The Importance of Precision in Humour Classification

Joana Costa[1], Catarina Silva[1,2], Mário Antunes[1,3], and Bernardete Ribeiro[2]

[1] Computer Science Communication and Research Centre
School of Technology and Management, Polytechnic Institute of Leiria; Portugal
{joana.costa,catarina,mario.antunes}@ipleiria.pt
[2] Department of Informatics Engineering, Center for Informatics and Systems of the University of Coimbra (CISUC); Portugal
{catarina,bribeiro}@dei.uc.pt
[3] Center for Research in Advanced Computing Systems (CRACS), Portugal

Abstract. Humour classification is one of the most interesting and difficult tasks in text classification. Humour is subjective by nature, yet humans are able to promptly define their preferences.
Nowadays people often search for humour as a relaxing proxy to overcome stressful and demanding situations, having little or no time to search contents for such activities. Hence, we propose to aid the definition of personal models that allow the user to access humour with more confidence on the precision of his preferences.
In this paper we focus on a Support Vector Machine (SVM) active learning strategy that uses specific most informative examples to improve baseline performance. Experiments were carried out using the widely available Jester jokes dataset, with encouraging results on the proposed framework.

Keywords: Support Vector Machine, Active Learning, Text Classification, Humour classification

1 Introduction

Humour classification is one of the most interesting and difficult tasks in text classification. However, despite the attention it has received in fields such as philosophy, linguistics, and psychology, there have been few attempts to create computational models for humour classification [1].

Modern societies turn human course of life a *fast forward* version of itself. It is not only overwhelming work times, but also the pressure they convey. Most people feel that there is not enough time to de-stress, and even when there is, the mind is constantly being overstimulated by the mass media, that take part of everyday life and expose people to so much information.

While it is merely considered a way to induce amusement, humour also has a positive effect on the mental state of those using it and has the ability to improve their activity [1, 2].

With these constraints in mind, we propose a framework to aid the definition of personal models that allow the user to access humour with more confidence on the precision of his preferences. When searching for amusing content with little or no time, the user is more interested in spending a nice time than grasping everything that would be possible. In other words, in a computer science point of view, the precision of the displayed content is more relevant than its recall performance.

Active learning designs and analyses learning algorithms that can effectively filter or choose the samples to be labeled by a supervisor (a.k.a. oracle or teacher). The reason for using active learning is mainly to expedite the learning process and reduce the labeling efforts required by the teacher [3]. Another strong reason is the possibility for each user to define personal labels, thus constructing a customized learning model that better fits his preferences.

The SVM active learning framework we propose is a certainty-based method using the definition of the specific most informative examples to improve baseline performance, with two major guidelines: (i) the number of active examples has to be necessarily small; and (ii) precision is a critical factor.

The rest of the paper is organized as follows. We start in Section 2 by describing the background on SVM, active learning and humour classification and proceed into Section 3 by presenting the active learning framework for humour classification. Then, in Section 4 we introduce the Jester benchmark and discuss the results obtained. Finally, in Section 5 we delineate some conclusions and present some directions for future work.

2 Background

In what follows we will provide the background on Support Vector Machine (SVM), active learning and humour classification, which constitute the generic knowledge for understanding the approach proposed ahead in this paper.

2.1 Support Vector Machines

SVM is a machine learning method introduced by Vapnik [4], based on his Statistical learning Theory and Structural Risk Minimization Principle. The undelying idea behind the use of SVM for classification, consists on finding the optimal separating hyperplane between the positive and negative examples. The optimal hyperplane is defined as the one giving the maximum margin between the training examples that are closest to it. Support vectors are the examples that lie closest to the separating hyperplane. Once this hyperplane is found, new examples can be classified simply by determining on which side of the hyperplane they are.

The output of a linear SVM is $u = \mathbf{w} \times \mathbf{x} - b$, where $\mathbf{w}$ is the normal weight vector to the hyperplane and $\mathbf{x}$ is the input vector. Maximizing the margin can

be seen as an optimization problem:

$$\begin{aligned} &minimize \quad \frac{1}{2}||\mathbf{w}||^2, \\ &subjected\ \ to\ \ y_i(\mathbf{w}.\mathbf{x}+b) \geq 1, \forall i, \end{aligned} \tag{1}$$

where $\mathbf{x}$ is the training example and y_i is the correct output for the ith training example. Intuitively the classifier with the largest margin will give low expected risk, and hence better generalization.

To deal with the constrained optimization problem in (1) Lagrange multipliers $\alpha_i \geq 0$ and the Lagrangian (2) can be introduced:

$$L_p \equiv \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{l} \alpha_i(y_i(\mathbf{w}.\mathbf{x}+b)-1). \tag{2}$$

In fact, SVM constitute currently the best of breed kernel-based technique, exhibiting state-of-the-art performance in diverse application areas, such as text classification [5–7]. In humour classification we can also find the use of SVM to classify data sets [1, 8].

2.2 Active Learning

The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns. An active learner may pose queries, usually in the form of unlabeled data instances to be labeled by an oracle [9].

Active learning methods can be grouped according to the selection strategy: committee-based and certainty-based [10]. The first group determines the active examples combining the outputs of a set of committee members. As in [11], most effort is done in determining the examples in which the members disagree the most as examples to be labeled. The certainty-based methods try to determine the most uncertain examples and point them as active examples to be labeled. The certainty measure depends on the learning method used.

2.3 Humour classification

Humour research in computer science has two main research areas: humour generation [2, 12] and humour recognition [1, 8, 13]. With respect to the latter, research done so far considers mostly humour in short sentences, like *one-liners*, that is jokes with only one line sentence, and the improvement of interaction between applications and users.

Humour classification is intrinsically subjective. Each one of us has its own perception of fun, hence automatic humour recognition is a difficult learning task that is gaining interest among the scientific community.

Classification methods used thus far are mainly text-based and include SVM classifiers, *naïve Bayes* and less commonly decision trees.

In [8] a humour recognition approach based in *one-liners* is presented. A dataset was built grabbing *one-liners* from many websites with an algorithm and the help of web search engines. This humorous dataset was then compared with non-humorous datasets like headlines from news articles published in the Reuters newswire and a collection of proverbs.

Another interesting approach [13] proposes to distinguish between an implicit funny comment and a not funny one. A 600,000 web comments dataset was used, retrieved from the Slashdot news Web site. These web comments were tagged by users in four categories: funny, informative, insightful, and negative, which split the dataset in humorous and non-humorous comments.

3 Proposed approach

This section describes the proposed SVM active learning strategy. The SVM active learning framework we propose is a certainty-based method, i.e. it determines the most uncertain examples and point them as active examples to be labeled. As the certainty measure depends on the learning method used, for SVM we used the margin as the determining factor. When an SVM model classifies new unlabeled examples, they are classified according to which side of the Optimal Separating Hyperplane (OSH) they fall. As can be gleaned from Fig. 1, not all unlabeled points are classified with the same distance to the OSH. In fact, the farther from the OSH they lie, i.e. the larger the margin, more confidence can be put on their classification, since slight deviations of the OSH would not change their given class.

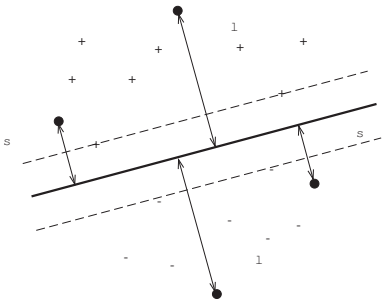

Fig. 1: Unlabeled examples (black dots) with small and large margins.

Our active learning approach includes a certain number of unlabeled examples from the testing set (only the features, not the classification) in which the SVM has less confidence (smaller margin, see Fig. 1) after they are correctly classified by the supervisor. Thus, an example $(\mathbf{x}_i, y_i)$ will be included if Equation (3) holds.

$$(\mathbf{x}_i, y_i) : \rho(\mathbf{x}_i, y_i) = \frac{2}{\|w\|} < \Delta \tag{3}$$

This number of examples can not be large, since the supervisor will be asked to manually classify them. After being correctly classified, they are integrated in the training set. This approach can be regarded as a form of active learning, where the information introduced by each example in the classification task is inversely proportional to its classification margin.

Despite not being fully automated, the active learning method has the potential to efficiently improve classification performance, since a user must classify the margin-based chosen examples. These examples help customize the learning machine regarding personal classification preferences.

4 Experimental Setup

4.1 Data set

The Jester dataset contains 4.1 million continuous ratings (-10.00 to +10.00) of 100 jokes from 73,421 users and is available at: http://eigentaste.berkeley.edu. It was generated from Ken Goldberg's joke recommendation website, where users rate a core set of 10 jokes and receive recommendations from other jokes they could also like. As users can continue reading and rating and many of them end up rating all the 100 jokes, the dataset is quite dense. The dataset is provided in three parts: the first one contains data from 24,983 users who have rated 36 or more jokes, the second one data from 23,500 users who have rated 36 or more jokes and the third one contains data from 24,938 users who have rated between 15 and 35 jokes. The experiments were carried out using the first part as it contains a significant number of users and rates for testing purposes, and for classification purposes was considered that a joke classified on average above 0.00 is a recommendable joke, and a joke classified below that value is non recommendable. The jokes were split into two equal and disjoint sets: training and test. The data from the training set is used to select learning models, and the data from the testing set to evaluate performance.

4.2 Pre-processing methods

A joke is represented as the most common, simple and successful document representation, which is the vector space model, also known as *Bag of Words*. Each joke is indexed with the *bag* of the terms occurring in it, i.e., a vector with one component for each term occurring in the whole collection, having a value that takes into account the number of times the term occurred in the joke. It was also considered the simplest approach in the definition of term, as it was defined as any space-separated word. Considering the proposed approach and the use of text-classification methods, pre-processing methods were applied in order to reduce feature space. These techniques, as the name reveals, reduce the size of the joke representation and prevent the mislead classification as some words, such as articles, prepositions and conjuctions, called *stopwords*, are non-informative words, and occur more frequently than informative ones. These words could also

mislead correlations between jokes, so *stopword* removal technique was applied. *Stemming* method was also applied. This method consists in removing case and inflection information of a word, reducing it to the word stem. Steaming does not alter significantly the information included, but it does avoid feature expansion.

4.3 Performance metrics

In order to evaluate a binary decision task we first define a contingency matrix representing the possible outcomes of the classification, as shown in Table 1.

	Class Positive	Class Negative
Assigned Positive	a	b
	(True Positives)	(False Positives)
Assigned Negative	c	d
	(False Negatives)	(True Negatives)

Table 1: Contingency table for binary classification.

Several measures have been defined based on this contingency table, such as, error rate ($\frac{b+c}{a+b+c+d}$), recall ($R = \frac{a}{a+c}$), and precision ($P = \frac{a}{a+b}$), as well as combined measures, such as, the van Rijsbergen F_β measure [14], which combines recall and precision in a single score:

$$F_\beta = \frac{(\beta^2 + 1)P \times R}{\beta^2 P + R}. \tag{4}$$

F_β is one of the best suited measures for text classification used with $\beta = 1$, i.e. F_1, an harmonic average between precision and recall (5).

$$F_1 = \frac{2 \times P \times R}{P + R}. \tag{5}$$

4.4 Results and discussion

To test the proposed approach an experimental setup with three different experiments was defined:

1. Baseline SVM
2. Active Learning SVM with random active examples (Random AL SVM)
3. Active Learning SVM with margin-based active examples (Margin AL SVM)

Keeping in mind our initial guidelines: (i) the number of active examples has to be necessarily small; and (ii) precision is a critical factor, we defined a set of only 10 active examples, following the initial dataset construction procedure (see Section 4.1).

In the first experiment the SVMLight[1] package was used with linear kernels and default parameters. For the second experiment 30 runs were carried out, by randomly selecting 10 active examples and average values are presented. For the third experiment, the proposed SVM margin-based active learning strategy was deployed (see Section 3). Table 2 summarizes the performance results obtained.

