A Framework for Comparing Text Categorization Approaches

Isabelle Moulinier
LAFORIA-IBP-CNRS
Université Paris VI
4 place Jussieu
F-75252 Paris Cedex 05 – FRANCE
moulinie@laforia.ibp.fr

Abstract

For the past few years, text categorization has emerged as an application domain to machine learning techniques. Several approaches have already been proposed. This paper does not present yet another technique. It is rather an attempt to unify the approaches encountered so far. Moreover this state-of-the-art enables us to stress a shortcoming in earlier research: the lack of evaluation of inductive learners in the categorization process. We present a first attempt to remedy this lack. We expose an experimental framework, that fits in with our unified view of text categorization methods. This framework allows us to conduct a set of tentative experiments in order to assess which characteristics allow a learner to perform well on the text categorization task.

Introduction

Text categorization, which is often defined as the content-based assignment of one or more predefined categories to texts, has become important in two aspects. On an information retrieval (IR) point of view, information processing needs have increased with the rapid growth of textual information sources, such as Internet. Text categorization can be used to support IR or to perform information extraction, document filtering and routing to topic-specific processing mechanisms (Hayes *et al.* 1990; Riloff & Lehnert 1994). On a machine learning (ML) point of view, recent research has be concerned with scaling-up (e.g. data mining (Holsheimer & Siebes 1994)). Text categorization is a domain where large data sets are available and which provides an application field to ML (Lewis & Catlett 1994; Cohen 1995). Indeed, manual categorization is known to be an expensive and time-consuming task. Hand-crafted knowledge engineered systems such as CONSTRUE (Hayes & Weinstein 1990) also have such drawbacks. ML approaches to *classification* (text categorization is a classification task) suggest the construction of categorization means using induction over pre-classified samples. They have been rather successfully applied in various studies, e.g. (Lewis & Ringuette 1994; Apté, Damerau, & Weiss 1994; Wiener, Pedersen, & Weigend 1995).

In this paper, we are primarily concerned with the analysis of these earlier studies on text categorization. Our presentation is two-folded. We first show that, even though the nature of the inducer used in each approach may differ, most approaches have common characteristics in the whole categorization process. Then, we discuss the issue of choosing one technique rather than another. Actually, many approaches have been suggested; these include numerical learning such as Bayesian classification (Lewis & Ringuette 1994), or symbolic learning like in (Moulinier & Ganascia 1995). However, no assessment has be conducted on whether a given learning technique was superior to another on the text categorization task, even though the sketch of an answer can be found in (Lewis & Ringuette 1994). We first design an experimental framework which fits in with our unifying view of text categorization systems. In that framework, we compare several learners in order to try and extract major characteristics of both data and learners, that lead to good performances on the text categorization task.

In the next section, we present a unifying view of research in text categorization. An experimental framework for comparison is given next, while preliminary experiments are reported and discussed in the last section.

Text Categorization: a Unifying View

Text categorization is at the meeting point between ML and IR, since it applies ML techniques for IR purposes. In the following, we adopt a ML point of view. Many existing text categorization systems share certain characteristics. Namely, they all use induction as the core of learning classifiers. Moreover, they require a *text representation* step that turns textual data into learning examples. This step involves both IR and ML

