

The Role of Named Entities in Text Classification

Quintin Armour, Nathalie Japkowicz and Stan Matwin

School of Information Technology and Engineering

University of Ottawa

Ottawa, Canada K1N 6N5

{qarmour, nat, stan}@site.uottawa.ca

Abstract

Named entities are typically associated with names of people, places and organizations and constitute a group of textual elements present in almost any type of document. The general techniques used to extract them and their variable-length property also makes them an attractive type of attribute to study in text classification. In this paper, several data sets are characterized as being either dependent or independent of named entities with a Naive Bayes based ranking technique. Using this characterization, results are presented which find named entities to be in fact useful in classification tasks, and that accuracy can be improved by considering them as a special type of attribute. Namely, the inclusion of regular terms, named entity representation and the frequency with which a classifier is retrained all have an impact on the classification of documents where named entities are important.

1 Introduction

A central problem in information retrieval is the manner in which a document is represented. The quality of the representation directly impacts the way documents are organized. A better organization of documents leads to faster and more accurate document retrieval. Traditional methods for document representation or *indexing* include such methods as the well-known *bag-of-words* (BOW), and other *n*-gram based or phrase-based approaches (Caropreso et al., 2001). This paper considers the utility of *named entities* for document indexing in the related problem of text classification. Named entities are a kind of phrasal element which describe *named* elements such as people, places and organizations. These entities are semantically rich, multi-word elements and occur in almost every type of document of interest to those performing text classification. They are therefore of con-

siderable interest, especially in the classification of news articles and email where they are most prominent.

The majority of research in text classification to date involves some kind of evaluation with the Reuters-21578 data set¹. In (Bekkerman et al., 2001), the authors argue that the Reuters data set in particular, since it was manually labeled, favours the development of keyword-based text classifiers. In other words, if the Reuters data set only works well with keyword features, then any more sophisticated approach probably will not seem too successful when tested on this *standard* data set. This conjecture regarding the Reuters data set is supported in the work by (Lewis, 1992) which shows that a phrasal and word-cluster approach to text classification performs poorly on the Reuters data set. The main reason for this decay in accuracy, according to Lewis, is the sparseness of the document-feature matrix. The reasoning behind this argument is that if content words, like nouns and verbs, are not very frequent by nature, then sequences of content words are even less likely to occur since there are greater number of possible combinations. However, in (Croft et al., 1991) the authors mention that “words may be associated and co-occur, but not be part of the same phrasal concept. NE (named entity) extraction preserves the concept.” In other words, named entities may not suffer the same effects of data sparseness as other phrasal approaches since it is known that the words belong to the same concept and the semantics can be preserved. This paper provides us with an indication that named entities are useful elements to extract from text because they behave essentially like multi-word nouns.

Such work has generated some interest in considering named entities for text classification tasks. In (Cooley, 1999), the possibility of using named entities as a way of reducing the dimensionality of the feature space in the classification of news articles is investigated. Cooley argues that since named entities are present in almost all text documents and certainly in all news articles, that

¹<http://www.daviddlewis.com/resources/testcollections/reuters21578/>

this may be a successful tool in adding some meaning to the classification process. In his results, he finds out that the classifier trained only on named entities, although it reduced the feature space, does not produce sufficiently high accuracy to be of any use. The limitation of the work is that the results are based only on one data set and do not investigate the reasons for which named entities produced poor results. There are also some characteristics of classification algorithm he uses that can explain the results he obtained.

In a later work (Clifton and Cooley, 1999), a tool developed to perform *Topic Detection and Tracking* (TDT) is presented. This tool makes good use of named entities by incorporating them as the keystone in the main algorithm. In the algorithm, named entities are grouped to form frequent item sets using a method from association rule mining. These item sets are used in a technique to cluster documents sharing a common topic. Based on the success of this technique, named entities can be considered as being linked to the topic of a news article. This fact suggests that named entities would in fact be useful in text classification tasks.

As evidence of the current topicality of named entities in text classification, In a recent technical report (Bekkerman et al., 2004), the author presents an evaluation of text classification methods on the Enron corpus². He mentions in his future work that “[n]amed entities may be highly relevant features. It would be desirable to incorporate a named entity extractor into the foldering system.” The incorporation of named entities into text classification tasks is exactly the purpose of this work.

