

Information Gain Feature Selection for Ordinal Text Classification using Probability Re-distribution

Rahman Mukras, Nirmalie Wiratunga, Robert Lothian, Sutanu Chakraborti,
and David Harper

School of Computing
The Robert Gordon University
St Andrew Street
Aberdeen UK, AB25 1HG
{ram,nw,rml,sc,djh}@comp.rgu.ac.uk

Abstract. This paper looks at feature selection for ordinal text classification. Typical applications are sentiment and opinion classification, where classes have relationships based on an ordinal scale. We show that standard feature selection using Information Gain (IG) fails to identify discriminatory features, particularly when they are distributed over multiple ordinal classes. This is because inter-class similarity, implicit in the ordinal scale, is not exploited during feature selection. The Probability Re-distribution Procedure (PRP), introduced in this paper, explicates inter-class similarity by revising feature distributions. It aims to influence feature selection by improving the ranking of features that are distributed over similar classes, relative to those distributed over dissimilar classes. Evaluations on three datasets illustrate that the PRP helps select features that result in significant improvements on classifier performance. Future work will focus on automated acquisition of inter-class similarity knowledge, with the aim of generalising the PRP for a wider class of problems.

1 Introduction

Comparative experimental studies have consistently shown Information Gain [4] based feature selection to result in good classifier performance [18, 6, 9]. This performance success and IG's well found principles in Information Theory [17], both make it a popular feature selection algorithm for text classification.

The idea behind IG is to select features that reveal the most information about the classes. Ideally, such features are highly discriminative and occur in a single class. However, in text classification it is often the case that features would occur in more than one class. This consequently forces IG to select the best features amongst those that occur in multiple classes. In this situation, IG would be sufficient to select such *multi-class features*, provided that all pairs of classes are equally similar. However, if some pairs of classes are more similar than others (for example, when there is an ordinal relationship between them),

then it matters which classes a feature is distributed across. In this case, a feature that occurs in two similar classes is relatively more discriminative than one that occurs in two dissimilar ones. As a result of this, a feature selection algorithm that does not address inter-class similarities (in particular IG) would be inadequate for selecting such multi-class features.

In this paper we address the problem of selecting features that occur in multiple classes for the domain of ordinal text classification, which is known to have acute inter-class similarities [14, 3]. Ordinal text classification involves classifying a document into an ordinal scale consisting of three or more classes. This type of problem is characterised by classes that possess a similarity that decays with the ordinal distance between them. For instance, a textual movie review accompanied by a rating of 1 (on a 10 point scale) is expected to be more similar to one rated at 2 than another at 10. Consequently, in this domain, a direct correlation can be drawn between inter-class distance and inter-class similarities. In our solution, we capitalise on this fact by using the distance between classes as a metric of the similarities between them. Our approach is termed as the Probability Re-distribution Procedure. It identifies a feature that occurs in similar classes, and ‘favours’ it over another that occurs in less similar ones. This is achieved by revising the features distribution across the classes, so that it yields a relatively higher IG score.

This paper proceeds as follows: Section 2 introduces a motivating example after which the PRP is discussed in Section 3. The evaluation of the PRP on three datasets is then presented in Section 4. Discussions on future work appear in Section 5 followed by related work in Section 6. Finally the conclusion is presented in Section 7.

2 Information Gain Feature Selection for Ordinal Classes

In an ordinal problem, class ordering is directly related to inter-class similarities and hence is an important factor to consider in selecting a discriminative feature. To illustrate this using an example, assume that we have two features, f_1 and f_2 , that are distributed across a set of equally sized ordinal classes with the proportions shown in Fig. 1. If one were to disregard class ordering, then these

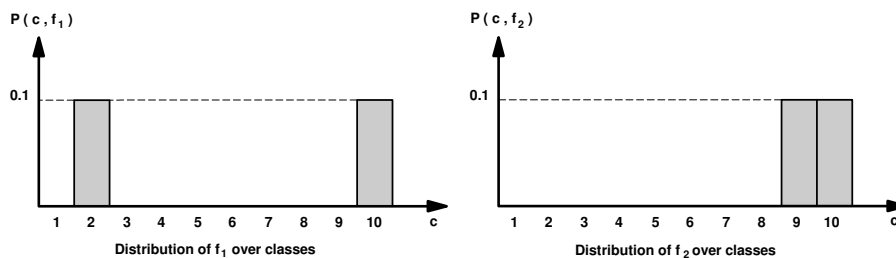


Fig. 1. An example of two features distributions that would obtain the same IG score.

