Classifying Racist Texts Using A Support Vector Machine

Edel Greevy
Princip Project, SALIS
Dublin City University
Dublin 9, Ireland
edel.greevy@dcu.ie

Alan F. Smeaton
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
asmeaton@computing.dcu.ie

ABSTRACT

In this poster we present an overview of the techniques we used to develop and evaluate a text categorisation system for the PRINCIP project which sets out to automatically classify racist texts. Support Vector Machines (SVM) are used to automatically categorise web pages based on whether or not they are racist. Different interpretations of what constitutes a term are taken, and in this poster we look at a bag of words (BOW) vs. a bigram representation of a web page within a SVM.

Keywords

Text Categorisation/Classification, Machine Learning, Support Vector Machines.

1. INTRODUCTION

PRINCIP is primarily a linguistics-based project which aims to build a classification system for racist pages on the web through the corpus-based analysis of racist content. Linguistic patterns identified during analysis of web pages can be formulated into rules and used in a categorisation system to allow for detection of illicit content on the web. Text Categorisation (TC) is concerned with the automatic assignment of documents to predefined categories. Modern TC borrows and applies many techniques from two established fields of research: Information Retrieval and Machine Learning.

Current methods of filtering racism rely heavily on either keywords or the manual labelling of offensive material by Label Bureaus. In order to implement successful filtering systems, a considerable human effort is required, not only in the initial stages of filter construction but also in an ongoing basis as the targets of racism change, as language evolves, existing websites are edited or new websites are added. Automatic text categorisation techniques are reported to have been successful when applied to other domains such as news story categorisation and such methods led to vast improvements in productivity as well as savings in terms of time and manpower. Given the fluidity of racism on the web, this is one area that may benefit from the application of automatic text categorisation techniques.

In our work we tackle the issue of racist texts in the PRINCIP project by building an automatic classification system using Support Vector Machines to allow for the detection of racism automatically.

2. DETECTING RACIST TEXTS

Detecting racism on the Internet is not just a topic-based problem as in news story classification, rather it is more similar to genre detection as described in [1], in that we are not really concerned with the topic itself but we are trying to identify features that will discern an author's attitude in relation to the topic, something which is orthogonal to the actual topic. In their work on genre detection, Finn *et al.* [1] found the distribution of parts of speech (POS) to outperform the bag of words (BOW) approach for genre detection. Our own experiments in PRINCIP revealed there to be differences in some lexical, collocation and POS distributions across racist and non-racist documents [2]. Based on this analysis of our domain of racist texts, we have decided to compare various feature representations, looking at bag-of words and bi-grams of words patterns in our wok on detecting racist texts. An SVM will be trained on each representation in order to identify the most productive method and representation for detecting racism.

3. SUPPORT VECTOR MACHINES

We use Support Vector Machines to learn the features of the training sets and classify new unseen documents. SVMs are a very powerful learning method that "since its introduction has already outperformed most other systems in a wide variety of applications" [3]. SVMs overcome many of the problems associated with efficiency of training such as overfitting. They are capable of generalising well in high dimensional spaces such as ours where there is a rich representation of words, bi-grams etc., meaning solutions can always be found efficiently even for training sets with many thousands of examples. The compact representation of the hypothesis being learned (in our case the categorisation of documents) means that evaluation on unseen input is very fast thereby making it efficient when it comes to testing.

Given an input document *d*, in order to arrive at output class *c* the SVM has to learn the relationship between the input and output pairings. The function that does so is known as the *target function*. This enables the machine to make a decision about the *target* class of the unseen documents.

The input/output pairings are represented by a vector $x_i \varepsilon R^n$, $i = 1,\ldots, n$ and the associated class y_i where y_i is 1 if a document d belongs to category c and -1 otherwise. The task of the machine is to choose the mapping $x_i \dashrightarrow y_i$ that minimises the risk of error.

SIGIR'04, July 25–29, 2004, Sheffield, UK.