The goal of this paper is twofold. The first goal is to identify the situations in which named entities are the key features needed for accurate text classification. The second goal is to study how, in these same situations, named entities and their properties can be used to achieve better classification accuracy. The situations which are hypothesized to improve the utility of named entities include the following scenarios:

- Named entities used in the absence of other elements
- Named entities represented as a single entity instead of by their component words
- Named entities in classification tasks where the negative effects of time are reduced by more frequent classifier retraining

The first point will discover whether or not named entities are negatively affected by the presence of other types of features. In particular, we refer to *regular terms* which we defined as those elements that are not named entities. For the purposes of this paper, regular terms consist

²<http://www-2.cs.cmu.edu/~enron/>

of single word components. The second point will investigate which feature representation is best for named entities in those tasks where they are identified as being useful. The final point will show whether or not named entities are more significantly affected by the passage of time in comparison to other types of features. If so, the possibility of retraining the classifier to remedy the loss in accuracy over time will be studied. These goals will be achieved as follows. First a method for characterizing a data set will be developed. This method will enable us to evaluate an attribute list for a given data set and state whether or not the data is dependent on it. Second, each of the three situations described above will be evaluated in the context of the characterizations made.

This paper will be organized as follows. In Section 2, the method for data set characterization will be presented. Next, in Section 3, the results for three experiments considered for named entities in text classification are provided. Finally, in Section 4 we state our conclusions along some plans for future work.

2 Data Set Characterization

This section describes the results of a method developed to characterize a data set given an attribute list. The method is required in order to conclude whether an attribute list is adequate in its representation of a data set. In this paper, we consider two types of attributes: named entities and regular terms. The named entities are extracted from the documents using GATE (General Architecture for Text Engineering) (Cunningham et al., 2002). The regular terms are obtained by creating a BOW from the texts with named entities removed. This mechanism to evaluate an attribute list in the context of classification will allow us to measure and categorize the data sets studied according to how well they are represented by named entities. The results are compared with those obtained using regular terms. For further details and for a comparison to other methods, refer to (Armour, 2005).

Before presenting the data set characterization method, we will first briefly introduce the data sets used in this work. There are six data sets of interest:

- Hockey: data collected from Yahoo! News related to hockey. The articles relate to the subjects of junior and professional hockey, and form a binary classification task. This data was part of an existing hierarchy, since these articles fall into the larger domain of sports.
- Movies: data collected from Yahoo! News related to entertainment. The articles discuss events related to television and film.
- Rt-earn: data from the Reuters-21578 data set. The top 10 categories, as defined by (Estabrooks et al.,

2004), were used and the *earn* category was labeled as the positive class.

- Rt-acq: same as *Rt-earn* except the *acq* category is used as the positive class.
- lokay-m: data from the flat-format Enron data set³. The *tw_commercial_group* folder is used as the positive class and the *corporate*, *articles*, *personal* and *enron_t_s* are used as the negative class.
- farmer-d: data from the Enron data set. The *logistics* folder is used as the positive class and the *tufco*, *wellhead* and *personal* are used as the negative class.

The method for data set characterization first requires the ranking of classifications provided by the Naive Bayes classifier. The Naive Bayes classifier is known to provide a good ranking of examples as stated in (Zhang and Su, 2004) and (Zadrozny and Elkan, 2001). The claim we make here is that if a classifier is able to produce a good ranking with the provided attribute list, then the data set is dependent on that list. This claim makes intuitive sense since a larger percentage of easily classified examples implies that the attribute list is a suitable representation for the classification task. Ranking can be accomplished by ordering the classified examples according to the value obtained from the following certainty equation:

$$cert(E_i) = 1 - \frac{score(C_2|\vec{A})}{score(C_{max}|\vec{A})} \quad (1)$$

where $score(C_{max}|\vec{A})$ is the value output by the Naive Bayes algorithm for the predicted class and $score(C_2|\vec{A})$ is the value of the next most likely class. These values are related to the conditional probabilities, but in no way approximate them. The computation of the right hand side of Equation 1 results in a certainty value, $cert(E_i)$, between 0 and 1. Once the examples are ranked, accuracy can be computed on subsets of more and more certain classifications by progressively discarding low certainty examples. If the ranking procedure is successful, accuracy will increase as the low certainty examples are discarded. This relationship between ranking and accuracy can be described by the following set of equations:

$$P(E_1|\vec{A}) > P(E_2|\vec{A}) > \dots > P(E_n|\vec{A}) \quad (2a)$$

$$cert(E_1) > cert(E_2) > \dots > cert(E_n) \quad (2b)$$

$$E[acc([E_1])] > E[acc([E_1, E_2])] > \dots > E[acc([E_1, E_2, \dots, E_n])] \quad (2c)$$