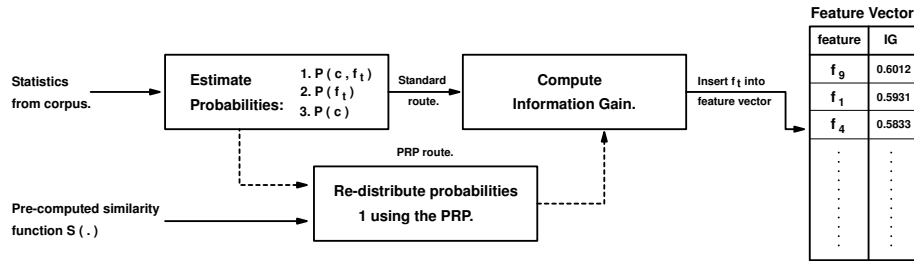


Fig. 2. The Probability Re-distribution Procedure for a given feature f_t .

two features would be identical. However, with the inclusion of class ordering, they are different. Firstly, feature f_1 occurs in two distant, hence dissimilar, classes and this lowers its discriminative ability. On the other hand, feature f_2 occurs in two neighbouring, hence similar, classes and this makes it relatively more discriminative than feature f_1 .

The problem of using IG in ordinal classes is that the joint distribution, $P(c, f_t)$ (between the random variable over the classes c , and the feature in question f_t), it employs carries no explicit information about class ordering. This forces IG to treat ordinal classes as having no inter-class similarities. IG would thus make no distinction between the discriminative power of features such as f_1 and f_2 in Fig. 1.

3 The Probability Re-distribution Procedure

The idea behind the PRP is to revise the joint distributions so that features that occur in similar classes are given priority over those occurring in less similar ones. Fig. 2 gives a general overview of the PRP. It deviates slightly from the standard IG feature selection procedure and is composed of two steps.

The first step is performed prior to feature selection. It involves initialising the *similarity function* $S(c_i, c_j)$ with the similarity values between all pairs of classes $c_i, c_j \in \{c_1, c_2, \dots, c_M\}$ (subscripts denote class ordering). In this study, the similarity between two classes c_i and c_j was defined using a function that decays linearly with the distance between them,

$$S(c_i, c_j) = 1 - \frac{|i - j|}{M}. \quad (1)$$

Note that other functions could also be used, and as part of future work, we intend to experiment with the exponential and stepwise decay functions.

The second step of the PRP is the revision of $P(c, f_t)$ so as to yield a new distribution $P'(c, f_t)$. This is performed for each class c_j as follows:

$$P'(c_j, f_t) = \sum_k P(c_k, f_t) S(c_k, c_j). \quad (2)$$

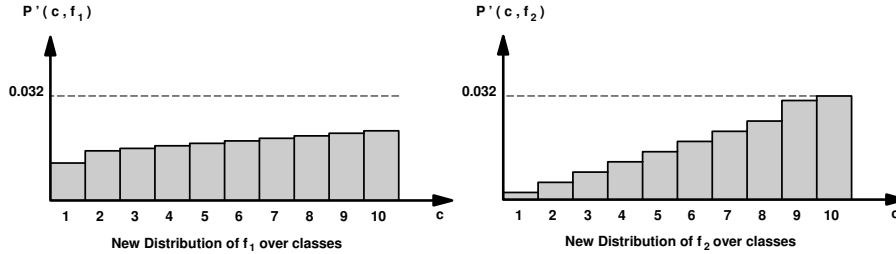


Fig. 3. The revised distributions of features f_1 and f_2 from Fig. 1.