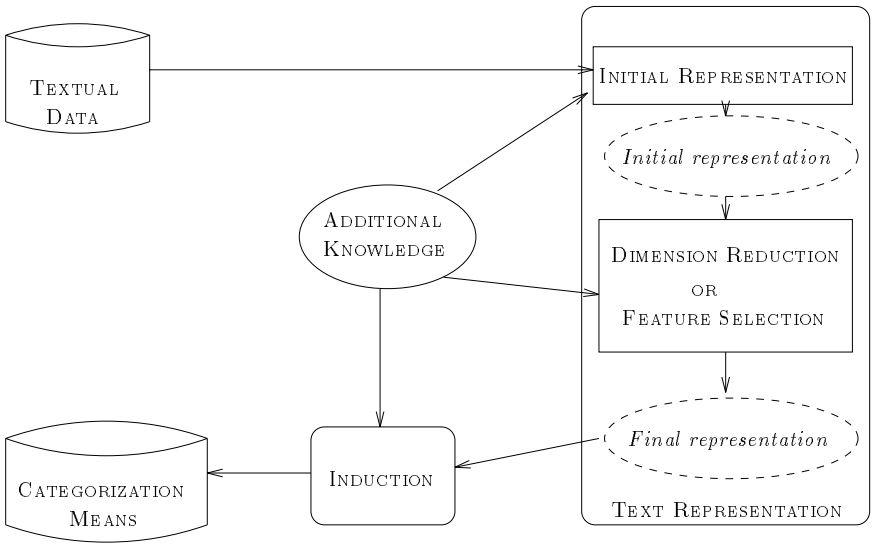

Figure 1: A unified framework for text categorization

techniques. Finally, additional knowledge may be provided to enhance the whole categorization task. We present these three aspects in the remainder of this section. Their relationships are summarized in Figure 1.

Text Representation

Text representation in categorization differs from its homologue in IR. It is more specific, as it requires further processing. In fact, we can distinguish two steps when representing a text. The first step is the standard IR representation, for instance a boolean model as in (Lewis & Ringuette 1994; Apté, Damerau, & Weiss 1994; Moulinier & Ganascia 1995) or a frequency model as in (Fuhr *et al.* 1991; Wiener, Pedersen, & Weigend 1995). Nevertheless, this step may not be sufficient to produce tractable data for learners. Indeed, the feature set that results from such a representation can be numbered in hundreds of thousands. Even though some studies have reported working with such a number of features (Yang 1994; Creecy *et al.* 1992), few inductive learners can handle such a number of features. For instance, typical experiments in ML hardly ever deal with more than a hundred of features. Therefore, a second step is unavoidable: it consists in the reduction of that original feature set, commonly known as *dimensionality reduction* in pattern recognition.

We can distinguish two axes for dimensionality reduction: its *scope* and its *nature*. The scope of reduction is concerned with the universality of the resulting feature set, whereas its nature describes how the features are selected. In (Apté, Damerau, & Weiss 1994), two alternatives to the scope of reduction are suggested: category-oriented, or *local*, and overall, or *global*, feature set reduction. The so-called global reduction (Maron 1961; Apté, Damerau, & Weiss 1994) provides an inductive learner with the same feature set for each category, while local reduction selects a specific feature set for each category (see for instance (Apté, Damerau, & Weiss 1994; Lewis 1992; Wiener, Pedersen, & Weigend 1995)). The nature of reduction can also be qualified by two different means: *filtering* and *construction*. Filtering aims at reducing the number of features by selecting the best ones according to some criterion; such criteria include mutual information (Lewis & Ringuette 1994; Moulinier & Ganascia 1995), frequency (Apté, Damerau, & Weiss 1994), term ranking (Fuhr *et al.* 1991; Wiener, Pedersen, & Weigend 1995) or expert's judgment (Maron 1961). Construction has a lesser impact in text categorization. Instead of selecting a subset of the original feature set, new features are constructed as combinations of original features. Latent Semantic Indexing (LSI) (Deerwester *et al.* 1990), as used in (Wiener, Pedersen, & Weigend 1995), is such a constructive approach.

Inductive Construction of Categorizers

Once texts are turned into learning examples, inductive learners are used to induce categorizers. Since the ideas behind these learners are well known in ML, we only review those used in text categorization experiments.

In most categorization systems, induction is performed by a numerical learner. Linear regression (Biebericher *et al.* 1988; Fuhr *et al.* 1991), Bayesian classifiers (Maron 1961; Lewis 1992), k-nearest neighbors (Masand, Linoff, & Waltz 1992; Yang 1994), neural nets (Wiener, Pedersen, & Weigend 1995) and threshold computation (Liddy, Paik, & Yu 1994) are instances of such learners. Recent studies have introduced symbolic learners in order to build categorizers: decision tree constructors (Fuhr *et al.* 1991; Lewis & Ringuette 1994), relational k-DNF learners (Cohen 1995) and production rule inducers (Apté, Damerau, & Weiss 1994; Moulinier & Ganascia 1995).

We now outline a couple of differences between these learners, that may be significant for the text categorization task. First, numerical and symbolic learners differ their abilities to handle structured features and produce understandable classifiers. The instance language, i.e. the feature set issued from text representation, is known to strongly bias the inductive learner (Michalski 1983). Symbolic learners usually deal with a structured instance language but perform rather poorly when they are confronted with numerical data. On the other hand, numerical learners can not easily deal with structured features. Moreover, symbolic learners are often said to produce interpretable classifiers. However, text categorization is a domain where classifiers are quite verbose: a categorization system may include several thousands of rules (Moulinier & Ganascia 1995), which can hardly be considered as interpretable.

Finally, we believe that resistance to noise may be critical for the text categorization task, since textual databases are usually rather large and are bound to be noisy. Some symbolic learners like ID3 (Quinlan 1986) or CHARADE (Ganascia 1993) are said to construct consistent descriptions of concepts, i.e. a description is generated when all examples covered by this description belong to the same concept. Such learners are not noise-resistant. However, most ML techniques provide some means to take noise into account.

What Impact Has Knowledge ?

Our third concern is the analysis of the use and impact of knowledge during the whole categorization process. As shown in Figure 1, additional knowledge may appear during any of the two major subtasks of categorization, i.e. text representation and induction.

There is no single definition for knowledge. We therefore distinguish three facets to the term knowledge. In IR and numerical learning, knowledge is often extracted from data. For instance, a frequency-based model can be considered as adding knowledge to a boolean model. We call the second facet *domain knowledge*. Such a kind of knowledge is provided by an external interaction and refers to a specific application. For example, machine-readable dictionaries are sources of domain knowledge. Lastly, an inductive bias can be considered as a knowledge source for the learner or the reduction method.

These three facets of knowledge are mostly evoked during the text representation step. Local selection, LSI or even the frequency based model can be considered as adding knowledge extracted from data to a global text representation based on a boolean model. Domain knowledge has been used by (Liddy, Paik, & Yu 1994), where a machine-readable dictionary was employed to build the initial representation. We also consider the assignment of a greater weight to words appearing in the headlines of a news-story (Apté, Damerau, & Weiss 1994) as domain knowledge. In (Cohen 1995), the expressive power of a relational formalism, i.e. language bias, enables the representation to take into account the positions of words inside a document.

There has been little research conducted on the use of knowledge during the inductive phase of categorization. Nevertheless, a noticeable attempt is presented in (Wiener, Pedersen, & Weigend 1995), where the authors group categories according to semantic characteristics and induce categorizers of these sub-domains. In (Fuhr *et al.* 1991), the authors used knowledge to guide an indexing system: for instance, knowledge enabled the discrimination among candidate keywords issued from the inductive step.

Most experiments reported in text categorization, which used additional knowledge in the representation and induction steps, show that an enriched categorization system outperforms a naive approach. However, few studies have reported experiments, where varying amounts of knowledge were involved. For instance, (Wiener, Pedersen, & Weigend 1995) reported an enhancement of 5% using LSI and a hierarchical net over boolean features using a flat network. Similarly, (Apté, Damerau, & Weiss 1994) reported a increase of 5% between a locally reduced representation based on frequency and weight assignment, and a global boolean representation.

Limitations of this Unifying View

There remain some text categorization approaches that do not fit into the preceding schema. The processing step between the initial representation and the final representation does not always imply dimensionality reduction. For instance, in (Creecy *et al.* 1992), the authors expand the initial representation and their learner has to deal with over 4 million features. Reduc-

tion of the training set, as opposed to dimensionality reduction, has also been eluded in this schema. Sampling, as described in (Lewis & Catlett 1994) and used by (Cohen 1995), is one such approach for reducing the number of training examples.