³http://www.cs.umass.edu/~ronb/enron_dataset.htm

where $P(E_i|\vec{A})$ is the true probability of correctly classifying example E_i ; $cert(E_i)$ is the certainty of classification of example E_i ; and $E[acc(S)]$ is the expected accuracy achieved using the set of examples S . These equations can be interpreted as follows: if the certainty measure correctly ranks the examples, then the ranking should correspond to an ordering of the probability of correct classification. If this ranking of probabilities occurs, then the relationship expressed in Equation 2c is true. As a result of this equation being true, the accuracy on smaller and smaller sets of high confidence examples is expected to increase. We can then use the percentage of examples that were used to compute high accuracy values to reward a particular attribute list. For example, if an attribute list is a suitable representation for a data set, then it will have few low certainty examples and therefore use a larger percentage of examples in achieving a high accuracy. This relationship is described graphically in Figure 1 where accuracy is shown to increase with certainty as the set of examples on which it is calculated becomes smaller and smaller (coverage decreases as 1-coverage increases). We then develop a measure to characterize the

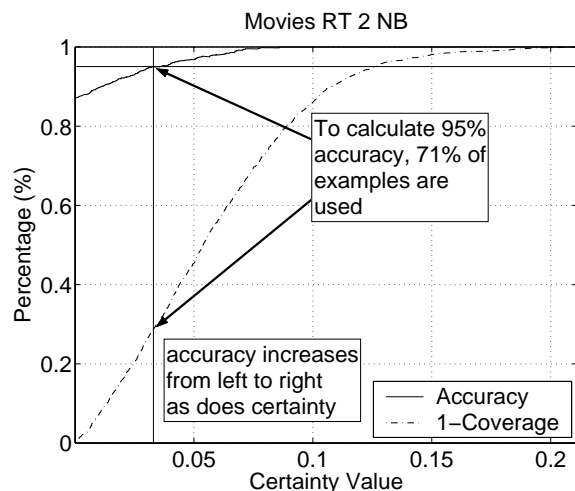


Figure 1: Annotated characterization curve

data set by computing the result of a weighted sum. The sum assigns more weight to the high certainty sets which result in high accuracy. The formula is given by Equation 3.

$$dep(E_i) = \sum_{i=1}^{20} w_i \cdot P_{(100-5i)} \quad (3)$$

where P_x is the percent of the total number of documents used to calculate an accuracy value of x and $w_{i+1} = \frac{w_i}{2}$. This approach is similar to the *area under the curve* (AUC) measure applied in (Zhang and Su, 2004), except the different areas are weighted to give more importance

Data set	> 80%		> 85%		> 90%		> 95%		overall	
	NE	RT	NE	RT	NE	RT	NE	RT	NE	RT
Hockey	100	93	100	70	100	5	86	0	91.0	78.0
Movies	93	100	73	100	55	92	35	71	78.3	86.9
Rt-earn	85	100	68	100	48	91	18	66	72.9	87.3
Rt-acq	100	100	97	100	81	100	54	100	80.4	97.3
lokay-m	100	99	100	70	91	10	64	6	85.3	79.1
farmer-d	67	88	48	68	36	29	29	20	68.8	76.2

Table 1: Percentage of Examples in Accuracy Range for NB Classifier

Data set	score		Dependency		Overall
	NE	RT	NE	RT	
Hockey	0.897	0.185	high	none	1
Movies	0.573	0.839	med	high	2
Rt-earn	0.449	0.794	none	high	2
Rt-acq	0.745	1.000	low	high	2
lokay-m	0.777	0.254	high	low	1
farmer-d	0.460	0.398	low	low	4

Table 2: Characterization of data sets with Naive Bayes

to the high accuracy subsets of ranked data. By weighting the points in this manner two situations are avoided:

- Low overall accuracy curves producing high scores
- High overall accuracy curves producing low scores

Both of these situations are possible since when evaluating rankings with AUC the values are normalized to the best and worst possible curves given the overall accuracy. The results of the data set characterization, on attribute lists constructed from named entities (NE) and regular terms (RT), are provided in Table 1 and Table 2.

