Text Classification of NSFW Reddit Posts

Anonymous Author(s)
Affiliation
Address
email

Abstract

Filtering inappropriate content on the Internet correctly remains an active challenge in the machine learning community. This problem is exacerbated by the high volumes of new content added every second to media websites. The Reddit website is one of the largest growing collections of user-submitted content. This paper examines several current text classification techniques to classify Reddit posts as Not Safe For Work (NSFW) using the sparse data given in a single post.

Figure 1: A wordle [5] of common terms in a filtered set of NSFW posts

1 Introduction

1.1 Overview

Since the early days of the Internet, various strategies have been used to filter or flag inappropriate content before users can view it. Many software packages exist that create a virtual firewall for specific content and will stop certain websites from being loaded. More commonly, websites are now flagging user generated content so that users won't accidentally view inappropriate content in their workplace. This flag has been named Not Safe for Work or NSFW. This flag covers not only pornographic images, but also content containing disturbing or other lewd images or text.

Early technologies relied on managed lists of inappropriate URLs to block access. Given the huge rise in new content, and easy strategies to circumvent such lists, technology began to use primitive

filtering techniques. These involved looking for certain keywords in the website and searching for associated websites [3]. Machine learning methods began to be applied including analysis of linked documents and embedded URLs.

Text classification has many different applications, from suggesting keywords for documents, to classically detecting spam emails. It has been applied in various ways for content filtering where the problem is modelled as a binary classification problem. This approach then makes use of the vast research in classification algorithms in machine learning.

The Reddit website is a website designed for user-generated content with over eight million regular users. A user can post a topic with a title, and either a link to a webpage or a short piece of text to elicit discussion [14]. A user can also tag their post as NSFW or later another user can do so. The website receives a huge amount of new volume constantly. This means that not every NSFW post will be flagged appropriately either because the posting user decided not to, or it was only viewed by a small number of users who also failed to flag it. A user can click on a link unaware of where it will take them, and be presented with content that could cause disciplinary action in many workplaces. Due to the huge volume of traffic, it is important that the classification can be successful from only the details in the Reddit post and not necessarily content at the associated URL. Therefore a fast and accurate algorithm for detecting and flagging NSFW posts would be a greatly valuable addition to the Reddit infrastructure.

This paper examines text classification techniques and appropriate machine learning classifiers applied to this binary classification problem. Firstly various feature extraction methods are tested based on the 'bag of words' concept to generate feature vectors for each post. This involves examining which fields of a Reddit post are the most important for successful classification. Then several different binary classifiers are examined the various benefits. In the end we present a reliable method for NSFW classification of new Reddit posts.

1.2 Related Work

Many existing commercial software tools including NetNanny [8] have been developed for filtering internet content. These use a mix of heuristics and internally administered black-lists for blocking content such as [1]. Early approaches [7] for examining the content of website included examining linked pages, the content of images and specific keywords. Newer machine learning approaches use a full text and image classification strategy on the content of the website such as [11] and [2].

Most text classification problems involve large corpora such as full news articles. The recent rise of Twitter has changed researchers focuses to using text classification techniques on short text sequences of less than 50 words. Much research has involved sentiment prediction from messages on Twitter such as [6]. Other researchers have specifically examined Reddit as an interesting source of predictive machine learning data such as [13] which attempted to predict post popularity. However the area of filtering for inappropriate content using short texts remain an interesting field of research.

2 Testing Methodology

2.1 Data

Selecting the appropriate data was very important for the study. A Reddit scraper was created to pull all posts with associated meta-information from the Reddit website. The scraper collected posts between 21st December 2012 and 8th January 2013. The scraper was executed almost three months later in early April. This allowed the Reddit community sufficient time to have viewed the posts and tagged for NSFW if appropriate. Over 1 million reddit posts were scraped during those dates.