Prior to using this Equation, it is important to note that the similarity values between the first class and the rest must be normalised, i.e $\sum_k S(c_i, c_k) = 1$. This property ensures that the area under $P'(c, f_t)$ equates to that under $P(c, f_t)$.

The overall effect of Equation 2 is that it returns a distribution that peaks at classes in which feature f_t is found, and decays linearly otherwise. An important aspect to note is that a feature that is concentrated in a few neighbouring classes would obtain a pyramid-shaped distribution. On the other hand, one that is concentrated around two distant classes would obtain a flat distribution as the linear decaying effect from the two classes cancel each other out. Such a distribution would therefore yield a relatively lower IG score.

To illustrate this idea more clearly, consider Fig. 3 which depicts the revised distributions of features f_1 and f_2 from the example introduced in Fig. 1. Note that the new distribution assigned to feature f_1 is almost flat, which is expected as f_1 occurs in two distant classes. In contrast to this, the distribution assigned to feature f_2 is relatively much steeper, as it occurs in two neighbouring classes. These new distributions clearly indicate that feature f_2 is more discriminative than feature f_1 .

4 Evaluation

The PRP was evaluated on three ordinal datasets with equal class distributions. The following list describes the steps that were taken to prepare these three datasets:

1. **The Edmunds Dataset:** This dataset was compiled from consumer reviews on used motor vehicles from the *Edmunds.com* website. Each review contained an integer rating between 0 to 99 (99 being the most positive) allocated by the consumer. These were used as the class labels. We, however, noted that consumers rarely gave a rating below 48 and hence we only used reviews within the range of 48 to 99. We also found some of the classes to be highly sparse and hence we collapsed the dataset to 26 classes by combining each even class with its neighbouring odd one. Finally each class was trimmed to 100 reviews, so as to obtain equal class distributions.

Table 1. MSE values when selecting features using IG, and IG with the PRP.

		10–350 dimensions			350–3000 dimensions		
		Actors	Movies	Edmunds	Actors	Movies	Edmunds
Naïve Bayes	IG	2.249	1.872	51.49	1.944	1.594	45.35
	IG+PRP	2.127	1.828	53.24	1.834	1.501	42.48
SVM	IG	2.524	2.184	74.07	2.206	1.892	57.53
	IG+PRP	2.570	2.351	73.28	2.288	1.921	58.92

- The Actors Dataset:** This dataset¹ was first used in an ordinal text classification study by Chakraborti *et al* [3]. It was obtained from reviews on the actors and actresses sub-topic of the *Rateitall.com* opinion website. Each review contained an integer rating ranging from 1 to 5 assigned by the reviewer. These were used as the class labels. All reviews that had less than 10 words were discarded. We then restricted the number of reviews per author per rating to at most 14, so as to avoid any authors bias from dominating the corpus (Pang *et al* [15] employed a similar approach). Finally, we formed 5 equally distributed classes containing 500 reviews each.
- The Movies Dataset:** We obtained this dataset from reviews on the movies sub-topic of the *Rateitall.com* opinion website. The compilation process was identical to that of the actors dataset.

The three datasets were pre-processed in a similar fashion using the standard procedure of tokenization, stop-word removal and finally stemming. The evaluation was performed using two classifiers: Multinomial Naïve Bayes [10], and Joachim’s SVM^{light} implementation [11] of Support Vector Machines (SVM). We used a one-vs-all approach to convert SVM^{light} into a multi-class classifier.

The evaluation metric used was the Mean Squared Error (MSE). We chose this principally because the problem is ordinal in nature. It, therefore, makes sense to prefer an incorrect prediction that is closer to the true class label over another that is more distant. This type of information is captured well by the MSE which emphasizes the deviation of the prediction from the true class label. All our experiments were performed using 10 fold cross validation and thereafter the paired *t*-test was used to assess significance.