The first table lists the percentage of examples in each subset for each attribute type, i.e. the largest percentage of examples used in calculating an accuracy of at least x percent (only the top four points are listed for each). The second table provides the overall score as computed by Equation 3, along with a label. The label (high, med, low, none) comes from a discretization of the range of possible values into four equal sets. The last column compares the values for each and determines which attribute list best characterizes the data set. The value corresponds to one of the following: (1) Named entity dependent; (2) Regular term dependent; (3) Dependent on both named entities and regular terms; and (4) Independent of named entities and regular terms. The results show that the *Hockey* and the *lokay-m* data set are named entity dependent. For the *Hockey* data set, it is apparent that regular terms are not of much use since both classes deal with the same general subject matter. The best way to discriminate between the two classes is with attributes such as the names of

teams, players, administrators, and cities. For the *lokay-m* data set people and organization names also play an important role in deciding which class label to assign. The other data sets, with the exception of *farmer-d* are found to be regular term dependent. With the *farmer-d* dataset, the results for regular terms are poor since the different folders deal with similar topics and often with an implied context. The Reuters data sets are all regular term dependent. This fact is in agreement with the research conducted by (Bekkerman et al., 2001). The work suggests that since the data was hand-labeled, the underlying classification process is keyword-based. The same is the case for the *Movies* data set where keywords appear regularly enough to help in the classification. Named entities also occur very frequently, but it is usually in combination with the more statistically significant keywords. The *farmer-d* data set is not well represented by either attribute type and would require a more advanced form of indexing.

3 Effect of Named Entities

Given the characterizations found in the previous section, we want investigate the behaviour of named entities in classification tasks. The characterizations will allow us to justify and target the evaluations performed in this section. For example, when we evaluate different representations for named entities, we will only test them on named entity dependent data sets because it is on these data sets where different representations would have an effect. In this section we are interested in three different areas where named entities could potentially impact classification accuracy. These areas are:

1. The relative frequency between named entities and regular terms
2. The different possible representations for named entities
3. The greater dependence on time for named entities than for regular terms

Each of these areas of interest will be addressed in turn with specially crafted experiments to evaluate them. These experiments and associated results are presented and discussed in the remainder of this section.

3.1 Attribute Frequency

It is well-known that attribute frequency plays a significant role in the training of classifiers by machine learning algorithms. In this experiment, we will study the effect that named entity frequency has on the learned classifier, given that their frequency is usually lower than that of regular terms. For this task, we will use regular term dependent and named entity dependent data sets and evaluate them using three different attribute lists: (1) regular

Data set	NE	RT	Combined
Movies	78.3	86.9	86.6
Rt-earn	72.9	87.3	87.2
Rt-acq	80.4	97.3	97.5
Hockey	91.0	78.0	82.6
lokay-m	85.3	79.1	81.8

Table 3: Accuracy Achieved with NB Classifier for Different Attribute Types

Dataset	NE	RT	Combined
Movies	64.1	88.7	88.8
Rt-earn	76.6	97.9	97.5
Rt-acq	85.4	96.4	96.2
Hockey	92.7	91.5	91.3
lokay-m	79.6	83.3	83.4

Table 4: Accuracy Achieved with SVM Classifier for Different Attribute Types

terms; (2) named entities; and (3) a combined list. The data from each data set is split into two equally sized sets, one for training and one for testing. The belief is that regardless of the task, regular terms will govern the performance of the combined list since they are more frequent.

In Tables 3 and 4 we present the results obtained on the selected data sets using a Naive Bayes classifier and a Support Vector Machine classifier. The results show that for both the Naive Bayes and SVM classifiers, our hypothesis is true. The performance of the combined attribute list is closer to the result obtained for the regular term classifier. In other words, the presence of named entities in the combined list makes no difference in terms of classification accuracy even for the tasks deemed to be named entity dependent.

3.2 Attribute Representation

For the named entity dependent tasks, it is necessary to evaluate whether there is any benefit to representing named entities by their entire description. The alternative is to represent them by their component words (NE-BOW) which has the benefit of simplicity and a potential reduction in the number of features. The hypothesis for this experiment is, however, that attribute representation for named entities will have an effect on classification accuracy.

For this purpose, only the named entity dependent data sets are used for experimentation. The data from each data set is split into two equally sized sets, one for training and one for testing. The results of the experiments on the two named entity dependent data sets are provided in Tables 5 and 6 for the two classification algorithms being considered.