Finally, for the sake of simplicity, we have not included feedback into the whole categorization system. Clearly, however, all systems perform hand-driven feedback, when tuning parameters to optimize some evaluation criterion. We are not aware of automatic feed-back in the context of text categorization.

A Framework to Compare Learners

Very few studies have conducted a thorough comparison between learners on the text categorization task. In (Lewis & Ringuette 1994), two learning approaches are compared: Bayesian classification and decision tree construction; (Wiener, Pedersen, & Weigend 1995) experimented on several neural net models. However, most studies report some performance improvements of a given approach over others. Hence, there has been no conjecture on the properties a learner ought to possess so that it performs well on the text categorization task. Moreover, comparing existing approaches is inconclusive to assess learners inasmuch as no clear distinction can be made between the exact roles of text representation and inductive learning. In this section, we propose an experimental framework in order to compare individual learners, and not the whole categorization system.

Text Representation and Learning Scheme

Text representation is a two-stage process. The first stage is concerned with the initial text representation. We are confronted with an alternative: we can either hold text representation constant or choose the text representation that is best suited to each learner. We choose to have a unique representation for all learners and use a naive boolean model.

In a second stage, this boolean representation is reduced using local filtering based on the mutual information criterion. For each category, we select the features that obtain the n top-most scores using the mutual information criterion between the given category and a feature.

To end up, we obtain the following learning scheme. Since learners are typically used for single-class prediction, the assignment of n categories to a document is transformed into n assignments decisions on each single category. The original text database is translated in terms of locally selected features for each category.

Evaluation Criteria

Evaluation criteria in IR and in ML differ. We choose to assess our experiments with an IR criterion, since accuracy, a measure commonly used in ML, is biased by the high disproportion between the assignment and the non-assignment of categories. Thus, we consider recall and precision as evaluation measures. We use micro-averaging (Lewis 1992, Sec. 6.4) as a means of cumulating performances over all categories. However, since recall usually goes up (respectively down) when precision goes down (respectively up), it is rather tricky to assess performances on the basis of these two measures. Among several summarizing measures that have been proposed, we choose the F_β-measure (Lewis 1995) as an evaluation criterion:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R},$$

where R denotes recall, P precision and β varies from 0 to infinity.

Experimental Results

The Reuters Corpus

We carried out our experiments on the Reuters dataset of financial newswire stories from the year 1987, also identified as Reuters-22173[1]. The original corpus is composed of 21450 manually indexed stories divided into a learning set (14704 documents) and a testing set (6746 documents). Among 689 subjects of interest, including topics, places or company names, we worked on a set of 135 categories that were provided together with the formatted version of the corpus. We decided to overlook stories without category assignment, since we could not possibly learn from them[2]. This left us with 7789 learning and 3875 testing examples described by 22791 words provided by Lewis' processing (Lewis 1992, p. 99).

Which Learners ?

Our experiments were conducted on four learners which illustrate symbolic and numerical learning. An implementation of ID3 (Quinlan 1986) and CHARADE (Ganascia 1991), a production rule learner, represent symbolic learners, while a k-nearest neighbors algorithm called IB, and a Bayesian approach, NaiveBayes, are instances of numerical learners. The ID3, IB and NaiveBayes are those implemented in the MLC++ library (Kohavi *et al.* 1994).

[1] The Reuters dataset can be obtained by anonymous ftp from /pub/reuters1 on ciir-ftp.cs.umass.edu.

[2] Overlooking stories without category assignment was a misunderstanding of the original corpus labels.

Learner	# features	F_1	Break-even point
ID3	75	78.1	
CHARADE	75	78.2	
IB	10	75.1	
NaiveBayes	10	70.2	
Neural Nets (boolean model)	–		77.5
Neural Nets (enriched model)	–		82.5
Swap-1 (local boolean model)	–		78.5
Swap-1 (enriched model)	–		80.5

Table 1: Micro-averaged performances of learners on the Reuters testing set.