	TP	FP	TN	FN	Precision	Recall	F1
Baseline SVM	35	8	4	3	81.40%	92.11%	86.42%
Random AL SVM	32	6	6	6	84.36%	84.74%	83.81%
Margin AL SVM	36	5	7	2	87.80%	94.74%	91.14%

Table 2: Performances of Baseline and Active Learning Approaches.

Focusing on precision values, we can see that there is a trend for improvement: 81.40%, 84.36% and 87.80%. This can become a determining factor in humour classification, since users are typically more interested in a strong confidence of amusement (low false positive values) than in the guarantee of getting all jokes (low false negative values).

Comparing both active learning strategies, we can see that although both present improvements in precision, the random approach achieves it at the expense of recall values, while the proposed margin-based active learning permits the improvement of both recall and precision.

5 Conclusions and Future Work

In this paper we have described a framework for humour classification, based on an SVM active learning strategy. Our aim was to evaluate the use of such a strategy to increase the overall humour classification precision. For that purpose we have conducted a set of experiments with the Jester benchmark data set, by comparing the baseline SVM model with a two-fold active learning approach: (i) using a set of arbitrary examples; and (ii) using a set of the most relevant examples.

The preliminary results obtained are very promising. We were able to observe that the proposed active learning strategies have increased the overall precision measure, i.e. have reduced the false positive examples, when compared with the baseline SVM classification.

Regarding the recall, we have also observed that only in the active learning approach with the most relevant examples we were able to maintain an appropriate false negative rate, hence not worsening the overall classification results (e.g. $F1$). That is, for the specific case of joke classification with the Jester data

[1] http://svmlight.joachims.org/

set, using an active learning approach, we increase the amount of jokes correctly classified as having fun and thus *recommended* for reading. Such an active learning approach may also benefit other different application domains, like the recommendation systems for books or movies.

Our research is now focused on introducing crowdsourcing information into the active learning processing. That is, instead of using the results obtained from a supervisor we intend to use knowledge acquired from the end-users volunteer participation.

References

1. R. Mihalcea and C. Strapparava, "Making computers laugh: investigations in automatic humor recognition," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 531–538. 1, 3
2. O. Stock and C. Strapparava, "Getting serious about the development of computational humor," in *IJCAI'03*, 2003, pp. 59–64. 1, 3
3. Y. Baram, R. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," in *Proceedings of ICML-2003, 20th International Conference on Machine Learning*, 2003, pp. 19–26. 2
4. V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1999. 2
5. T. Joachims, *Learning Text Classifiers with Support Vector Machines*. Kluwer Academic Publishers, Dordrecht, NL, 2002. 3
6. S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002. 3
7. M. Antunes, C. Silva, B. Ribeiro, and M. Correia, "A Hybrid AIS-SVM Ensemble Approach for Text Classification," *Adaptive and Natural Computing Algorithms*, pp. 342–352, 2011. 3
8. R. Mihalcea and C. Strapparava, "Technologies That Make You Smile: Adding Humor to Text-Based Applications," *Intelligent Systems, IEEE*, vol. 21, no. 5, pp. 33–39, 2006. 3, 4
9. B. Settles, "Active learning literature survey," CS Technical Report 1648, University of Wisconsin-Madison, 2010. 3
10. C. Silva and B. Ribeiro, "On text-based mining with active learning and background knowledge using svm," *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, vol. 11, no. 6, pp. 519–530, 2007. 3
11. A. K. McCallum and K. Nigam, "Employing EM and pool-based active learning for text classification," in *Proceedings of ICML-98, 15th International Conference on Machine Learning*. Morgan Kaufmann Publishers, San Francisco, US, 1998, pp. 350–358. 3
12. K. Binsted and G. Ritchie, "An implemented model of punning riddles," *arXiv.org*, vol. cmp-lg, jun 1994. 3
13. A. Reyes, M. Potthast, P. Rosso, and B. Stein, "Evaluating Humor Features on Web Comments," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, may 2010. 3, 4
14. C. van Rijsbergen, *Information Retrieval*. Butterworths Ed., 1979. 6