SVM Approach to Forum and Comment Moderation

Adam Maus

University of Wisconsin – Madison, Computer Science Department
2120 University Ave. Madison WI, 53726
amaus@wisc.edu

Abstract

Social networks, blogs, and forums bring users together to build a community usually based on a system of communication through comments. Based on the degree of complexity, these systems may or may not have a moderator whose primary purpose is to remove spam or abusive comments within these systems. When a moderator is used, it is most often a human, whose time and energy must be exerted to read each and every comment making it a tedious job for a large website. A support vector machine (SVM) approach is proposed for comment moderation. Using a training corpus obtained from the popular website Youtube.com, a support vector machine is used to classify comments as abusive or not. Baseline accuracy is found by performing 10-fold cross validation on unprocessed data. Different experiments are performed on the data by preprocessing it to find if certain variations provide a more accurate estimate.

Background

The rise of social interaction with websites has dramatically increased in the past few years. Interaction with these websites usually involves entering comments about a posting on the website. The comments can take the form of any written text whether that is in English or otherwise. In many cases, the commenting system is an important aspect of creating a community on the website. This system also usually involves the use of anonymous posting which gives users the ability to abuse the poster or others on the system without reprimand.

Abuse on a commenting system ranges from spam to comments that are inappropriate. Users may find this content extremely offensive and the website can get a bad reputation because of a few users. Therefore the moderator has an important task in protecting the integrity of a website. They enforce certain rules about what can and cannot be posted. For example, an abusive comment could attack a user using derogatory terms, the moderator would then decide if this comment should even be posted. Usually moderators are human readers who must read through each of the comments to classify them. For many large websites such as Youtube.com, human moderators are simply not possible because users are constantly leaving comments and the task of reading them all is enormous. For websites such as this, ratings are often used as a form of moderation.

Some commenting systems use these ratings to allow users to do moderation of inappropriate comments but often the system can easily be bypassed by users or a group that rate inappropriate comments well. This is a problem for a website such as Youtube.com because many inappropriate comments are not flagged or given a poor rating. Likewise, appropriate comments are in some cases given a poor rating without any reason. For a large website, a feasible solution would be to use a classification algorithm to moderate comments [1].

Support Vector Machines

A classifier can be used to moderate a website and operates faster than having human moderators or users flag comments. A support vector machine (SVM) was chosen for the classification of abusive comments. SVM is used to construct a hyperplane in *n*-dimensional space and can divide the space inhabited by the vector representation of the comments. The hyperplane is created by solving a convex programming problem and different hyperparameters are used to control characteristics of how the model is created. These hyperparameters can be optimized to find the highest accuracy for a given training corpus [2].

The LIBSVM toolbox [3] was selected for this project to create the SVM and a tuning algorithm was used to find optimal hyperparameters of the model. A radial basis function kernel was used to represent the hyperplane,

$$e^{-\gamma|u-v|^2}$$

In a SVM with a radial basis function kernel, there is an adjustable γ, an ε that serves as the termination criterion [4], and an adjustable C parameter that represents the cost for a misclassified vector.

Instead of a normal grid search [5], a stochastic algorithm was used to tune the parameters. The SVM is first trained on the training corpus using default parameters specified by the LIBSVM toolbox. $\gamma = 0.1$, $\varepsilon = 0.1$, and $C = 0.1$. The accuracy of the model with these parameters was found using 10-fold cross-validation. After obtaining an initial accuracy the parameters were altered slightly by taking the highest accuracy model's parameters and adding a Gaussian random number to them. The σ for the Gaussian distribution was initialized to be 1.0, after each iteration the accuracy of the new model was found, if the modified model had a higher accuracy than the best model so far, σ was increased by 2. If worse, it was decreased by 0.95. This serves as a shrinking neighborhood around the best parameters, if the tuning is going well, the neighborhood stays large so there is a larger area of parameter space to explore. The parameters are tuned using this procedure for 300 iterations. This stochastic procedure was repeated several times and the highest accuracy model was taken.

Data Collection

The corpus studied was manually collected and classified from Youtube.com. 2665 English comments were collected, 1451 were classified as positive or neutral comments. The other 1214 comments were considered to be abusive or spam. The majority of the comments were over 5 words long; words in this case are strings of text that were separated by a space. These comments were part of an experiment to test the consistency of the classifier on longer length comments.

For preprocessed dataset experiments, each comment was separated by symbols where each symbol became a word. In another experiment, references to proper nouns were removed. The idea behind this was that it should be easier to generalize the comments if proper noun references were removed. An example of this would be, "I hate Monday." or "I hate Mick." Which became "I hate <proper>."

The preprocessed data was then processed into a bag of word vector that counted the number of times a word appeared in a particular comment. It was also constructed so that all words were lowercase to remove inconsistencies in capitalization.

Experimental Results

To experiment on the corpus, 1000 positive and negative comments were chosen to be training data for the SVM model and held consistent when performing experimental preprocessing on this data. A different dataset was randomly created for the 5+ words experiments. The different datasets were used as input for the SVM model and the parameters tuned using the procedure described above. Table 1 shows the accuracy of the best model for each experimental preprocessing of the dataset.

Input Data	Accuracy
Baseline (1000 of each)	85.75
Split Data	85.45
Split + Proper Nouns Removed	85.85
Baseline (1000 of each) with 5+ words	84.45
Split Data with 5+ words	86.95
Split + Proper Nouns Removed with 5+ words	85.80
Split Bag of Bigrams (1000 of each)	78.90

Table 1: Experimental Results using SVM

Conclusions and Future Work

Classification using the SVM was at best 86.95% accurate using 10-fold cross-validation on the training corpus with data preprocessing. However, this is not significantly better than the baseline accuracy of 85.75%. Even though different experiments were performed on this dataset, it did not significantly raise the accuracy from the baseline. Using the method on comments taken from Youtube.com proves to be accurate but not accurate enough to use in a system because of the relatively high error rate. This would mean that between one and two comments are misclassified for every ten.

It is difficult to ascertain where difficulty in classification may lie. More experiments on data taken from the training corpus could be done to find the average accuracy of the model on the data using the trained parameters. Different kernels could also be used; a linear or polynomial kernel may provide a better estimation. These were only tested on the last day but baseline accuracies of 85.75% and 85.90% were achieved on a linear kernel and polynomial with degree 2, respectively.

The training corpus was not rigorously examined data by a group of people that classify a comment as abusive or not. Therefore objectivity of one person does not necessarily represent the feelings of a group. Along those lines, this problem is fuzzy in the sense that abuse in a comment can range from a sentence that is directly derogatory towards the author to one that is rather indirect and subtly abusive. A better model may be to use support vector regression which can take features and classify it into a real number of how abusive a comment is.

The SVM model was shown to be fairly accurate and it would be interesting to see if it is possible to significantly increase the accuracy using different text preprocessing and/or models.

References

1. Arnt, A.; and Zilberstein, S. 2003. Learning to Perform Moderation in Online Forums. In Proceedings of the IEEE/WIC International Conference on Web Intelligence, 637-641.
2. Ahn, H.; Lee, K.; and Kim, K. 2006. Global Optimization of Support Vector Machines Using Genetic Algorithms for Bankruptcy Prediction. In *Neural Information Processing*, ed. I. King, 420-429. Berlin: Springer.
3. Chang, C.; and Lin, C. 2001. LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
4. Hearst, M. A. 1998. Support Vector Machines. *IEEE Intelligent Systems*, 13(4): 18-28.
5. Hsu, C.; Chang, C.; and Lin, C. 2003. A practical guide to support vector classification, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei.