Decision Support Systems 68 (2014) 26–38

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Sentiment analysis: Bayesian Ensemble Learning

E. Fersini *, E. Messina, F.A. Pozzi

University of Milano-Bicocca (DISCo), Viale Sarca, 336, 20126 Milan, Italy

ARTICLE INFO

Article history:
Received 26 February 2014
Received in revised form 15 October 2014
Accepted 17 October 2014
Available online 24 October 2014

Keywords:
Sentiment analysis
Polarity classification
Ensemble learning

ABSTRACT

The huge amount of textual data on the Web has grown in the last few years rapidly creating unique contents of massive dimension. In a decision making context, one of the most relevant tasks is polarity classification of a text source, which is usually performed through supervised learning methods. Most of the existing approaches select the best classification model leading to over-confident decisions that do not take into account the inherent uncertainty of the natural language. In this paper, we pursue the paradigm of ensemble learning to reduce the noise sensitivity related to language ambiguity and therefore to provide a more accurate prediction of polarity. The proposed ensemble method is based on Bayesian Model Averaging, where both uncertainty and reliability of each single model are taken into account. We address the classifier selection problem by proposing a greedy approach that evaluates the contribution of each model with respect to the ensemble. Experimental results on gold standard datasets show that the proposed approach outperforms both traditional classification and ensemble methods.

© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The amount of textual data available on the Web has proliferated in the last few years, affecting not only ICT industries but also business companies and public services. Governments grasp citizen-generated text to be aware of public opinions for policy establishment [1], market analysts take advantage of product/service reviews for strategic analysis and commercial planning [2] and e-learning systems capture the student sentiment to adapt teaching resources and methodologies [3]. Considering that the current technological progresses enable an efficient storing and retrieval of huge amount of data, the key point is now on methodologies for extracting information and creating knowledge from raw sources. In the context of Big Data, social media represent an emerging challenging sector: the natural language expressions of people can be easily reported through blogs and short text messages, rapidly creating unique contents of huge dimensions that must be efficiently and effectively analyzed to create actionable knowledge for decision making processes. The massive quantity of continuously contributing texts, which should be processed in real time in order to take informed decisions, calls for two main radical advancements: (1) a change of direction in the research, i.e. from data-constrained to data-enabled paradigm and (2) the convergence to a multi-disciplinary area that takes advantage of psychology, sociology, natural language processing and machine learning. A potential leverage towards novel decision support systems is represented by the transformation of qualitative data from user-generated contents to quantitative information when making decisions. In this context, the extraction of this subjective information is crucial to create structured and actionable knowledge to be used by either a decision support system or a decision maker. The knowledge embedded in user-generated contents has been shown to be of paramount importance from both user and company/organization points of view: people express opinions on any kind of topic in an unconstrained and unbiased environment, while corporations and institutions can gauge valuable information from raw sources. In order to make qualitative textual data effectively functional for decision processes, the quantification of "what people think" becomes a mandatory step. We approached this issue as a polarity detection task aimed at classifying texts as positive and negative. In particular, we propose a Bayesian Ensemble Learning approach that takes advantage of multiple classifiers to predict the sentiment orientation of user-generated contents. The contribution of the paper is two-fold, i.e. a novel ensemble learning methodology to improve the performance of the polarity classification task and a model selection strategy able to radically reduce the search space of candidate ensembles. The investigation results in an effective and efficient paradigm suitable for polarity detection both in well-formed scenarios (reviews) and social media environments (Twitter).

The paper is organized as follows. In Section 2, a literature review is reported to highlight the position of the paper with respect to the current state of the art. In Section 3, the proposed ensemble learning approach is described. The experimental settings to compare the proposed approach with the state-of-the-art techniques have been

* Corresponding author.
E-mail addresses: fersini@disco.unimib.it (E. Fersini), Messina@disco.unimib.it (E. Messina), federico.pozzi@disco.unimib.it (F.A. Pozzi).

http://dx.doi.org/10.1016/j.dss.2014.10.004
0167-9236/© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

reported in Section 4, while datasets and evaluation measures are described in Section 5. Computational results are discussed in Section 6 and time complexity analysis is reported in Section 7. Finally, in Section 8 conclusion and future research are outlined.