Reddit data can also be scored by users with either an up-vote or a down-vote. This allowed for a metric of a minimum number of users who had viewed the page. By thresholding posts that had a minimum number of total votes, a more robust set of data was generated with a higher likelihood that posts had been appropriately tagged as NSFW. Figure 2 shows the quality of the classifier increase as the minimum total votes is increases which shows the theory. If the threshold was set too high, a large proportion of posts would be omitted. After analysis of the data, the threshold was decided

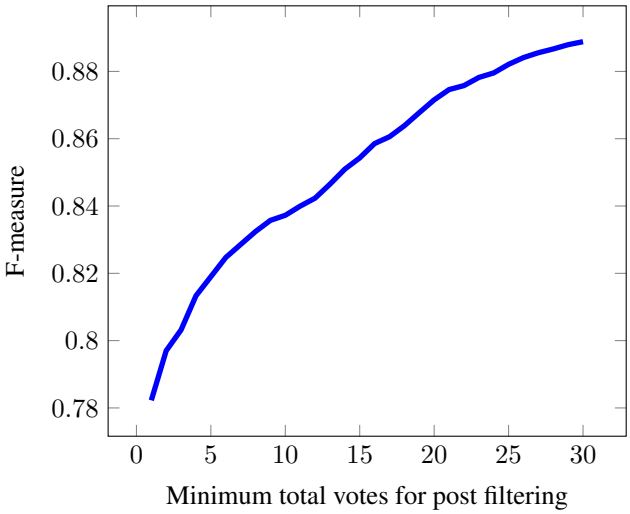

Figure 2: More popular posts are more likely to be correctly annotated, thereby creating a more robust test set

to be the median of the data which was 10. This meant that only posts of above average popularity would be tested, and thereby give a more reliable dataset.

500,000 posts with this threshold were selected. Only 8.6% of these posts were tagged as NSFW which was representative of the normal data. This huge class skew creates an additional challenge in both classification and also proper testing.

2.2 Testing

Because of the hugh class skew, a basic accuracy metric would not be appropriate. This is due to the simple idea that if a classifier tags all posts as Safe for Work, it would be 95% accurate as only 5% of posts are NSFW. It is most important that posts that are NSFW are tagged appropriately, as the potential cost to a user clicking on incorrectly labelled Safe for Work material could be large. On the other hand, Safe for Work posts cannot be tagged NSFW to often to affect the quality of posts on Reddit.

$$F_{measure} = 2\frac{precision*recall}{precision+recall}$$

The F-measure (shown above), trades off precision and recall (also known as sensitivity). Due to the need to balance the success of positive classifications as well as negative classifications, we selected the F-measure as our target metric.

Furthermore the classifiers were testing using 5-fold cross-validation. This meant that in each test-case, the training data had 400,000 posts and the test data had 100,000 posts.

3 Features

3.1 Extraction

Several properties of each Reddit post were captured. The most visible to the user on Reddit is the title and would be a key indicator for the metric. The subreddit, which is the category of the post decided by the user, and the author's username of the post were also captured.

In order to extract numerical vectors from these text features, the 'bag of words' concept was used. This method finds each unique word in the data set, and then for each post counts the number of each word present to generate a very large and normally very sparse feature vector. The idea for 'bag of

words' comes from the simple idea that an email with the word 'viagra' anywhere in it would more than likely be spam.

The 'bag of words' method has been adapted in several ways. Firstly, the technique can be extended to bigrams. This method searches for all word pairings instead of single words and is able to add more contextual information to the feature vector. It has been suggested that often bi-grams are enough to capture the core concept of phrases, e.g. 'United States of America' can be captured by the bigram 'United States'. Therefore the use of tri-grams rarely gives additional gain. This was tested and is also shown in Figure 3.