Data set	NE	NE-BOW
Hockey	91.0	91.8
lokay-m	85.3	78.2

Table 5: Named Entity Representation Comparison for NB Classifier

Data set	NE	NE-BOW
Hockey	92.7	93.1
lokay-m	79.6	75.2

Table 6: Named Entity Representation Comparison for SVM Classifier

From these results, we note that for the one data set, *lokay-m*, the choice of representation has an effect. This effect was present in both the results for Naive Bayes and Support Vector Machines. The reason for why *this* data set was affected and not the other can be explained by considering the nature of the respective tasks. In the *Hockey* data set, the named entities present are often unambiguous given the restricted context of the task. For example, the name “Martin Brodeur” is unambiguous if represented by the term “brodeur” since there is only one player with this name. For the Enron data set, however, the task is much more general in nature and named entities overlap if represented by single-word tokens.

3.3 Attribute Time Dependence

The dependence of an attribute on time means that it could potentially negatively affect the performance of a classifier on unseen data. The performance can decrease over time if the testing data changes significantly from the data on which the classifier was trained. It is necessary to investigate how named entities behave in relation to time in order to determine if their behaviour is any different from that of regular terms. In this section, we first present a discussion about the causes of accuracy degradation over time. We present three different reasons for accuracy loss: (1) Overfitting; (2) Concept drift; and (3) Novelty introduction. We will then evaluate the effect of time on the different types of attributes. Each data set is split into four equal parts D_i where $i = 1, 2, 3, 4$. The data for each data set is ordered in time such that the data in D_1 occurs before the data in D_2 , etc. The evaluation will be performed by considering three different scenarios:

1. One classifier: A classifier trained on the earliest set of data and tested on later sets of unseen data. For example, a classifier trained on data D_1 is evaluated on sets D_2, D_3, D_4 .
2. Retrained classifier: A classifier trained on a set of data and tested on the data set occurring immedi-

ately after it. For example, a classifier trained on data D_1 is tested on D_2 and one trained on D_2 is tested on D_3 , etc.

3. Attribute density: The density of attributes will be compared to see if named entities occurrences tend to decrease over time more quickly than regular terms. Attribute density is defined as the total number of attribute occurrences, m , divided by the total number of possible occurrences $n \times d$, where n is the number of attributes and d is the number of documents.

Such measures are described in (Klinkenberg, 1999) as indicators of concept drift. This discussion will allow us to argue about the governing principle behind named entity classification accuracy decay.

3.3.1 Classifier Performance Decay

As mentioned, we will consider the three different causes of classifier performance decay, namely that of overfitting, concept drift and novelty introduction. The purpose of this discussion is to briefly define each cause, for the purpose of determining which has the greatest impact on named entities. For overfitting, we argue that it can be defined as being related to whether or not a trained classifier is overly general or overly specific. The concept of overfitting does not apply for named entities since there is no distinction for the notions of general and specific. For regular terms, overfitting is possible because such notions are valid. For example, the term “banana” is the generalization of the terms “green banana” and “yellow banana.” Within the context of named entities, however, each term stands on its own. Concept drift on the other hand, is defined as when the definition of a concept changes over time. The meaning of the term “hockey” may change from that of “ice hockey” to that of “field hockey” depending on whether it is winter or summer. In other words, it is when an attribute appears in the training and test sets, but its meaning has changed since the classifier was trained. Concept drift is also associated with the problem of when an attribute does not occur in the testing set. Novelty introduction refers to when an attribute appears in the testing, but was previously unseen in the training set. Such attribute behaviour can degrade performance because the best attributes for classification were not available in the training phase.

3.3.2 One Classifier

The results presented in Table 7 report the performance of a classifier trained on the set D_1 and tested on the sets D_i where $i = 1, 2, 3, 4$. In the first column, the training accuracy obtained is shown and in subsequent columns the difference in accuracy is given for the other data sets. An average difference is reported in the last column of

Data set	NE				RT			
	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4
Movies	92.5	-10.8	-13.3	-19.4	94.3	-6.8	-8.5	-7.2
Rt-earn	74.3	-4.6	-10.1	-5.7	88.8	-2.0	-6.1	-1.1
Rt-acq	82.1	-6.8	-10.7	-7.2	95.9	-1.4	-1.5	+0.4
farmer-d	72.0	-3.9	-8.9	-11.7	82.3	-7.1	-1.3	-7.2
Hockey	93.0	-3.5	+0.6	-6.3	81.4	-2.1	+2.3	-11.3
lokay-m	85.9	-7.6	-3.0	-0.9	80.6	-4.7	-7.0	-0.8
Average	n/a	-6.2	-7.6	-8.5	n/a	-4.0	-3.7	-4.5