Results

We ran several series of experiments; for each algorithm, we used varying sizes of feature sets. The evaluation was conducted on the set of 3875 testing examples. Results, reported in Table 1, show the best performances with regards to the F_1 criterion. Results from earlier experiments on the same corpus complete this summary. Neural nets refer to the experiments presented by (Wiener, Pedersen, & Weigend 1995), while Swap-1 is a production rule learner used in (Apté, Damerau, & Weiss 1994). In both cases, the enriched model takes into account various kinds of knowledge sources, while the (local) boolean model is very close to our naive framework.

These earlier experiments were not evaluated using the same criterion. However, the break-even point and the F_1 measure may be compared since $F_1(P^*, P^*) = P^*$, where P^* is the precision obtained at the break-even point. Finally, it is worth noticing that the performances of four learners out of six (ID3, CHARADE, Swap-1 and Neural Nets) are very close, when these learners are given a similar text representation.

Discussion

The difference between some learner's microaveraged performance is not really significant. Let us, for instance, consider CHARADE and ID3. The microaveraged F_1 is roughly the same; however, these two learners have distinct behavior: while CHARADE favors recall, ID3 favors precision. Moreover, the gap between the values of recall and precision is wider using the decision tree technique (cf Table 2).

Furthermore, to get a better insight, we looked at the behavior of each learner on individual categories. Results on a subset of 17 categories are reported Figure 2. This subset groups the most frequently assigned categories on the training set, as well as some randomly selected ones. The number of positive training examples is also given. It is worth noticing that no learner outperforms the others on all categories, even though in most cases symbolic learners show better performances.

Table 1 may give rise to a unfortunate association between rule learners and large feature sets, as opposed to statistical learners and small feature sets. In Table 2, we show that this association does not hold: ID3 performs well with few feature and IB performs equally well with 75 features. However, CHARADE performances are greatly deteriorated by a small set of features.

An alternative hypothesis was that symbolic learners performed their own selection of features during the learning phase, whereas numerical ones did not. In Table 3, we report the number of features that appear in the descriptions (either tree or rule set) learned by ID3 and CHARADE for each of the 17 categories. A striking difference between these two learners can be seen: while ID3 does have some kind of feature selection, since it does not use all features (even with only ten features) in the decision tree. CHARADE, on the other hand, uses most of the available features; this characteristics is emphasized by the use of redundancy during learning.

Another direction would need further investigation. In our framework, we learn to decide whether a category can be assigned to a text or not. Since text categorization is more concerned with the assignment than with the non-assignment, it would be interesting to assess the ability of learners to learn a concept with few positive examples. At first view, numerical learners are less sensitive to irregular distributions. For instance, IB and NaiveBayes perform rather well on the `veg-oil` and `palm-oil` categories, whereas CHARADE does poorly.

Finally, the poor performance of all learners on the `yen` category is striking (cf. Figure 2). Our belief on this particular case is that induction is flawed by representation. Indeed, documents from category `yen` and `dlr` often use a similar vocabulary. However, there are less examples labeled with category `yen`. Thus, it