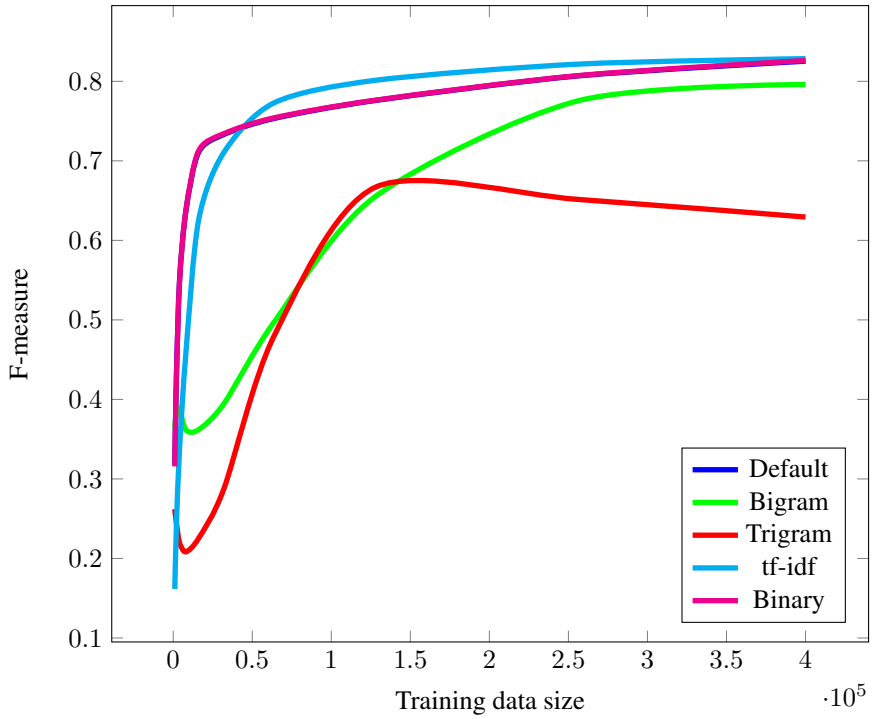

Figure 3: A comparison of different text extraction techniques

The frequency of words and varied text lengths can also skew the feature vector. The tf-idf method normalises the data for length of text and frequency of common words [9]. This is beneficial to this problem so that certain words are not over-weighted and the result of this is shown in Figure 3.

A binary feature extractor was also tested. This would only test for a word appearance and not count word occurrence. In the figure, the binary classifier overlaps closely with the default single word tokenizer and gives no additional benefit. It should be noted that in all these cases, the lower-case text was used. Also stop-words (common words in the English language) and punctuation were also removed.

After cross-validation the tf-idf tokenizer was selected as the best feature extractor. Because of the very large data-sets used, the feature hashing approach was applied [15]. This uses the hash of words to calculate the column ID and does not require on a dictionary reducing the memory and computation requirements. The extraction methods were implemented using the Scikit Learn libraries [12].

3.2 Selection

The 'bag of words' methodology creates huge numbers of features. Specifically using our data set, a basic single word 'bag of words' algorithm creates over one million features.

Traditionally feature selection is an excellent way to prune the number of features to a manageable level. This is beneficial for several reasons. Often classifiers will not be as successful with a very high number of classifiers. This is due to the very high dimensionality of the problem, and the challenge of creating a relevant fit around the given data. It can also be useful for greater understanding in the problem to identify which features are important in the classification and which are not. However it is more challenging in text classification. This is both due to the extremely large number of vectors and also the incredible sparseness of the data.

It has been noted that the results of feature selection in a text classification study vary [4] and depends heavily on the data-set used. We tested the chi-squared technique for feature selection in order to reduce the number of features. The results are shown in Figure 4. These interesting results show the reliance on a small subset of the features for the majority of successful classifications. Further analysis using a non-hashing vectorizer revealed that a significant proportion of the remaining features were related to subreddits. This shows that subreddits are very important in successful classification.

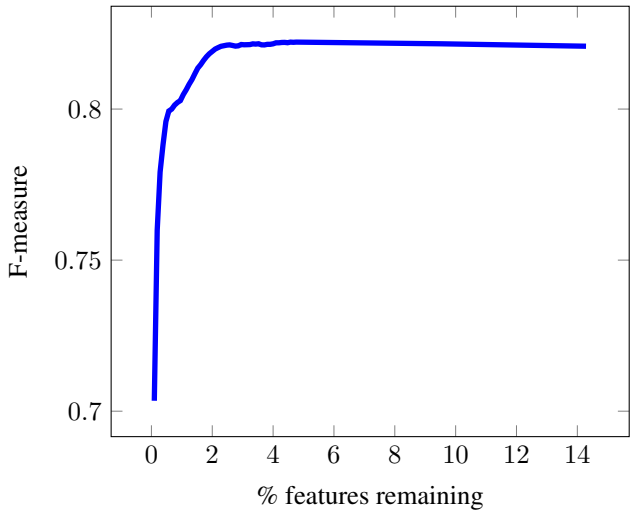

Figure 4: Results of chi-squared removal of different proportions of the feature space

4 Classification

Various binary classifiers were tested on the dataset given using the reduced feature set. Cross-validation was used to adjust the various parameters for the best results for this data set. The results of the different optimised strategies outlined in this section are shown in Figure 5.

The Multinomial Naive bayes method and Bernouilli Naive Bayes method use the basic probabilities of word occurrences as well as the class frequencies to calculate the probability that a post is NSFW. These techniques used the Scikit Learn python library [12].

Neural Networks mimic the behaviour of neurons in the human brain. Each neuron takes in multiple inputs and only fires (giving a particular output) when certain constraints are met on the inputs. A multi-layer network of neurons is built, where each input feature is linked to a neuron and neurons are interlinked on several layers until a single output is given. This output decides whether the feature data given should be classified as SFW or NSFW. Support vector machines, a method for splitting a data-set by transforming data and splitting with a hyperplane, and logistic regression are also tested.

The Vowpal Wabbit tool [10] from Yahoo Research and Microsoft Research was used to test logistic regression, support vector machine and neural networks approaches on the very large data set used. Using cross-validation the size of the hidden layer in the neural network (i.e. the number of additional neurons between the inputs and the output used) is adjusted for optimal results.

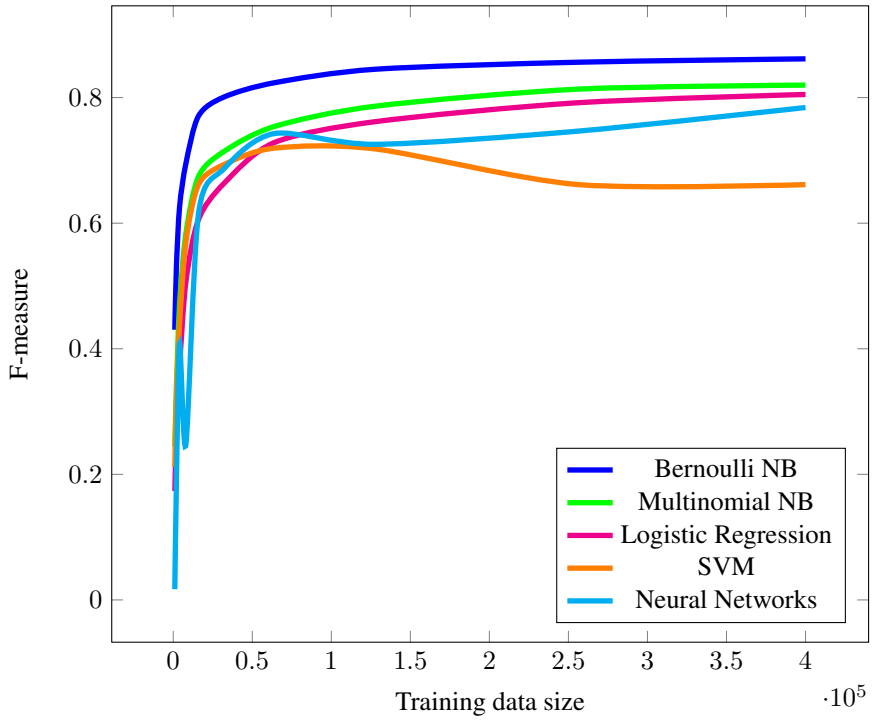

Figure 5: Results of each classifier using tf-idf feature extractor

Classifier	
Feature extractor	tf-idf
Feature selection limit	22000
Model	Bernoulli NB
Results	
True Positive	35088
False Positive	6636
True Negative	453650
False Negative	4626
Sensitivity	0.883
Specificity	0.986
F-Measure	0.862

Table 1: Detailed results for the best classifier for full data size

5 Results and Discussion

As had been shown the use of tf-idf tokenizer had given the best set of features for this data set. This feature set was further pruned using chi-squared-based feature selection. The results in Figure 5 show that the Bernoulli Naives Bayes implementation gives the highest success. The specific results from this classifier are shown in Table 1.