Table 7: Decay in NB classifier accuracy over time

Data set	NE		RT	
	D_3	D_4	D_3	D_4
Movies	-0.7	+6.1	+1.3	-2.4
Rt-earn	+0.6	-0.4	-2.3	+0.4
Rt-acq	+2.3	-2.1	+0.7	+1.1
farmer-d	+4.7	+16.0	-4.8	+2.3
Hockey	+1.5	-0.1	+1.4	-3.3
lokay-m	-4.6	-2.8	+3.8	+1.5
Average	+0.63	+2.78	+0.02	-0.07

Table 8: Accuracy for NB classifier if regularly retrained

the table. The results obtained with an SVM classifier are omitted since they agree with those of the Naive Bayes classifier. From these results, the accuracy of classifiers trained on named entities are observed to decay faster over time than those trained on regular terms. In other words, named entities are more sensitive in general than regular terms to the effects of time.

3.3.3 Retrained Classifier

Presented here are the results for the experiment where concept drift is expected to be less pronounced. As previously described, the training data for each experiment is the set of data that occurs immediately before it. As was the case with Table 7, the results in Table 8 report only the difference in classification accuracy for the sets D_3 and D_4 . The difference in this case, however, indicates the change in accuracy versus the value reported for the one classifier scenario. The values in this table serve to quantify the effect of classifier retraining.

From these results, named entities seem to benefit the most from the retraining since they were the ones to suffer the most from the aging of the classifier. This result is deduced from the average difference values reported in Table 8. However, the results show that a named entity dependent data set did not make any special use from the updated regular terms and the converse is also true. In other words, for a regular term dependent data set, named entities show an improvement with retraining, but it is not enough to make their accuracy higher than that of regular terms. The main result here is therefore that named entity

Data set	NE				RT			
	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4
Movies	1.01	0.81	0.72	0.72	3.30	3.16	3.20	3.03
Rt-earn	0.49	0.44	0.46	0.42	1.65	1.66	1.70	1.47
Rt-acq	0.49	0.44	0.46	0.42	1.65	1.66	1.70	1.47
farmer-d	1.66	1.95	1.62	1.35	2.36	2.53	2.19	2.44
Hockey	1.32	1.26	1.35	1.27	5.28	5.10	6.37	6.36
lokay-m	1.63	1.32	1.15	1.04	2.60	2.19	2.36	2.26

Table 9: Attribute density over time

trained classifiers can be potentially helped more so than regular term classifiers by classifier retraining. The benefit, although for these data sets it does not translate into better performance for the named entity dependent sets, could be observed in other cases where concept drift is more pronounced. The conclusion is that named entities are negatively impacted by time, but the degree to which this is true depends on the nature of the data set.

3.3.4 Attribute Density

The results for attribute density are displayed in Table 9. Similar to the other results for time analysis, the values obtained are presented in four sections to show how the attribute density changes over time. The density being reported is a percentage of attribute occurrences over the total possible number occurrences. The interesting details about the results in Table 9 are twofold. The first is how much more frequent regular terms are than named entities. The second is that these results support the earlier conclusion that named entity occurrences decay more noticeably than regular term occurrences. This fact is observed even for named entity dependent data sets. The fact that the attributes tend to disappear after a while supports the theory that it is concept drift and novelty introduction which are to blame for the decrease in the classification accuracy of these data sets.

4 Conclusions and Future Work

This paper presents several findings regarding the behaviour and role of named entities in text classification tasks. The first conclusion is that named entities are in fact useful in text classification. In such tasks, a general approach to the problem may not offer an appropriate solution based on the choice of classification algorithm and attribute representation. The data sets found to be named entity dependent in this paper were the ones where a context was well-defined as part of the classification task. To determine if named entities would be useful for classification, an example ranking technique is used. If the examples can be properly ranked, then the attribute list is a good representation for the data set. After correctly characterizing a classification task, it is possible to tailor the

representation in order to further improve performance. In a named entity dependent task, for example, it is beneficial on two levels to throw out regular terms: (1) accuracy is improved and (2) training and classification time is reduced as a result of the feature reduction. This result is in contrast to the findings in (Cooley, 1999) which report named entities to not be of any use in classification.