2. Literature review

Sentiment analysis (SA) addresses polarity classification, the task aimed at classifying texts as positive, negative or neutral, at different levels: document [4], sentence [5,6] and feature/aspect [7]. The state-of-the-art approaches for polarity classification can be divided into: unsupervised, semi-supervised and supervised. Most unsupervised learning approaches are usually composed of two phases: the first is the creation of a sentiment lexicon in an unsupervised manner and the second is the evaluation of the degree of positivity/negativity of a text unit via some function based on positive and negative indicators. A common approach employed by a number of researchers is to take into consideration some *seed words*, for which the polarity is already known. The relationships that exist between the set of seed words and the other co-occurring words are helpful to determine their polarities, as in [8] where the polarity of a given word or phrase is determined by considering the difference between the Pointwise Mutual Information of the phrase with the words "poor" and "excellent". Regarding the semi-supervised learning framework, most of the studies [9,10] address the polarity classification by expanding an initial set of sentiment words through synonyms and antonyms retrieved by thesauruses. Although there are relevant unsupervised and semi-supervised methods in the literature, most of the approaches for document-level polarity classification focus on supervised learning, thanks to their predictive power [7]. The common characteristic of these approaches concerns with the identification of the model which classifies the polarity of text sources with the highest accuracy as possible. However, none of the classification algorithms consistently perform better than others and there is no consensus regarding which methodology should be adopted for a given problem in a given domain. In order to overcome this limitation, an ensemble of different classifiers could lead to more robust and accurate classification. The idea behind ensemble mechanisms is to exploit the characteristics of several independent learners by combining them in order to achieve better performance than the best baseline classifier. Two necessary conditions should be satisfied to achieve a good ensemble: accuracy and prediction diversity. The state of the art about ensemble learning for SA basically comprises traditional methods such as Majority Voting, Bagging and Boosting (see [11] for a comprehensive study). Majority Voting is the most widely used ensemble technique, which is characterized by a set of "experts" that classifies the sentence and determines the final polarity by selecting the most popular label prediction to increase the accuracy but not explicitly addressing diversity. Further approaches aimed at accounting for diversity are represented by Bagging and Boosting. In Bagging, diversity is obtained by using bootstrapped replicas of the training data: different training data subsets are randomly drawn, with replacement, from the entire training dataset. Each training data subset (i.e. bag) is used to train a different baseline learner of the same type. Regarding Boosting, it incrementally builds an ensemble by training each new model to emphasize those instances that previous models misclassified. Although the presented approaches are widely used in sentiment analysis, they suffer from several limitations that the proposed paper intends to overcome:

- *Single learner generalizes worst than multiple models*. Many machine learning approaches have been investigated for sentiment classification purposes [12,4]. However, within the sentiment classification research field, there is no agreement on which methodology is better than others: one learner could perform better than others in respect of a given application domain, while a further approach could outperform the others when dealing with a given language or linguistic register. The uncertainty about which model represents the optimal one in different contexts has been overcome by introducing a novel ensemble learning approach able to exploit the potentials of several learners when predicting the sentiment orientation.
- *Ensembles assume independent and equally reliable models*. Classifiers enclosed in traditional ensemble learning approaches are assumed to be independent and equally trustworthy [13,11], which is not true in a real situation. For instance, consider several poor classifiers which make highly correlated mistakes predicting positive sentences as negative and a good classifier that correctly predicts the sentiment orientation. Assuming these classifiers as independent and equally reliable could lead to biased decisions. The proposed paper accounts for dependencies and accuracies of learners that would help to evaluate the contribution of each model in an ensemble and to smooth weak classifiers when making polarity predictions.
- *The search of the optimal ensemble comes with a cost*. One of the major challenges is concerned with online big data, where ensembles are attempting to come up with a reasonable trade-off between classification accuracy and computational time. Traditional state of the art approaches mainly focus on dealing with data and/or models to obtain the highest recognition performance, disregarding the computational complexity issue. A contribution of this paper is to derive not only an effective, but also an efficient methodology.
- *Lack of investigations across several domains*. Traditional ensemble approaches have shown their potential on predicting the sentiment orientation either on well-formed texts [14] or on noisy contents [13]. The investigation performed in this paper contributes not only on corroborating the strength of the proposed solution on well-written texts, but also to highlight its benefits on short and informal messages.