4.1 The Effect of Probability Re-distribution on Information Gain

To assess the effect of the PRP on IG, we compared classifier performance on features returned by IG against those returned by IG in conjunction with the PRP (IG+PRP). Fig. 4 illustrates the results of our comparison at increasing dimensions (number of features). The graphs on the left column relate to Naïve Bayes performance, and those on the right to that of SVM. Table 1 provides a summary of the significance tests. Each cell of Table 1 contains the average MSE

¹ Available from <http://www.comp.rgu.ac.uk/staff/ram/downloads.html>

value of selected points within the specified dimensions. For each column, the performances significantly better ($p < 0.05$) than the rest, are shown in bold.

An overall view of Fig. 4 shows that the effect of IG+PRP is felt more by Naïve Bayes than by SVM. In particular, it can be seen that, as a result of the PRP, the performance of Naïve Bayes significantly improves at dimensions above 350. This is further supported by significance tests in Table 1. These results serve as evidence that features occurring in similar classes should be preferred over those occurring in less similar ones. In this regard, we performed a supplementary experiment to investigate the features that were most affected by the PRP. The most outstanding one we found was the number *2.7* from the Edmunds dataset which was demoted by about 8,500 positions by the PRP. A study of the corpus revealed that *2.7* was used in the context of the *2.7* litre engine which had several negative reviews attributed to it. It also occurred in several positive reviews in the context of trading in a *2.7* litre car for a much better one. Hence, due to its bi-polar occurrence, its distribution was revised so as to yield a lower IG score.

The graphs in Fig. 4 illustrate that there is no significant difference between IG and IG+PRP when the dimensions drop below 350. We speculate that this is due to the lack of sufficient features to describe the inter-class relationships as the error quite high at the low dimensions.

4.2 The Effect of Probability Re-distribution on the Classifiers

Two distinct aspects about the results are that: (1) SVM does not respond to the PRP, and (2) The performance of SVM is consistently poorer than that of Naïve Bayes. The second aspect, in particular, is in direct contrast to previous text classification studies where SVM has been found to be the better of the two [11, 13]. We try to justify these two phenomenon as follows.

Firstly, it has been previously been argued that SVM does not necessarily require feature selection as it can handle large numbers of features [11]. Furthermore, some studies have illustrated that blind feature selection may even be detrimental to SVM performance [2, 1]. If these two points are the case, then the feature vectors generated by IG and IG+PRP would be more or less synonymous to SVM, hence explaining the likeness in performance between them. However, the problem with this explanation is that, it would also be expected for the performance of SVM to approach, or even exceed, that of Naïve Bayes as the number of features grows. There is, however, no such trend in our results. We conjecture that this is due to the one-vs-all classification paradigm under which SVM operates.

Given a multi-classification task of say n classes, the one-vs-all approach builds n binary classifiers to discriminate each class against the rest. A new example is then classified by passing it to these n classifiers and assigning it the label of the classifier that prefers it best. This strategy works well when each class has a sufficiently distinct language. We, however, argue that it breaks down when applied to problems where the language-to-class boundaries are not as distinct as it disregards the similarities between the classes in the *all* cluster. As a consequence, SVM would inherently perform poorer than Naïve Bayes,