4. RESULTS

4.1 About the dataset

During the PRINCIP project a corpus of 3 millions words was collected. For this study, the corpus was split into datasets of varying sizes with an equal number of racist and non-racist documents in each set.

Table 1. The size of the datasets

	Set 1	Set 2	Set 3	Set 4
No. Docs in Training Set	200	400	600	800
No. Docs in Test Set	60	100	150	200

Bag of Words

Figure 1 illustrates how recall improves as the training set increases. A steady increase was reported with the precision/recall figures for the final dataset achieving 92.55%/87.00%, a considerable improvement on precision/recall for set 1.

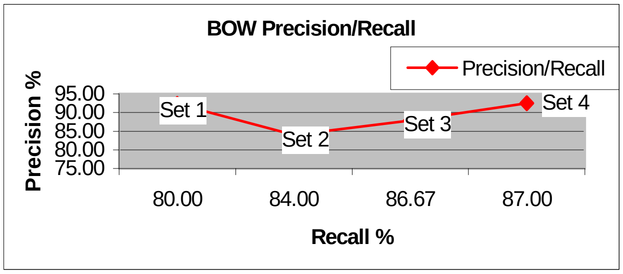

Figure 1. Precision and recall figures for the BOW representation using the linear kernel function

Bi-grams

From figure 2 you will see that for each dataset precision improved dramatically compared to the BOW representation whereas the recall figures dropped by between 10-15%. As the training set is increased precision drops slightly while recall improves reaching 75% in experiment set 3.

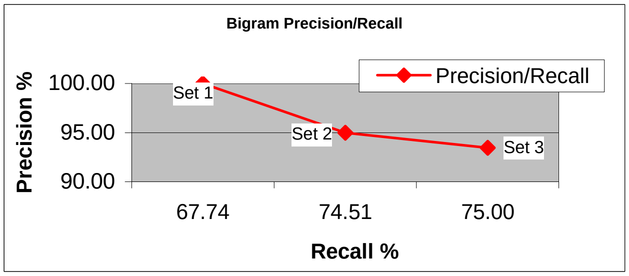

Figure 2. Precision and recall figures for the bigrams

4.4 Experimenting with Kernel Functions

By conducting numerous experiments using linear, polynomial, radial basis function and sigmoid tanh as kernel functions, we learned that the polynomial kernel function improved precision and recall on set 4 for the BOW representation from 92.55%/87.00% to 92.78% precision and 90% recall.

The sigmoid tanh kernel function produces the best precision and recall figures for the bigram representation (see figure 2) improving on the linear function by between 1-4%.

4.5 Summary

The accuracy on the test set for the BOW representation outperformed bigrams: for set 3 for the accuracy on the test set for the BOW was 87.33% while bigrams gave an accuracy of 84.77%.

Both the bag of words approach and the bigrams have their advantages with BOW resulting in high recall and bigrams giving high precision. It would be interesting to see if adjusting the cost-factor during the training of the bigram representation, has any influence on recall.

Though this may be computationally expensive, it would also be interesting to see what affect the BOW and bigrams together as a representation, would have on precision and recall.

5. CONCLUSION

We have shown that it is possible to build an automatic classification system for the detection of racism on the web.

Our future research involves training Support Vector Machines for tri-gram word sequences and part of speech tags - so as to identify the most effective method that will allow for the classification of racist documents on the web.

6. REFERENCES

[1] Finn A., Kushmerick N. and Smyth B. (2002). Genre classification and domain transfer for information filtering. In *Proceedings of the European Colloquium on Information Retrieval Research (Glasgow).*

[2] Lechleiter H. and Greevy E. (2003). The Language of Open Racism: A Corpus Linguistic Analysis. Societas Linguistica Europaea Conference. Lyon, France, September 4th 2003.

[3] Cristianini N. and Shawe-Taylor J. (2000). *Support Vector Machines and other kernel-based learning methods.* Cambridge University Press.

.