Early text classification: a Naïve solution*

Hugo Jair Escalante
INAOE
Puebla, 72840, Mexico
hugojair@inaoep.mx

Manuel Montes y Gomez
INAOE
Puebla, 72840, Mexico
mmontesg@inaoep.mx

Luis Villaseñor Pineda
INAOE
Puebla, 72840, Mexico
villasen@inaoep.mx

Marcelo L. Errecalde
Universidad Nacional de San Luis
San Luis, D5700HHW, Argentina
merrecalde@gmail.com

Abstract

Text classification is a widely studied problem, and it can be considered *solved* for some domains and under certain circumstances. There are scenarios, however, that have received little or no attention at all, despite its relevance and applicability. One of such scenarios is early text classification, where one needs to know the category of a document by using partial information only. A document is processed as a sequence of terms, and the goal is to devise a method that can make predictions as fast as possible. The importance of this variant of the text classification problem is evident in domains like sexual predator detection, where one wants to identify an offender as early as possible. This paper analyzes the suitability of the standard naïve Bayes classifier for approaching this problem. Specifically, we assess its performance when classifying documents after seeing an increasingly number of terms. A simple modification to the standard naïve Bayes implementation allows us to make predictions with partial information. To the best of our knowledge Naïve Bayes has not been used for this purpose before. Throughout an extensive experimental evaluation we show the effectiveness of the classifier for early text classification. What is more, we show that this simple solution is very competitive when compared with state of the art methodologies that are more elaborated. We foresee our work will pave the way for the development of more effective early text classification techniques based in the naïve Bayes formulation.

*This work was supported by CONACyT grants No. CB-2014-241306 and PN-247870.

1 Introduction

Text classification is the task of assigning documents to its correct categories (Sebastiani, 2008). This is one of the most studied topics within natural language processing. Advances in the last two decades have made significant progress and nowadays the text classification problem is considered to be solved in some scenarios and under certain circumstances (e.g., news classification with plenty of data). There are, however, settings of the text classification problem that have received little attention despite the wide applicability they may have. One of such scenarios is that of *early text classification*, which deals with the development of predictive models that are capable of determining the class a document belongs to as soon as possible. A text is assumed to be processed sequentially, starting at the beginning of the document and reading input words one by one. It is desired to make predictions with as low information as possible.

The early text classification topic has received little attention in the community, and there exist only a few works that have approached similar scenarios (Dulac-Arnold et al., 2011) (please note that in this work the problem is not stated as one of early recognition). Despite its low popularity, this topic has a major potential in practical applications. For instance, consider the problem of detecting sexual predators in chat conversations. Here, the goal is to sequentially read a conversation and to determine as fast as possible whenever a sexual predator is involved; clearly, a detection using the whole conversation can only be used for forensics rather than for prevention. Other sample applications include,

Proceedings of NAACL-HLT 2016, pages 91–99,
San Diego, California, June 12-17, 2016.

any kind of conversation analysis that requires of a fast response, (e.g., cyber-bullying prevention, adaptive/intelligent answering systems); trending-topic discovery (e.g., analyzing comments on social networks and determining as soon as possible whenever a topic will become a trend); content filtering (e.g., filtering inappropriate/ilegal content in local networks), author profiling (e.g., knowing the age, gender or interest of a person by using as few written information as possible) etcetera.

This paper explores the suitability of one of the most popular methods for text classification, i.e., naïve Bayes (McCallum and Nigam, 1998; Sebastiani, 2008), to approach the early-classification setting: *early naïve Bayes*. Specifically, we evaluate the capabilities of this classifier to make predictions when *seeing* an increasing number of terms from documents. A simple modification to the standard naïve Bayes implementation allows us to make predictions with partial information. Despite its simplicity, the proposed extension obtains competitive performance in standard text classification tasks and in sexual predator detection. In fact we show that the proposed modification compares favorably with the only existing work that addresses a similar task. Hopefully, our work will motivate research on further extensions to this classifier for early text classification.

The remainder of this paper is organized as follows. Next section reviews related work on early text classification and on extensions to naïve Bayes to face closely related problems. Then, Section 3 describes naïve Bayes classifier and the modification we propose to make early predictions. Section 4 reports experimental results that show the effectiveness of the proposal. Section 5 presents conclusions and discusses future work directions.

2 Related work

This section reviews related work on both: early text classification and extensions to naïve Bayes to face similar problems.

2.1 Early text classification

To the best of our knowledge, the early text categorization problem has been approached only in (Dulac-Arnold et al., 2011); although the authors' main focus was not on making predictions earlier but on improving the classification performance with a *sequential reading approach*. In that work, the authors process documents in a sentence-level basis. Every time t, the authors read a sentence and attempt to determine the class of the document, where multi-label classification is allowed. They proposed a Markov decision process (MDP) to approach the problem, where two possible actions were allowed: read next sentence, or classify. Each sentence has to be represented by its *tfidf* representation and a classifier is trained to learn good/bad state-action pairs (10,000 examples were randomly generated) on a high-dimensional space.

The performance of their method was evaluated in standard text classification data sets. Although the performance of such method is competitive (it was compared to a SVM classifier), it remains unknown whether a much more simpler approach would be as effective as the complex procedure in (Dulac-Arnold et al., 2011). In Section 4 we compare the proposed extension of naïve Bayes with the previous work. We show our proposal is competitive in terms of performance, but also has the following advantages: it is scalable in the number of categories (the MDP evaluated every possible state after reading each sentence, ours simply adds probabilities); it is able to make predictions with as low information as no-word (using priors-only information, but the most important aspect is that it can make predictions at anytime); it process documents in a word-level basis (i.e., one word added at a time, while the MDP requires processing whole sentences); training is much more efficient (same training complexity as an standard naïve Bayes classifier, the MDP requires of high-complexity training procedures) and the resultant model is way more simple.

Although the early text classification problem has not been studied elsewhere, it is worth mentioning works that have approached related tasks. In (Denoyer et al., 2001), the authors propose a hidden Markov model (HMM) to classify passages within documents. The task is information retrieval and a document is considered as relevant or irrelevant (i.e. two classes) to a given category/query. The document is decomposed into passages, each of which is considered by the HMM as relevant or irrelevant to the classification. No attempt is made to per-

form classification early, although it is interesting that the proposed model is a generalization of the multinomial naïve Bayes we consider in this work (again, for the two-class whole-document classification problem).

In (Dulac-Arnold et al., 2012) the authors extend the MDP proposed for sequential text classification to deal with any other type of data. The formulation is almost the same as in (Dulac-Arnold et al., 2011), although this time the MDP can decide what feature to sample from the instance under analysis (i.e., there is no sequential input). Furthermore, the MDP is equipped with a mechanism that aims to minimize the number of features to use for classification. Clearly, this extended MDP is not applicable to the early text classification domain (words cannot be chosen from documents, they appear sequentially).

Summarizing, it is remarkable the little attention that early text classification has received so far, this may be due to the fact that not so many applications in the past required to cope with this problem. Nowadays, however, the *online status* of the world population, requires of technology that can anticipate the prediction of certain events with the goal of preventing undesired effects or, on the other hand, to act as fast as possible to take the leadership on information technology.

2.2 Extending naïve Bayes

Naïve Bayes has been used extensively in text mining and within machine learning in general, because of its high performance in several domains, several modifications and extensions have been proposed to augment the scope of the classifier. Related to our work, the following extensions have been reported in the literature:

- **Alleviating independence assumption of Naïve Bayes.** This is perhaps the most studied topic in terms of extending the mentioned classifier. The independence assumption may be too strong for some domains/applications, therefore, several works have been proposed that try to relax it. Most notably TAN (Friedman et al., 1997), AODE (Webb et al., 2005), and WANBIA (Zaidi et al., 2013) extensions have reported outstanding results. Nevertheless, the focus here is on relaxing the attribute independence assumption, and not on working with partial information. One should note, however, that this extended versions of naïve Bayes can be well suited for early text classification, as attribute-dependency information can help the algorithm to classify texts earlier.

- **Anytime naïve Bayes.** The goal of this type of extensions is to provide naïve Bayes with mechanisms that allow it to make predictions at anytime (Yang et al., 2007; Hui et al., 2009). This means that the algorithm has to be ready to provide a prediction under time constraints: the classifier can spent increasing amounts of time for doing inference, but it must provide an answer when requested; usually accuracy increases as more time is allowed. This type of methods is related to our proposal in that the system has to be ready to make predictions at anytime, however, the granularity of information processing is different: in anytime classification a whole instance is seen, whereas in early text classification, part of an instance is available.

- **Incremental naïve Bayes.** Refers to developing learning and inference mechanisms to allow the classifier be trained in an online learning setting (Alcobé, 2002; Klawonn and Angelov, 2006). That is, reading a sample (or batch of samples at a time), the model makes predictions for the incoming samples and then it is provided with the correct labels, next, model parameters have to be updated accordingly. This type of methods are related to our proposal in that partial information is processed incrementally, although one should note that information units are instances and not words/attributes.

- **Naïve Bayes for incomplete information.** These extensions aim at helping naïve Bayes to deal with missing information, usually, at the attribute level. For instance by equipping the classifiers with mechanisms to work under highly-sparse representations (e.g., in short text categorization) (Shen et al., 2009; Cabrera et al., 2013; He and Ding, 2007; Yuan et

al., 2012). These methods are mostly based on smoothing attribute-class probabilities and often use co-occurrence statistics. Although not dealing with early text classification, this type of methods are relevant because smoothing plays a key role when working with partial information (everything not seen so far has to be smoothed).

Summarizing, there have been many attempts to improve and extend naïve Bayes to be robust against several limitations, however, to the best of our knowledge, it has not been used for early text classification before. This is somewhat surprising given that, as shown in the next section, the naïve Bayes classifiers can naturally deal with partial information.

3 Early text classification with Naïve Bayes

This section describes the way we use naïve Bayes classifier for early text classification.

3.1 Naïve Bayes classifier

We first describe the standard naïve Bayes classifier. Consider a data set: $\mathcal{D} = (\mathbf{x}_i, y_i)_{\{1,\ldots,N\}}$ with N pairs of instances ($\mathbf{x}_i$) and labels (y_i) associated to a supervised classification problem. Assuming that $\mathbf{x}_i \in \mathbb{R}^q$ and $y_i \in C = \{1, \ldots, K\}$ we have a $K-$class classification problem with numeric[1] attributes.

Under the naïve Bayes classifier, the class for an unseen instance $\mathbf{x}_T = \langle x_{T,1}, \ldots, x_{T,q} \rangle$ is given by:

$$\hat{C} = \arg\max_{C_i} P(C_i|\mathbf{x}_T) \tag{1}$$

From Bayes' theorem it follows that the posterior probability above can be estimated as:

$$P(C_i|\mathbf{x}_T) = \frac{P(\mathbf{x}_T|C_i)P(C_i)}{P(\mathbf{x}_T)} \tag{2}$$

The denominator can be removed from Equation (1) as it does not affect the decision:

$$P(C_i|\mathbf{x}_T) \approx P(\mathbf{x}_T|C_i)P(C_i) \tag{3}$$

[1]One should note that in text classification we can transform any document to a numeric vector with the bag of words representation, i.e., a vector of length q, where q is the vocabulary size and each element of the vector indicates the relevance of a term for describing the content of the document.

The assumption of naïve Bayes is that the probability of occurrence of attributes of $\mathbf{x}_T$ is independent given its class, that is:

$$P(C_i|\mathbf{x}_T) \approx \prod_{j=1}^{q} P(x_{T,j}|C_i)P(C_i) \tag{4}$$

The maximum likelihood estimation for the prior of class C_i is given by:

$$\hat{P}(C_i) = \frac{|X_i|}{N} \tag{5}$$

where X_i is the set of all instances in $\mathcal{D}$ that are labeled with class C_i. Hence, the key of the naïve Bayes classifier lies in the estimation of $P(\mathbf{x}_T|C_i)$, or more precisely of $\prod_{j=1}^{q} P(x_{T,j}|C_i)$. Depending on the type of data (e.g., binary, discrete, or real) a different distribution may be assumed for computing $P(x_{T,j}|C_i)$ (e.g., Bernoulli, Multinomial, or Gaussian, respectively). In text classification one of the most effective implementations is based in the multinomial distribution, when documents are represented by its term-frequency representation (i.e., we know for each document, the number of times each term from the vocabulary occurs) (McCallum and Nigam, 1998; Kibriya et al., 2005). Accordingly, we focus in this implementation, this means we assume w.l.o.g.: $\mathbf{x}_i \in \mathbb{Z}_+^q$ (i.e. the representation of a document is a vector of frequency values / integers).

