced nature of the data, with small occurrences of the critical positive (abusive) class. Using resampling, we obtained significant gains in positive recall but noted a critical point at which the negative class is then impacted. Secondly, given the textual nature of social media data, the management of feature numbers is a key issue. We noted no difference on class accuracies between bag-of-words and N-Grams text representations. As a simple technique, we found that a document frequency of just 0.1% threshold for document feature frequency was sufficient to preserve class accuracy, whilst achieving a 50% reduction on feature numbers. Both chi-square and singular value decomposition for further feature selection offer large reductions in the number of features, whilst maintaining good class accuracies. Only D3 has an obvious fall on the results using chi-square (positive recall falls from 0.79 to 0.57). We suggest that this decrease is due to the occurrence of multiple repeated posts across instances in D3 which is affecting the χ^2 score of each feature. Thirdly, the impact of generic structural features that can be applied across all the datasets is limited. To boost classification accuracies further, we need to consider using domain features specific to individual datasets in the further work. Finally, we note that there is a wide variation in detection accuracies across the eight datasets, such as 0.78 positive recall in D3 to 0 in D6 using the original datasets. There are a number of factors impacting this variation. Datasets from various sources have their own unique characteristics, such as different distributions of positive to negative instances, average length across all instances and distinct language phrasing, all of which can impact the performance of classifier. In addition, the very nature of abusive data leads to subjective labelling decisions. Our eight datasets have been gathered by a variety of researchers, using a variety of labelling methodologies. This can lead to potential inconsistent signals in the training data across the various datasets.

Our future work in this area is focussed on exploiting human-in-the-loop learning such as active learning to provide a continuous learning capability in this domain. Levels of confidence are associated with predictions and where there is a certain level of uncertainty, typically due to previously unseen or new types of abusive content, human users are used to confirm the prediction and provide new and better examples of abuse from which to learn. We will also consider enhancing classification performance by adding domain specific features where appropriate.

References

1. Jennifer Bayzick, April Kontostathis, and Lynne Edwards. Detecting the presence of cyberbullying using computer software. 2011.
2. Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 2015.
3. Nitesh V Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 853–867. Springer, 2005.
4. Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing (SocialCom)*, pages 71–80. IEEE, 2012.
5. Raviv Cohen and Derek Ruths. Classifying political orientation on twitter: It's not easy! In *ICWSM*, 2013.
6. Pádraig Cunningham. Dimension reduction. In *Machine learning techniques for multimedia*, pages 91–112. Springer, 2008.
7. Nadia FF da Silva, Eduardo R Hruschka, and Estevam R Hruschka. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179, 2014.
8. Maral Dadvar, FMG de Jong, RJF Ordelman, and RB Trieschnigg. Improved cyberbullying detection using gender information. 2012.
9. Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Advances in Artificial Intelligence*, pages 275–281. Springer, 2014.
10. Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Canadian Conference on Artificial Intelligence*, pages 275–281. Springer, 2014.

11. Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.
12. Laura P Del Bosque and Sara Elena Garza. Aggressive text detection for cyberbullying. In *Mexican International Conference on Artificial Intelligence*, pages 221–232. Springer, 2014.
13. Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*, 2011.
14. Pedro Domingos. A few useful things to know about machine learning. *CACM*, 55(10):78–87, 2012.
15. Vaishali Ganganwar. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Tech. and Advanced Eng.*, 2(4):42–47, 2012.
16. Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*, 2015.
17. Congrui Huang, Qiancheng Jiang, and Yan Zhang. Detecting comment spam through content analysis. In *Web-Age Information Management*, pages 222–233. Springer, 2010.
18. Henry Lieberman, Karthik Dinakar, and Birago Jones. Let's gang up on cyberbullying. *Computer*, 44(9):93–96, 2011.
19. Amrita Mangaonkar, Allenoush Hayrapetian, and Rajeev Raje. Collaborative detection of cyberbullying behavior in twitter data. In *Electro/Information Technology (EIT), 2015 IEEE International Conference on*, pages 611–616. IEEE, 2015.
20. Michael Mccord and M Chuah. Spam detection on twitter using traditional classifiers. In *Autonomic and trusted computing*, pages 175–186. Springer, 2011.
21. Vinita Nahar, Xue Li, and Chaoyi Pang. An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, 3(5):238, 2013.
22. Dan Olweus. Bullying at school. what we know and whal we can do, 1993.
23. Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326, 2010.
24. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
25. Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 241–244. IEEE, 2011.
26. Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.
27. Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385, 2008.
28. Sara Sood, Judd Antin, and Elizabeth Churchill. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1481–1490. ACM, 2012.
29. Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM Int Conf on Information and Knowledge Management*, pages 1980–1984. ACM, 2012.
30. Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Proc of the 2012 Conf of the Nth American chapter of the ACL: Human language technologies*, pages 656–666. ACL, 2012.
31. Jun-Ming Xu, Xiaojin Zhu, and Amy Bellmore. Fast learning for sentiment analysis on bullying. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10. ACM, 2012.
32. Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7, 2009.
33. Xiaohui Yu, Yang Liu, Xiangji Huang, and Aijun An. Mining online reviews for predicting sales performance: A case study in the movie domain. *Knowledge and Data Engineering, IEEE Transactions on*, 24(4):720–734, 2012.