The second conclusion relates to the representation of named entities. The work presented in this paper shows that if named entities are useful, the method used to represent them can have an impact on classification performance. The extent of this effect depends on the nature of the named entities given the context of the classification task. If the entities are unique or unambiguous then their representation has little effect. If named entities overlap, however, then using multi-word attributes can help reduce this confusion.

The third and final conclusion of this paper is that named entities, more so than regular terms, are dependent on the effects of time in classification. In other words, if a classification task is dependent on named entities, then time must be considered more closely and it may be necessary to retrain the document classifier more frequently than for a regular term dependent task. Classifiers trained on named entities cannot be thought of as overfitting the data since there is no generalization possible for such terms. They are affected more due to the fact that they tend to come in and out of prominence over time. Hence, the attributes appearing the training set may not appear in the testing set. If new information appears in the testing set and it degrades performance significantly, then the classifier must be retrained to ensure a quality classification.

In terms of future work, there are several new techniques which could be developed based on the conclusions presented. The first would be to reduce the cost of data set characterization. Named entity tagging is a costly endeavour, requiring several steps including parsing and part-of-speech tagging. It may be possible to develop a heuristic to extract likely named entities, or at least the most prominent ones, and use these to characterize the data set. With such a heuristic, data sets could be tested for named entity dependence at low computational cost.

Another technique would be to develop a specialized named entity classification algorithm to be used in cases where named entities are important. Such an algorithm would take into consideration the properties of named entities as reported in this paper. Namely, named entities tend to occur in bunches, and come in and out of prominence at unpredictable time intervals. The difficulty with the Naive Bayes algorithm when such a behaviour is observed is that it puts too much importance on when the attribute does not appear (especially if it was prominent in the training set). An improvement to this algorithm,

when considering named entities, would be to ignore the conditional probability when an attribute is absent.

A final possible future work would be to investigate the characterization of a data set not only on named entities but on the subtypes of named entities. In other words, we would like to train separate classifiers using the names of people, organizations and locations. For this paper, this was not possible due to the fact that named entity recognition is still not done with a high enough accuracy. The attribute types were therefore lumped together, and some semantic information was lost. If separated, a data set could be said to be dependent on people names or location names for example.

5 Acknowledgments

The authors would like to acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and Communications and Information Technology Ontario (CITO).

References

- Quintin Armour. 2005. The Role of Named Entities in Text Classification. Master's thesis, University of Ottawa.
- Ron Bekkerman, Ran El-Yaniv, Naftali Tishby, and Yoav Winter. 2001. On feature distributional clustering for text categorization. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 146–153. ACM Press.
- Ron Bekkerman, Andrew McCallum, and Gary Huang. 2004. Automatic categorization of email into folders: Benchmark experiments on Enron and SRI corpora. Technical Report IR-418, University of Massachusetts.
- Maria Fernanda Caropreso, Stan Matwin, and Fabrizio Sebastiani. 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In Amita G. Chin, editor, *Text Databases and Document Management: Theory and Practice*, pages 78–102. Idea Group Publishing, Hershey, US.
- Chris Clifton and Robert Cooley. 1999. Topcat: Data mining for topic identification in a text corpus. In *Principles of Data Mining and Knowledge Discovery*, pages 174–183.
- R. Cooley. 1999. Classification of news stories using support vector machines. In *IJCAI '99 Workshop on Text Mining*, August.
- W. Bruce Croft, Howard R. Turtle, and David D. Lewis. 1991. The use of phrases and structured queries in information retrieval. In *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 32–45, New York, NY, USA. ACM Press.
- H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.
- Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. 2004. A multiple resampling method for learning from imbalances data sets. *Computational Intelligence*, 20(1):18–36, February.
- R. Klinkenberg. 1999. Learning drifting concepts with partial user feedback. *Beitrag zum Treffen der GI-Fachgruppe 1.1.3 Maschinelles Lernen (FGML-99)*, Perner, Petra and Fink, Volkmar (ed.).
- David D. Lewis. 1992. An evaluation of phrasal and clustered representations on a text categorization task. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–50. ACM Press.
- Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 609–616. Morgan Kaufmann Publishers Inc.
- Harry Zhang and Jiang Su. 2004. Naive bayesian classifiers for ranking. In *ECML*, pages 501–512.