Assuming a multinomial distribution for the model we have that the maximum likelihood estimation for the term of interest is:

$$P(\mathbf{x}_T|C_i) \approx \prod_{j=1}^{q} \hat{P}(x_{T,j}|C_i)^{f_{j,T}} \tag{6}$$

where $f_{j,T}$ is the value of the j^{th} attribute in instance $\mathbf{x}_T$ (in text classification $f_{j,T}$ is the frequency of occurrence of the j^{th} term in document T), and

$$\hat{P}(x_{T,j}|C_i) = \frac{1 + F_{j,C_i}}{q + \sum_k^q F_{k,C_i}} \tag{7}$$

where F_{l,C_i} is the sum of values of the l^{th} attribute in documents of class C_i. The derivation from Equation (6) removes factorial terms that do not affect the final decision. For more details we refer the reader

to (McCallum and Nigam, 1998; Kibriya et al., 2005). In the description above we did not assume a text categorization problem because the same results apply to any type of (multinomial-distributed) attributes. In the following we use text-mining terminology, but we emphasize the description is generalizable to other problems.

3.2 Early Naïve Bayes

In early text classification we assume that during training we have full documents, therefore, the same training procedure as the standard naïve Bayes classifier is performed for estimating the necessary probabilities[2]. The difference comes at inference time: when classifying a new document we assume we read it in sequential order starting from the beginning (i.e. the first word from top to bottom and from left to right). W.l.o.g.[3], at time t we assume we have read the first $t-$terms in the document (i.e., one word is read at each time). Let d_T denote the document we want to classify, where it contains M_{d_T} words, then, $d_T = w_1, w_2, \ldots, w_{M_{d_T}}$.

We notice from Equations (5-7) that in fact we can make predictions for document d_T regardless the amount of information we have read from it: at time t we know that $d_T = w_1, \ldots, w_t$, therefore, we can generate a bag-of-words $\mathbf{x}_T$ representation for d_T as follows $\mathbf{x}_T = \langle \mathbf{x}_{T,1}, \ldots, \mathbf{x}_{T,q} \rangle$, where $\mathbf{x}_{T,j}$ indicates the frequency of occurrence of the j^{th} term in document d_T (i.e., a *tf* weighting scheme). Terms not occurring the d_T or not seen so far at time t are assigned values of $\mathbf{x}_{T,j} = 0$. With this representation we can use Equation (3) directly to classify the document. Actually, we can attempt to classify document d_T without having read any information! (i.e., with $t = 0$), of course the probability will be dominated by the priors, see Equation (5). Simply as this, we can use naïve Bayes to perform early classification.

We now briefly analyze what are the main components in play when making predictions early. At time t one can rewrite Equation (4) as:

$$P(C_i|\mathbf{x}_T) \approx P(C_i) \prod_{j:j\in d_T} P(x_{T,j}|C_i) \prod_{k:k\notin d_T} P(x_{T,k}|C_i) \quad (8)$$

the second product (over $j \in d_T$) accounts for the terms appearing in the document (probabilities are affected by the frequency of occurrence of such terms in d_T so far); the third product (on $k \notin d_T$) simply reduces to 1 (because of the exponent in Equation (6)). Therefore, for small values of t, the priors dominate the decision, as t increases the content of the document will dominate the other products. Therefore, the way these three components are estimated can be crucial for improving the performance of naïve Bayes in early classification.

Despite the simplicity of this early text classification approach, we will see in the next section that it compares favorably with a more complicated solution from the state of the art. We show its validity in a variety of problems. This paper motivates further work on extending this model for early text classification. For instance, one can define/modify adaptive priors that change as the value of t increases; we can implement the same idea with methods that take into account term-dependencies (see e.g., (Friedman et al., 1997; Webb et al., 2005; Zaidi et al., 2013)) in order to increase the predictive power of the classifier; also one can adopt advanced/alternative smoothing techniques to account for partial and missing information properly (Shen et al., 2009; Cabrera et al., 2013; He and Ding, 2007); as well as many other possibilities. The main goal of this paper is to show that naïve Bayes can be used for early text classification and that its performance is competitive with the single existing solution to this problem. We foresee our work will pave the way for development of a new type of models.