Starting from the idea proposed in [15], we overcome these limitations by developing a novel Bayesian Ensemble Learning approach, where the marginal predictive capability of each model is taken into account and a greedy selection strategy, based on backward elimination, is used to derive the optimal ensemble of classifiers.

3. Bayesian Ensemble Learning

3.1. Bayesian Model Averaging

The most important limitation of existing ensemble methods is that the models to be included in the composition have uniform distributed weights regardless of their reliability. However, the uncertainty left by data and models can be filtered by considering the Bayesian paradigm. In particular, all the possible models in the hypothesis space could be exploited by considering their marginal prediction capabilities and their reliabilities. Given a sentence s and a set C of independent classifiers, the probability of label $l(s)$ is estimated by Bayesian Model Averaging (BMA) as follows:

$$P(l(s)|C,\mathcal{D}) = \sum_{i \in C} P(l(s)|i,\mathcal{D})P(i|\mathcal{D}) \quad (1)$$

where $P(l(s)|i,\mathcal{D})$ is the marginal distribution of the label predicted by classifier i and $P(i|\mathcal{D})$ denotes the posterior probability of model i. The posterior $P(i|\mathcal{D})$ can be computed as:

$$P(i|\mathcal{D}) = \frac{P(\mathcal{D}|i)P(i)}{\sum_{j \in C} P(\mathcal{D}|j)P(j)} \quad (2)$$

where $P(i)$ is the prior probability of i and $P(\mathcal{D}|\cdot)$ is the model likelihood. In Eq. (2), $P(i)$ and $\sum_{j \in C} P(\mathcal{D}|j)P(j)$ are assumed to be a constant and

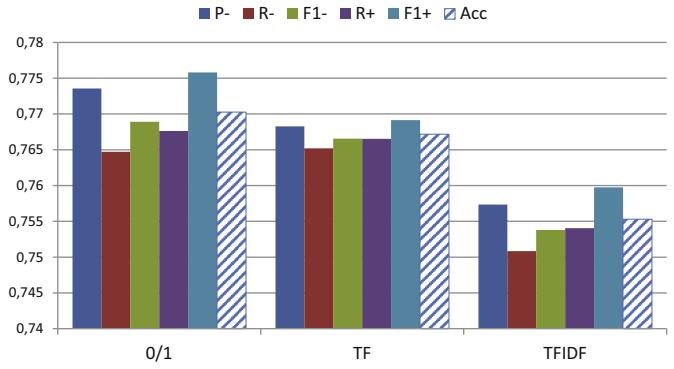

Fig. 1. Comparison of different weighting schemas on MovieData.

therefore can be omitted. Therefore, BMA assigns the optimal label $l^*(s)$ to s according to the following decision rule:

$$l^*(s) = \arg\max_{l(s)} P(l(s)|C,\mathcal{D}) = \sum_{i\in C} P(l(s)|i,\mathcal{D})P(i|\mathcal{D}) = \sum_{i\in C} P(l(s)|i,\mathcal{D})P(\mathcal{D}|i)P(i) = \sum_{i\in C} P(l(s)|i,\mathcal{D})P(\mathcal{D}|i). \quad (3)$$

The implicit measure $P(\mathcal{D}|i)$ can be easily replaced by an explicit estimate, known as F_1-measure, obtained during a preliminary evaluation of the classifiers i. In particular, by performing a cross validation each classifier can produce an averaged measure stating how well a learning machine generalizes to unseen data. Considering ϕ-folds for cross validating a classifier i, the measure $P(\mathcal{D}|i)$ can be approximated as

$$P(\mathcal{D}|i) \approx \frac{1}{\iota}\sum_{\iota=1}^{\phi} \frac{2 \times P_{i\iota}(\mathcal{D}) \times R_{i\iota}(\mathcal{D})}{P_{i\iota}(\mathcal{D}) + R_{i\iota}(\mathcal{D})} \quad (4)$$

where $P_{i\iota}(\mathcal{D})$ and $R_{i\iota}(\mathcal{D})$ denote precision and recall (Section 5) obtained by classifier i at fold ι. According to Eq. (3), we take into account the vote of each classifier by exploiting the prediction marginal instead of a 0/1 vote and we tune this *probabilistic claim* according to the ability of the classifier to fit the training data. This approach allows the uncertainty of each classifier to be taken into account, avoiding over-confident inferences.

3.2. Model selection strategy

A crucial issue of most ensemble methods is referred to the selection of the optimal set of models to be included in the ensemble. This is a combinatorial optimization problem over $\sum_{p=1}^{N} \frac{N!}{p!(N-p)!}$ possible solutions where N is the number of classifiers and p represents the dimension of each potential ensemble. Several metrics have been proposed in the literature to evaluate the contribution of classifiers to be included in the ensemble (see [16]). To the best of our knowledge these measures are not suitable for a Bayesian Ensemble, because they assume uniform weight distribution of classifiers. In this study, we propose a heuristic able to compute the discriminative marginal contribution that each classifier provides with respect to a given ensemble. In order to illustrate this strategy, consider a simple case with two classifiers named i and j. To evaluate the contribution (gain) that the classifier i gives with respect to j, we need to introduce two cases:

1 j incorrectly labels the sentence s, but i correctly tags it. This is the most important contribution of i to the voting mechanism and represents how much i is able to correct j's predictions;
2 Both i and j correctly label s. In this case, i corroborates the hypothesis provided by j to correctly label the sentence.

On the other hand, i could also bias the prediction in the following cases:

1 j correctly labels sentence s, but i incorrectly tags it. This is the most harmful contribution in a voting mechanism and represents how much i is able to negatively change the (correct) label provided by j.
2 Both i and j incorrectly label s. In this case, i corroborates the hypothesis provided by j leading to a double misclassification of s.

To formally represent the cases above, compute $P(i=1|j=0)$ as the number of instances correctly classified by i over the number of instances incorrectly classified by j (case 1) and $P(i=1|j=1)$ as the number of instances correctly classified by i over the number of instances correctly classified by j (case 2). Analogously, let $P(i=0|j=1)$ be the number of instances misclassified by i over the number of instances correctly classified by j (case 3) and $P(i=0|j=0)$ the number of instances misclassified by i over the number of instances misclassified also by j (case 4).

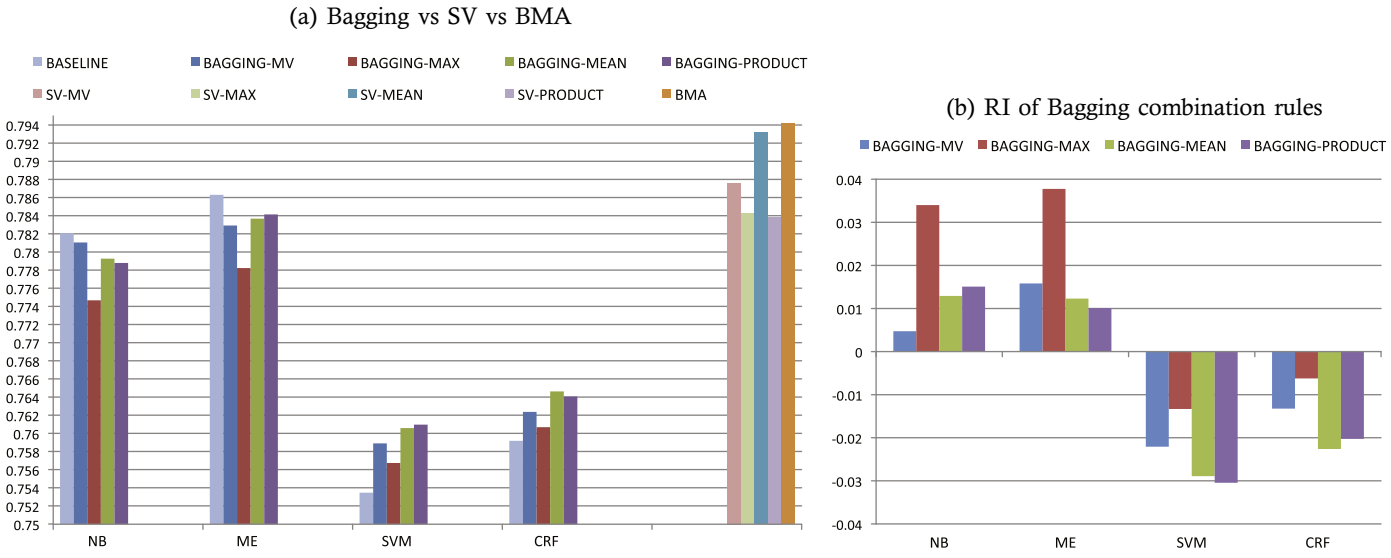

Fig. 2. Bagging performance on MovieData.

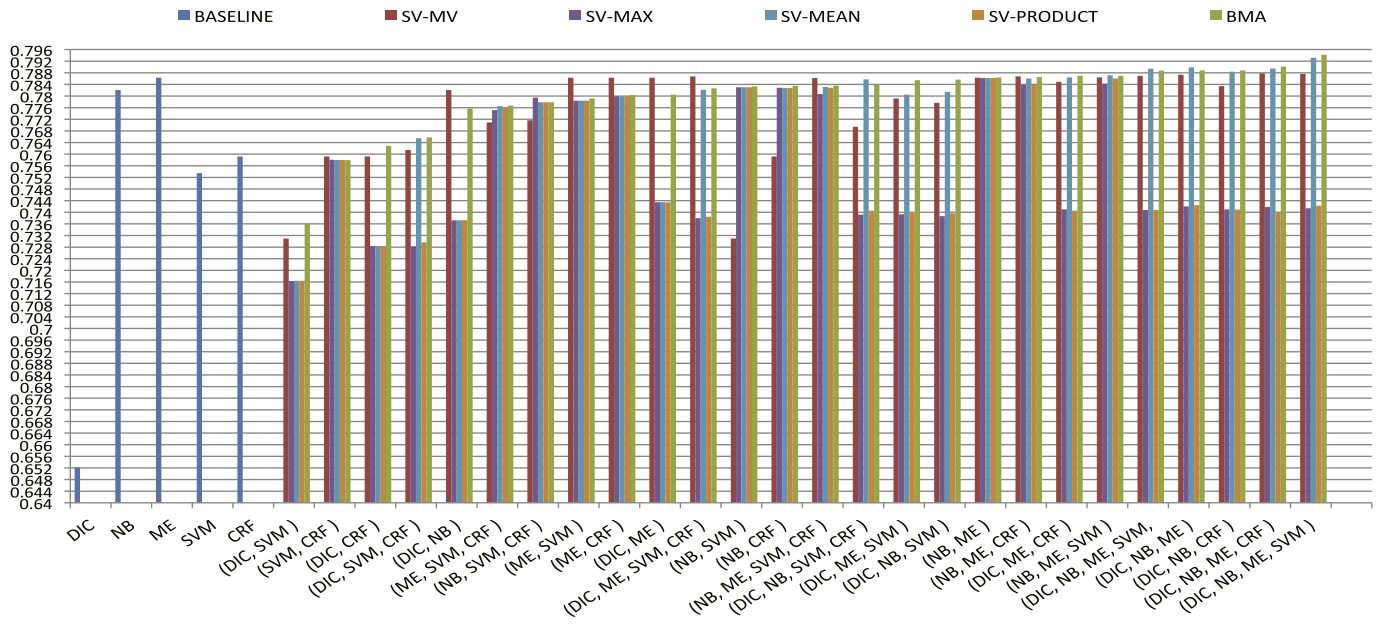

Fig. 3. Accuracy of baseline classifiers, SV and BMA on MovieData.

The contribution r_i^S of each classifier i belonging to a given ensemble $S \subseteq C$ can be estimated as:

$$r_i^S = \frac{\sum_{j \in \{S \setminus i\}} \sum_{q \in \{0,1\}} P(i=1|j=q)P(j=q)}{\sum_{j \in \{S \setminus i\}} \sum_{q \in \{0,1\}} P(i=0|j=q)P(j=q)} \quad (5)$$

where $P(j=q)$ is the prior of classifier j to either correctly or incorrectly predict labels. In particular, $P(j=1)$ denotes the percentage of correctly classified instances (i.e. accuracy), while $P(j=0)$ represents the rate of misclassified instances (i.e. error rate).

Once the contribution of each classifier has been computed, a further issue to be addressed concerns with the search strategy for determining the optimal ensemble composition. The greedy approaches presented in the literature can be distinguished, according to the search direction: *forward selection* [17] and *backward elimination* [18]. In *forward selection*, the initial ensemble S is an empty set. The algorithm iteratively adds to S the classifier $i \in \{C \setminus S\}$ that optimizes a given evaluation function. In *backward elimination*, the ensemble S initially contains all the classifiers of the complete set C and iteratively removes the classifier $i \in S$ that optimizes the evaluation function. The advantage of backward elimination is that recognizing irrelevant models is straightforward. Removing a relevant model from a complete set should cause a decline in the evaluation, while adding a relevant model to an incomplete set may have an immediate impact. According to this consideration, the proposed evaluation function r_i^S is included in a greedy strategy based on backward elimination: starting from an initial set $S = C$, the contribution r_i^S is iteratively computed excluding at each step the classifier that achieves the lowest r_i^S. The proposed strategy allows us to reduce the search space from $\sum_{p=1}^{n} \frac{n!}{p!(n-p)!}$ to $n-1$ potential candidates for determining the optimal ensemble, because at each step the classifier with the lowest r_i^S is disregarded until the smallest combination is achieved.

Another issue that concerns greedy selection is the stop condition related to the search process, i.e. how many models should be included in the final ensemble. The most common approach is to perform the search until all models have been removed from the ensemble and select the sub-ensemble with the lowest error on the evaluation set. Alternatively, other approaches select a fixed number of models. In this paper, we propose to perform a backward selection until a local maxima of average classifier contribution is achieved. In particular, the backward elimination will continue until the Average Classifier Contribution (ACC) of a sub-ensemble with respect to the parent ensemble will decrease. Indeed, when the average contribution decreases the parent ensemble corresponds to a local maximum and therefore is accepted as an optimal ensemble combination. More formally, an ensemble S is accepted as an optimal composition if the following condition is satisfied:

$$\frac{ACC(S)}{|S|} \geq \frac{ACC(S \setminus x)}{|S-1|} \quad (6)$$

where $ACC(S)$ is estimated as the average r_i^S over the classifiers belonging to the ensemble S. Note that the contribution of each classifier i is

Table 1
Model selection on MovieData.

Step	DIC	NB	ME	SVM	CRF	ACC	Accuracy
1	1.6618	1.9402	1.9294	1.6740	**1.6486**	1.7709	0.7887
2	**1.6662**	1.9747	2.0042	1.7486	–	**1.8485**	**0.7941**
3	–	1.5868	1.6073	**1.4891**	–	1.5611	0.7869
4	–	**1.1566**	1.2102	–	–	1.1835	0.7863

Bold-faced numbers denote the contribution r_i^S of the worst classifier that will be consequently removed from the ensemble.

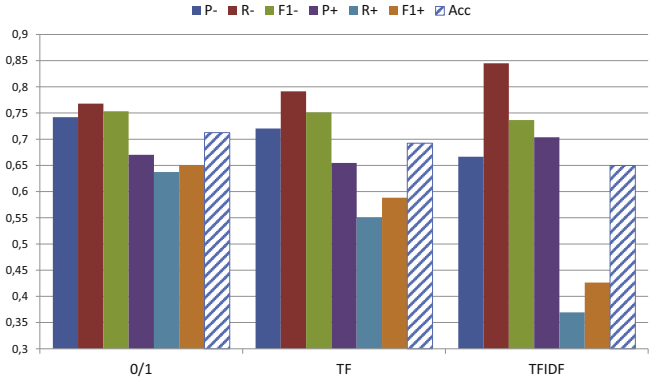

Fig. 4. Comparison of different weighting schemas on ProductData.