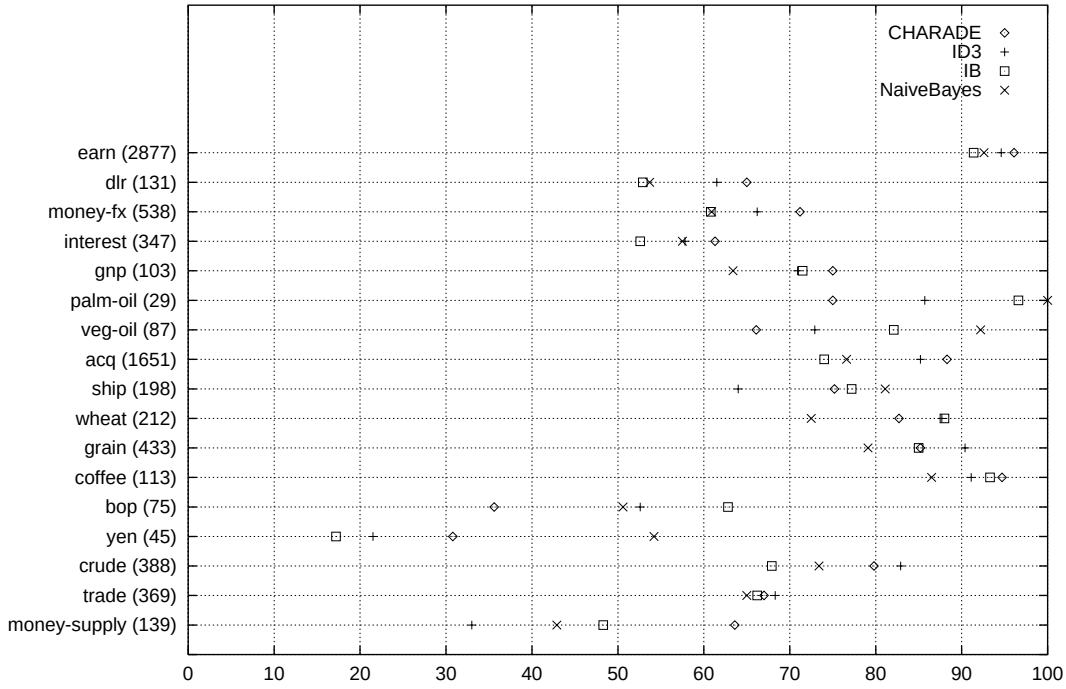

Figure 2: Comparison on a subset. The criterion is F_1 expressed as a percentage.

is rather hard to distinguish this category from the `dlr` category. This can be related to a kind of noisy data.

Conclusion

In this paper, we have presented a review of current research in text categorization and provided evidence that the properties of learners should be taken into account, in order to choose one particular learner to induce categorizers.

We argued that considering a single evaluation measure could not properly characterize the abilities of a given learner to the text categorization task. We also outlined differences between numerical and symbolic learners, in the language instance as well as in data distribution. Considering these two dimensions and the results reported in the last section, we believe that it would be interesting to study hybrid approaches to text categorization: data characteristics could guide the choice of a learner for each category.

Moreover, as ML algorithms currently have difficulties to deal with both large feature and example sets, future research should be dedicated to reducing these sets. One path has been pointed out by (Lewis & Catlett 1994) and consists in reducing the sample set. We prefer another path, which includes designing specific algorithms for dimensionality reduction and enhancing the initial text representation, using for instance linguistic knowledge.

Finally, we have not addressed the influence of noisy data on learning in a categorization context. The experiments we reported in this paper need to be further analyzed and developped in order to assess wether resistance to noise is important. However, we can clearly distinguish between two types of noise: noise may be present in the original textual dataset (i.e. two identical texts with different categories, as it appeared in the Reuters dataset), or it may be introduced by text representation, especially during the reduction step.

Acknowledgments

We wish to thank the anonymous referees and David Lewis for their helpful comments. The work presented here has benefited from many discussions with our colleagues at the Laforia laboratory.

References

Apté, C.; Damerau, F.; and Weiss, S. 1994. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems.*

Biebericher, P.; Fuhr, N.; Lustig, G.; Schwantner, M.; and Knorz, G. 1988. The automatic indexing system AIR/PHYS – from research to application. In *Proceeding of the 11th SIGIR.*

Learner	# features	Precision	F_1	Recall
ID3	75	81.2	78.1	75.4
ID3	50	80.7	77.8	75.2
ID3	10	80.9	75.2	70.3
CHARADE	75	76.9	78.3	76.9
CHARADE	50	79.5	77.6	75.6
CHARADE	10	86.4	59.0	44.8
IB	10	81.5	75.1	69.4
IB	4	80.2	73.6	68.0
IB	75	86.4	74.1	64.7
NaiveBayes	10	64.3	70.2	77.5
NaiveBayes	4	73.5	70.0	66.7
NaiveBayes	75	49.8	62.9	85.4

Table 2: Influence of the number of features on microaveraged performances.