4 Experiments and results

For experimentation we considered the data sets described in Table 1. We considered three standard thematic text categorization tasks (also used in (Dulac-Arnold et al., 2011)) and a data set for sexual predator detection (Inches and Crestani, 2012). All of the data and our code will be made available under request for future comparisons. In the subsections below we provide details on each data set

[2] One may also train naïve Bayes with partial documents, however, in that case the probability estimates associated to the model are not reliable because they are obtained from reduced documents. In preliminary experiments we corroborated this fact.

[3] One should note that we can take steps of any length, instead of processing word-by-word.

and report the corresponding experimental results obtained with them.

Text categorization					
Data set	**Classes**	**Terms**	**Red.V.**	**Train**	**Test**
Reuters-8	8	23583	2483	5339	2333
20-Newsgroup	20	61188	6894	11269	7505
WebKB	4	7770	3727	2458	1709
Sexual predator detection					
SPD	2	155886	6770	6588	15329

Table 1: Data sets considered for experimentation. Red. V. is the number of terms when a reduced vocabulary is used.

Text data sets were processed as follows: stop words were removed, then stemming was applied, next the bag-of-words representation was obtained using the TMG toolbox, a term-frequency (*tf*) weighting scheme was used (Zeimpekis and Gallopoulos, 2006). All of the data were processed in Matlab^R. For most experiments we used reduced vocabularies, that is, we used only a subset of the most frequent words/terms (see column 4 in Table 1), we proceeded like this for efficiency, nevertheless we also report results with full-vocabularies in text categorization data sets.

In addition to the comparison to the state of the art, we considered a linear SVM classifier as baseline, since this is a *mandatory* baseline in text classification (Joachims, 2008; Sebastiani, 2008). SVM was used in early classification similarly as the naïve Bayes model: it was trained with complete documents, and for making predictions, the bag of words of a document up to time t is obtained and feeded to the SVM classifier. In preliminary experimentation we compared SVM with *tf* and *tfidf* weighting schemes, we report the performance of SVM with the latter scheme because we obtained better results with this configuration.

In all of our experiments we report the performance of the early text classifiers when varying the percentage of the words in test documents (same procedure as in (Dulac-Arnold et al., 2011)). Macro-average f_1 measure was used for multiclass text categorization problems and f_1 of the minority class (i.e., predators) for the sexual predator detection data set. Ideally, the performance of a good early text classifier should draw a curve close to the $y-axis$ (see figures below): i.e., better performance with less information. A different problem, not evaluated in this paper, is that of triggering a prediction whenever the classifier is sure about the class of a document. Please note, however, that simple triggering mechanisms can be derived for our proposed formulation, e.g., after seeing a predefined number of words, or when the difference between the most probable and the second most probable class exceeds a threshold, and so on.

4.1 Early text categorization

First we analyze the performance of early naïve Bayes on thematic text classification. The first three data sets from Table 1 were considered, these are widely used benchmark data sets for text categorization; standard training/testing partitions[4] were used. Results of this experiment are shown in Figure 1.

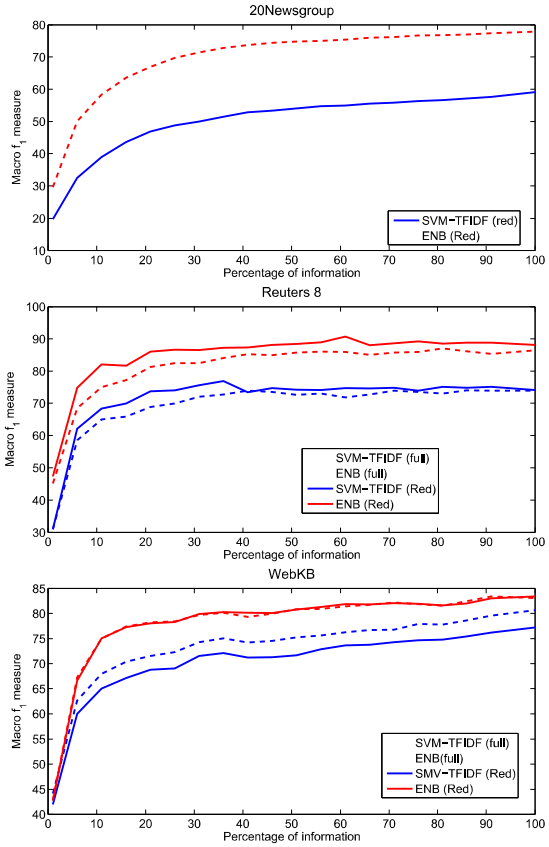

Figure 1: Early text classification on standard data sets.