This set of results together show the significance of the subreddit as a key predictor of a post's NSFW/SFW class. This is re-inforced by the evidence that bigrams do not offer improved performance, which suggests that the single words of the subreddit name are more important than word

Post Titles
My friends friend had an accident
Totally legit
Unimpressed Kitchen
So I googled reddit and down votes and this came up.
I don't think they are the reason why it sank ..
Somebody left this book at work

Table 2: Examples of false negative posts with ambiguous titles

combinations inside the title. This was also proposed by the large number of features that could be removed through feature selection. Interestingly a simpler Naive Bayes model gives better results than SVMs and neural networks which suggest that in this data-set these more complex models suffer from over-fitting.

5.1 The Challenging Subset

The results from the various classifier highlight that there remains a subset of posts that are difficult to classify. A cross-analysis of these posts shows several reasons for their challenging classification.

An initial look at the post titles highlight that many are linguistically ambiguous and would be very difficult for a human to identify the possible content. Some examples of the titles are shown in Table 5.1. Furthermore an analysis of the subreddit and author of the posts highlight the difficulty in using these metrics. A large proportion of these posts are from deleted users which causes the author name to be "[Deleted]". This author name has over 9000 posts attributed to it, 22% of which are tagged NSFW. Because of the large variability and high proportion of this author name, the author features become significantly less predicitve of NSFW posts. Many posts are also from subreddits with no clearly defined bias towards SFW or NSFW which makes the subreddit a more limiting predictor in these cases.

It may be possible to identify these challenging posts through the same metrics and flag them for further analysis. Then a deeper analysis of the linked URL or linked images could be done for a better classification. This would be interesting area of further research to improve the quality of classification for this challenging subset of data.

5.2 Title Only Classification

In order that this classification system could be used on a more diverse data set than just Reddit posts, we tested whether only using the title of the Reddit post would be sufficient to successfully classify posts. The same method of feature detection was used on the title only. This caused the F-measure to drop to 0.42.

After analysis of the failing posts, it can be shown that the problem with the challenging subset has been enlarged. The text in the title can contain very ambiguous language which causes the much lower success rate of the classifier. With the given success rate, the classifier could not be used a reliable metric for tagging posts. Furthermore the author and subreddit fields are certainly very important to the success of the classifer.

6 Conclusion

This paper introduces an effective classifier for NSFW posts on Reddit. It tests the various features that may be used and showed that the most effective results were gained from using a Bernouilli Naives Bayes classifier with a tf-idf feature extractor. While the sensitivity and specificity is not high enough to surpass human intervention, the classifer could be used as an excellent complement to suggest tags for posts.

References

[1] Brenda S Baker and Eric Grosse. Local control over filtered www access. In *Proceedings of the 4th World Wide Web Conference*, 1995.

[2] Thomas Deselaers, Lexi Pimenidis, and Hermann Ney. Bag-of-visual-words models for adult image classification and filtering. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[3] Rongbo Du, Reihaneh Safavi-Naini, and Willy Susilo. Web filtering using text classification. In *Networks, 2003. ICON2003. The 11th IEEE International Conference on*, pages 325–330. IEEE, 2003.

[4] George Forman. Feature selection for text classification. *Computational methods of feature selection*, pages 257–276, 2008.

[5] Wordle generating tool. Net Nanny: content-control software, http://www.wordle.com, 2013.

[6] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[7] Paul Greenfield, Peter Rickwood, Huu Cuong Tran, and Australian Broadcasting Authority. *Effectiveness of Internet filtering software products*. Australian Broadcasting Authority, 2001.

[8] ContentWatch Inc. Net Nanny: content-control software, http://www.netnanny.com, 2013.

[9] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, DTIC Document, 1996.

[10] J Langford, L Li, and A Strehl. Vowpal wabbit online learning project, 2007.

[11] Pui Y Lee, Siu C Hui, and Alvis Cheuk M Fong. Neural networks for web content filtering. *Intelligent Systems, IEEE*, 17(5):48–57, 2002.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] Jordan Segall and Alex Zamoshchin. Twitter sentiment classification using distant supervision. *CS229 Project Report, Stanford*.

[14] Troy Steinbauer. Information and social analysis of reddit. *CS224N Project Report, Stanford*.

[15] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1113–1120. ACM, 2009.