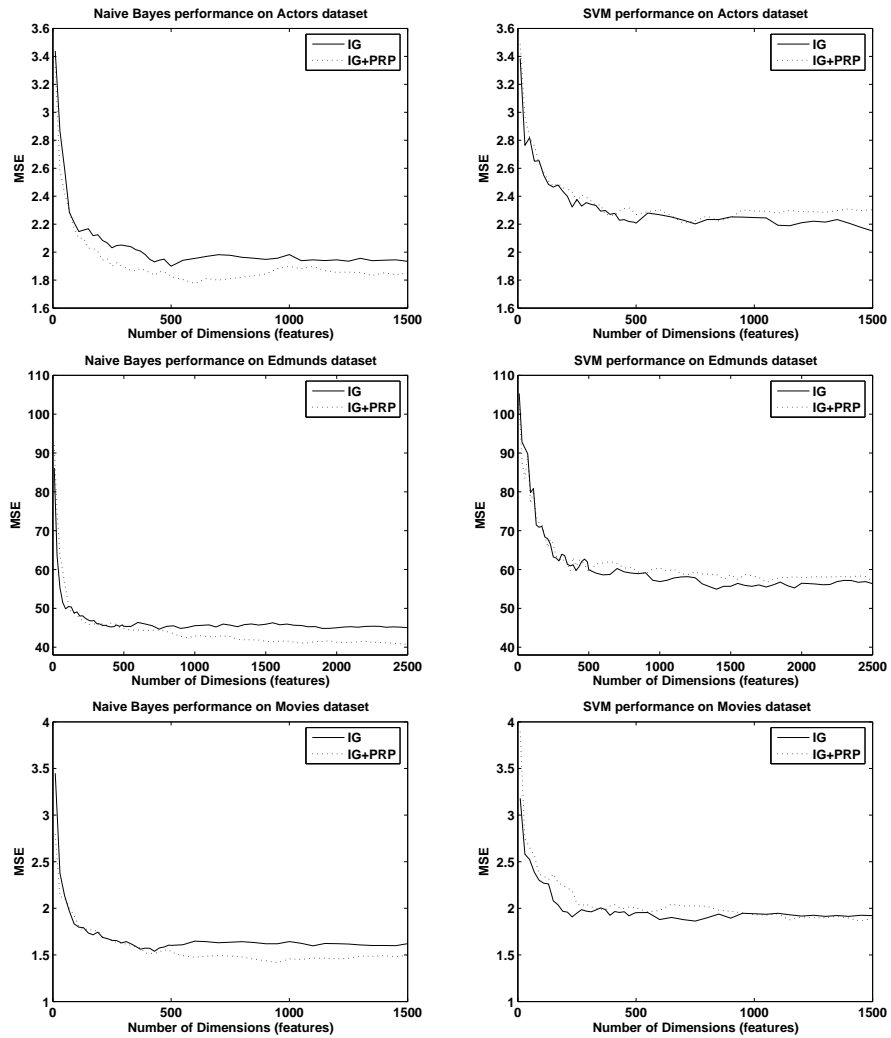


Fig. 4. Comparisons, at increasing dimensions, when selecting features using IG, and IG with the PRP.

which models each class independently and hence is able to appreciate the subtle language differences between them.

5 Discussion on Future Work

In theory the PRP can be tailored to suit the needs of a much broader class of problems. However, one limitation is that knowledge about the inter-class similarities is required prior to applying the PRP. This can be a setback in applications such as news filtering whereby the datastructure morphs rapidly, hence making human intervention impractical. In such scenarios the best option would be to automate the process of identifying inter-class relationships.

One possible solution is to employ confusion matrices. A confusion matrix compares a classifier’s predictions against the expert’s judgments on a class-by-class basis. The non-diagonal values in the matrix are indicative of the classes that the classifier finds hard to classify; the smaller the values the more separable the classes. This information could be utilised by mapping inter-class similarities to inter-class confusion. This mapping makes sense as classes that are easily separable should ideally be less similar to each other, and conversely those that are difficult to separate should be more similar to each other. In a previous study, Chakraborti *et al* [3] employed this principal to compute the Mutual Class Complexity (MCC) between two classes. The MCC has a value between 0 and 1 respectively signifying the simplicity and difficulty in separating two classes. This idea could be adopted into the PRP by setting the similarity function defined in Section 3 to be equal to the normalised MCC. In future work, we intend to integrate this idea into the PRP and evaluate the resultant on datasets that exhibit a wider variety of inter-class relationships.

The role of SVM for ordinal text classification is particularly interesting, because SVM is generally accepted as the de facto text classification algorithm. In our future work we intend evaluating different SVM classifier combination approaches for ordinal classification. In particular we identified three plausible alternatives to the one-vs-all approach from Machine Learning Literature.

The first is the all-vs-all approach [8] which builds $n(n - 1)/2$ classifiers to distinguish between *all pairs* of classes in an n -class problem. The fact that it considers all pairs of classes should better enable it to preserve the inter-class similarities for the ordinal task. The second approach uses error correcting output codes to perform classification. Here error correction techniques are used to enable the classification system to recover from incorrect predictions [5]. Essentially performance improvements are achieved by translating standard class labels into more granular binary codes of length m . Each class code is designed to be sufficiently different from all others in the hamming distance sense. Classification then involves training m classifiers for each bit position of the n codes. The last method is a simple approach to ordinal classification discussed by Frank and Hall [7]. Unlike the first two approaches, it is specifically formulated for the ordinal classification setting. Essentially it exploits the class ordering information to create $(n - 1)$ classifiers, where each classifier is devoted to learning the

classification at a given point on an incremental ordinal scale. As part of future work, we intend to compare our results here against those of these three approaches.

6 Related Work

To the best of our knowledge there has not been much work in feature selection for ordinal text classification. However a study by Schmitt *et al* [16] in Case Based Reasoning research has interesting similarities the approach presented here. In their study, Schmitt *et al* modifies the IG formulae in order to select the most useful attributes to present during online dialogues with consumers. A rough analogy of what they propose to do is revise the probability $P(c_j, f_t)$ to be the average similarity between all documents in the corpus to the feature f_t . This approach differs from ours in some respects, for example they do not ensure that the area under $P(c, f_t)$ is not distorted.

7 Conclusion

In this paper we show that inter-class similarity knowledge is crucial for feature selection in ordinal text classification. Empirical results on three ordinal datasets, confirm that exploiting this knowledge significantly improves classifier performance. This has strong implications on many ordinal problems, such as sentiment or opinion text classification, whereby feature selection algorithms, such as IG, are applied without exploiting inter-class similarities.

The solution we propose is the PRP which is performed prior to computing the IG score of a feature. The elegance in this is that the PRP is independent of IG. It therefore has the potential to be easily adopted into many feature selection frameworks (e.g. CHI, MI and Odds Ratio [18, 12]) without much effort. Future work will look at extending the PRP to more general text classification problems by automating the acquisition of inter-class similarity knowledge.

References

1. Ron Bekkerman, Ran El-Yaniv, Naftali Tishby, and Yoav Winter. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research*, 3:1183–1208, 2003.
2. Janez Brank, Marko Grobelnik, Nataša Milic-Frayling, and Dunja Mladenic. Feature Selection Using Linear Support Vector Machines. Technical report, Microsoft Research, June 2002.
3. Sutanu Chakraborti, Rahman Mukras, Robert Lothian, Nirmalie Wiratunga, Stuart Watt, and David Harper. Supervised Latent Semantic Indexing using Adaptive Sprinkling. In *IJCAI*, pages in–press, 2007.
4. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

5. Thomas G. Dietterich and Ghulum Bakiri. Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.
6. George Forman. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305, 2003.
7. Eibe Frank and Mark Hall. A Simple Approach to Ordinal Classification. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, pages 145–156, London, UK, 2001. Springer-Verlag.
8. Johannes Fürnkranz. Pairwise Classification as an Ensemble Technique. In *ECML '02: Proceedings of the 13th European Conference on Machine Learning*, pages 97–110, London, UK, 2002. Springer-Verlag.
9. Evgeniy Gabrilovich and Shaul Markovitch. Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5. In *The 21st International Conference on Machine Learning (ICML)*, pages 321–328, Banff, Alberta, Canada, July 2004.
10. Jaime Teevan Jason D. M. Rennie, Lawrence Shih and David R. Karger. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *ICML '03*, 2003.
11. Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *ECML*, pages 137–142, 1998.
12. Dunja Mladenic and Marko Grobelnik. Feature Selection for Unbalanced Class Distribution and Naive Bayes. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, pages 258–267, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.
13. Bo Pang and Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the ACL*, pages 271–278, 2004.
14. Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.
15. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.
16. Sascha Schmitt, Philipp Dopichaj, and Patricia Domínguez-Marín. Entropy-based vs. similarity-influenced: Attribute selection methods for dialogs tested on different electronic commerce domains. In *ECCBR '02: Proceedings of the 6th European Conference on Advances in Case-Based Reasoning*, pages 380–394, London, UK, 2002. Springer-Verlag.
17. C. E. Shannon. A Mathematical Theory of Communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, 2001.
18. Yiming Yang and Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420. Morgan Kaufmann Publishers Inc., 1997.