It can be seen in the top plot that the early naïve Bayes (ENB hereafter) classifier outperforms considerably the SVM baseline for the 20Newsgrup data set. For both methods, the performance in-

[4] As reported in: http://web.ist.utl.pt/acardoso/datasets/

creased monotonically and, as expected, better performance was obtained when more information is considered.

The middle and bottom plots in Figure 1 show results for Reuters 8 and WebKB, respectively; in these plots we show the performance of both methods, ENB and SVM, and when using all of the vocabulary (*full*) and a reduced one (for 20Newsgrup data set we were not able to run an experiment with the full vocabulary in reasonable times). Regardless of the vocabulary used, ENB outperforms SVM. However, using the full vocabulary had opposed effects in the two data sets. In Reuters 8, using the whole vocabulary reduced the performance of both methods mainly when using less than 50% of information; in WebKB the performance of ENB is virtually the same, but the performance of SVM increased when using the full vocabulary. This can be due to the specific characteristics of the data. Finally, in the three data sets it is somewhat evident that the predictive performance of ENB presents low variations after processing about 50% of the texts.

4.2 Comparison with related work

In this section we compare the performance of naïve Bayes with the MDP introduced in (Dulac-Arnold et al., 2011) using the same data sets from the previous section. For this comparison we replicated the experiment reported by the authors of (Dulac-Arnold et al., 2011). For each of the data sets, we used different percentages, $\{1\%, 5\%, 10\%, 30\%, 50\%, 90\%\}$, of documents for the training set and the remainder for the test set (this was not our choice, but the setting proposed by the authors of the reference paper). Five runs were performed, in each run the documents for training were randomly chosen. Average results are shown in Figure 2. The results of ENB are shown as graphs, whereas for the reference method we report the single-best reported result (shown as markers, one per training set size). Please note that in (Dulac-Arnold et al., 2011) the authors optimized the parameters of their method, called STC, whereas we have used default implementation/parameters for ENB.

From Figure 2, it can be seen that the percentage of training documents used for learning the model affects considerably the performance of ENB. In all

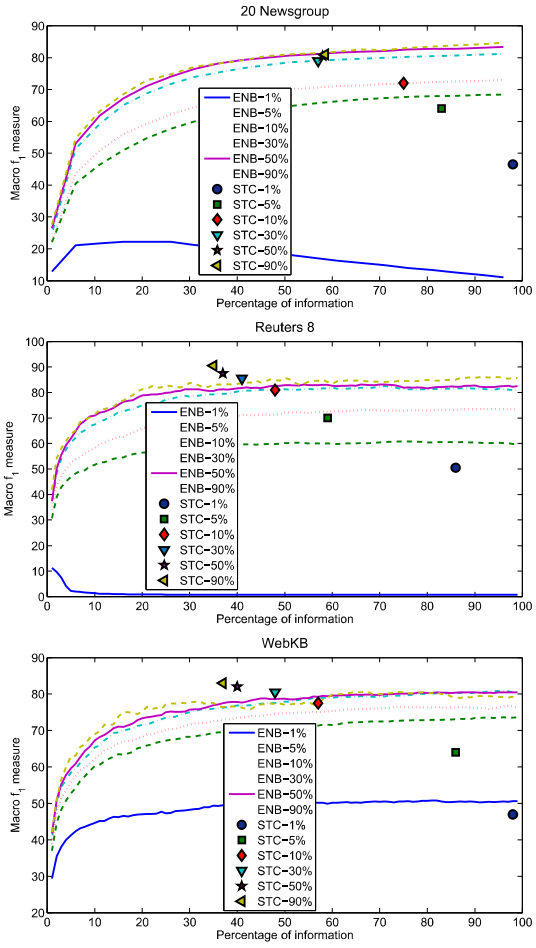

Figure 2: Comparison of ENB and the reference method STC.

three cases, using less than 30% of the samples for training results in low performance. This can be due to the fact that with small amounts of training documents, the estimated probabilities are not very representative of the classification task (and so, it is not convenient to estimate probabilities from partial information only). The best results were obtained when using 50% or 90% of instances for training the model. Also we can notice that the performance stabilizes after 40% of the information has been processed.