Cohen, W. 1995. Text categorization and relational learning. In *Proceedings of the Twelfth International Conference on Machine Learning.*

Creecy, R. H.; Masand, B. M.; Smith, S. J.; and Waltz, D. L. 1992. Trading MIPS and memory for knowledge engineering: Classifying census returns on the connection machine. *Communication of the ACM.*

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and R., H. 1990. Indexing by latent semantic indexing. *Journal of the American Society for Information Science* 41(6):391–407.

Fuhr, N.; Hartmann, S.; Lustig, G.; Schwantner, M.; and Tzeras, K. 1991. AIR/X — a Rule-Based Multistage Indexing System for Large Subject Fields. In *Proc. of RIAO'91.*

Ganascia, J.-G. 1991. Deriving the learning bias from rule properties. In Hayes, J. E.; Mitchie, D.; and Tyngu, E., eds., *Machine Intelligence 12.* Oxford: Clarendon Press. 151–167.

Ganascia, J.-G. 1993. TDIS: an Algebraic Formalization. In *International Joint Conference on Artificial Intelligence.*

Hayes, P., and Weinstein, S. 1990. CONSTRUE/TIS: a system for content-based indexing of a database of news stories. In *Second Annual Conference on Innovative Applications of Artificial Intelligence.*

Hayes, P.; Andersen, P.; Nirenburg, I.; and Schmandt, L. 1990. TCS: A Shell for Content-Based Text Categorization. In *Proceeding of the Sixth IEEE CAIA*, 321–325.

Holsheimer, M., and Siebes, A. 1994. Data mining. The search for knowledge in databases. Technical report, CWI.

Kohavi, R.; John, G.; Long, R.; Manley, D.; and Pfleger, K. 1994. MLC++: A Machine Learning Library in C++. In *Tools with Artificial Intelligence*, 740–743. IEEE Computer Society Press.

Lewis, D., and Catlett, J. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference on Machine Learning.*

Lewis, D., and Ringuette, M. 1994. A comparison of two learning algorithms for text categorization. In *Symposium on Document Analysis and Information Retrieval.*

Lewis, D. 1992. *Representation and Learning in Information Retrieval.* Ph.D. Dissertation, Graduate School of the University of Massachusetts.

Lewis, D. 1995. Evaluating and optimizing autonomous text classification systems. In *Proceedings of SIGIR-95.*

Liddy, E.; Paik, W.; and Yu, E. 1994. Text categorization for multiple users based oon semantic features from a machine-readable dictionary. *ACM Transactions on Information Systems.*

Maron, M. E. 1961. Automatic indexing: an experimental inquiry. *Journal of the Association for Computing Machinery* (8):404–417.

Masand, B.; Linoff, G.; and Waltz, D. 1992. Classifying News Stories unsing Memory Based Reasoning. In *Proc. of the 15th SIGIR.*

Michalski, R. 1983. A theory and methodoly of inductive learning. *Artificial Intelligence* (20):111–161.

Moulinier, I., and Ganascia, J.-G. 1995. Confronting an existing machine learning algorithm to the text categorization task. In *Working notes, IJCAI-95*

Category	CHARADE (75)	CHARADE (10)	ID3 (75)	ID3 (10)
earn	73	10	51	9
dlr	72	10	21	10
money-fx	75	10	41	10
interest	72	10	31	10
gnp	75	10	12	10
palm-oil	64	10	4	4
veg-oil	70	10	13	9
acq	73	10	50	9
ship	70	10	34	9
wheat	72	10	15	10
grain	74	10	38	10
coffee	66	10	5	5
bop	68	10	19	10
yen	65	10	13	9
crude	68	10	28	10
trade	74	10	39	10
money-supply	69	10	21	8

Table 3: Number of features actually used in the learned categorization tool by ID3 and CHARADE.

Workshop on New Approaches to Learning for Natural Language Processing.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning* 1:81–106.

Riloff, E., and Lehnert, W. 1994. Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems* 12(3):296–333.

Wiener, E.; Pedersen, J.; and Weigend, A. 1995. A neural network approach to topic spotting. In *Symposium on Document Analysis and Information Retrieval.*

Yang, Y. 1994. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of SIGIR-94.*