When comparing the ENB approach with the sequential text classification technique (STC) from (Dulac-Arnold et al., 2011), it can be seen that the MDP from the reference work and our ENB perform very similar (even when we only show best/optimized results for STC). This is a very interesting result: we obtained comparable performance

to a more complex model, with a much more simpler and efficient technique.

4.3 Sexual predator detection

We now evaluate the performance of ENB on the task of sexual predator detection. We used the development / test partitions of the data set used in the sexual predator competition from PAN'12 (Inches and Crestani, 2012), see Table 1. This corpus contains a large number of chat conversations, some of which include a sexual predator trying to approach a child[5]. The problem approached in the original competition was to identify sexual predators from many chat conversations. However, in this work, we approach the problem of detecting conversations with potential sexual predators in it. We proceeded in this way because the original task was one of forensic analysis: detect predators offline using all of the conversations in which they were involved (see (Villatoro-Tello et al., 2012) for our solution that obtained the best result in that challenge). Our ultimate goal, on the other hand, is to detect, as early as possible, conversations in which a sexual predator is involved, in such a way that sexual-attacks can be prevented and an alert for parents/police officers can be emitted. Based on our previous results from (Villatoro-Tello et al., 2012), and on the literature on non-thematic text classification we decided to represent chat conversations with 3-grams of characters (i.e., terms in this data set are sequences of 3-letters extracted from the training corpus); with this data set we used a reduced vocabulary and preprocessing processes described in (Villatoro-Tello et al., 2012). As suggested in (Inches and Crestani, 2012), for this experiment we report f_1 measure on the minority class (i.e., predators). Results of this experiment are shown in Figure 3.

On the one hand, we can see that this is a very difficult task, the performance of both models, SVM and ENB, is somewhat low, even when the whole information from documents is used (the highest performance is lower than 70% of f_1 measure). This is not a surprising result if we notice that this problem is highly imbalanced: the imbalance ratio for training and test partitions is of 12.1 and 9.56, respectively. Furthermore, the reduction of the vocabulary

[5]Police officers acted as children, predators are real.

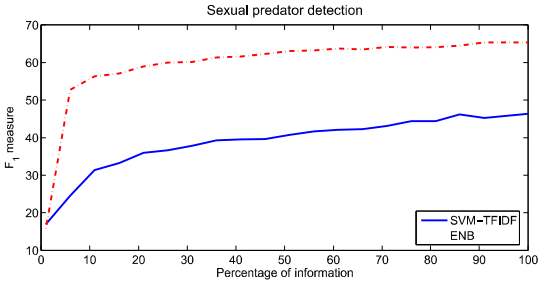

Figure 3: Early classification performance on detection of sexual predators.

may affect significantly this particular domain (the jargon used in chat conversations is quite diverse and rich). Despite the difficulty of the problem, we can see that again the ENB method outperforms the SVM model in most cases. Results shown in this section make evident the need of better methods for early text classification.

5 Conclusions

We described the use of naïve Bayes for early text classification. A minor modification to naïve Bayes allows us to make predictions using partial information. We show the effectiveness of this simple approach in three types of problems and compare its performance with the only existing state-of-the-art method. Our method compares favorably in terms of both effectiveness and earliness performance with the reference method, a much more complex model. Also, our method consistently outperformed an SVM baseline. Furthermore, we are the first in approaching the early classification of chat conversations for detecting sexual predators. Although results are encouraging, there is too much work to do yet. We foresee our work will pave the way for the development of more elaborated techniques based on naïve Bayes for early classification.

Future work is vast, for instance, exploiting research advances in extensions of naïve Bayes (see Section 2) for early text classification. Also, it is very important to develop spotting mechanisms that can be combined with the early naïve Bayes technique. Finally, theoretical analyses of the problem and the proposed method are very much needed.

References

J. R. Alcobé. 2002. Incremental learning of tree augmented naive bayes classifiers. In *IBERAMIA'02*, volume 2527 of *LNCS*, pages 32–41. Springer.

J. M. Cabrera, H. J. Escalante, and M. Montes y Gómez. 2013. Distributional term representations for short-text categorization. In *Proc. of CICLING*, volume 7817 of *LNCS*, pages 335–346. Springer.

L. Denoyer, H. Zaragoza, and P. Gallinari. 2001. Hmm–based passage models for document classification and ranking. In *Proc. of 23rd European Colloquium on Information Retrieval Research (ECIR'01)*.

G. Dulac-Arnold, L. Denoyer, and P. Gallinari. 2011. Text classification: A sequential reading approach. In *Advances in Information Retrieval, Proc. of 33rd European Conference on IR Research, (ECIR'11)*, volume 6611 of *LNCS*, pages 411–423. Springers.

G. Dulac-Arnold, L. Denoyer, P. Preux, and P. Gallinari. 2012. Sequential approaches for learning datum-wise sparse representations. *Machine Learning*, 89:87–122.

N. Friedman, D. Geiger, and M. Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning*, 29(2):131–163.

F. He and X. Ding. 2007. Improving naive bayes text classifier using smoothing methods. In *Proc. ECIR'07 Proceedings of the 29th European conference on IR research*, volume 4425 of *LNCS*, pages 703–707. Springer.

B. Hui, Y. Yang, and G. I. Webb. 2009. Anytime classification for a pool of instances. *Machine Learning*, 77:61–102.

G. Inches and F. Crestani. 2012. Overview of the international sexual predator identification competition at pan-2012. In *CEUR Workshop Proceedings, Working Notes for CLEF 2012 Conference*, volume 1178. CEUR.

T. Joachims. 2008. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98*, volume 1398 of *LNCS*, pages 137–142. Springer.

A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes. 2005. Multinomial naive bayes for text categorization revisited. In *AI 2004: Adv. Artificial Intelligence*, volume 3339 of *LNCS*, pages 488–499. Springer.

F. Klawonn and P. Angelov. 2006. Evolving extended naïve bayes classifiers. In *Proc. of Sixth IEEE International Conference on Data Mining Workshops, ICDM Workshops*, pages 643–647. IEEE.

A. McCallum and K. Nigam. 1998. A comparison of event models for naive bayes text classification. In *Proc. of AAAI/ICML-98 Workshop on Learning for Text Categorization*, pages 41–48. AAAI.

F. Sebastiani. 2008. Machine learning in automated text categorization. *ACM Computer Surveys*, 34(1):1–47.

D. Shen, J. Wu, B. Cao, J.T. Sun, Q. Yang, Z. Chen, and Y. Li. 2009. Exploiting term relationship to boost text classification. In *Proc. of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1637–1640. ACM.

E. Villatoro-Tello, A. Juarez-Gonzalez, H. J. Escalante, M. Montes y Gomez, and L. Villase nor Pineda. 2012. A two-step approach for effective detection of misbehaving users in chats. In *CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers*.

G. Webb, J. R. Boughton, and Z. Wang. 2005. Not so naive bayes: Aggregating one-dependence estimators. *Machine Learning*, 58:5–24.

Y. Yang, G. I. Webb, K. Korb, and K.M. Ting. 2007. Classifying under computational resource constraints: anytime classification using probabilistic estimators. *Machine Learning*, 69:35–53.

Q. Yuan, G. Cong, and N. M. Thalmann. 2012. Enhancing naive bayes with various smoothing methods for short text classification. In *Proc. of WWW Companion*.

N. A. Zaidi, J. Cerquides, M. J. Carman, and G. I. Webb. 2013. Alleviating naive bayes attribute independence assumption by attribute weighting. *Journal of Machine Learning Research*, 14:1947–1988.

D. Zeimpekis and E. Gallopoulos, 2006. *Grouping Multidimensional Data: Recent Advances in Clustering*, chapter TMG: A MATLAB toolbox for generating term-document matrices from text collections, pages 187–210. Springer.