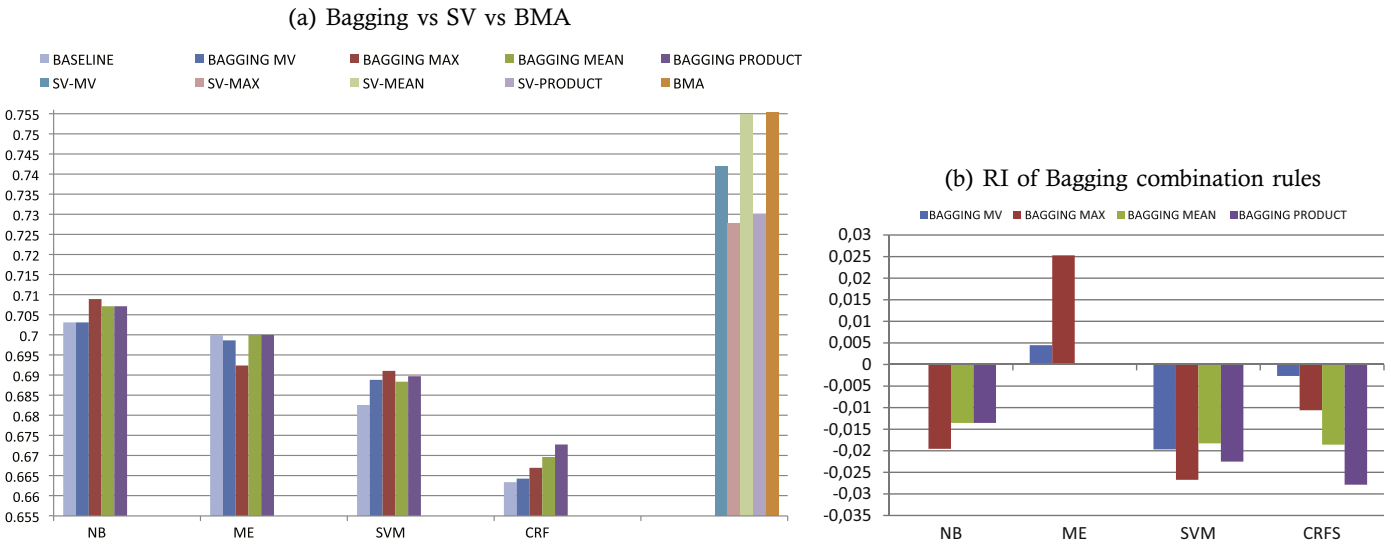

Fig. 5. Bagging performance on ProductData.

computed according to the ensemble S, that is iteratively updated once the worst classifier is removed. This leads to the definition of S characterized by a decreasing size ranging from $|S| = N, N-1,\ldots, 1$. In order to define the initial ensemble, the baseline classifiers in C have to show some level of dissimilarity. This can be achieved using models that belong to different families (i.e. generative, discriminative and large-margin models). As general remarks, this diversity helps ensembles to better capture different patterns of the natural language. Once this requirement is satisfied, the baseline classifiers to be enclosed in an ensemble can be arbitrarily selected.

4. Experimental investigation

In order to perform a comprehensive experimental evaluation, two issues need to be considered: (1) the identification of a suitable weighting schema for training the supervised classifiers and (2) the comparison of BMA with state of the art approaches, i.e. baseline classifiers and traditional ensembles.

4.1. Weighting schema

In order to derive the feature space used for learning, a vector space model has been adopted. A vector space model is an algebraic model for representing sentences as vectors of features, usually corresponding to terms. In our investigation, each sentence s is represented as a vector composed of terms for which a corresponding weight w can be computed. Concerning sentiment analysis, [4] pointed out that the overall sentiment of a text may not usually be expressed by multiple occurrences of the same terms. To verify this hypothesis in the context of ensemble learning, different weighting schemes have been investigated for computing w: Boolean (0/1), Term Frequency (TF) and Term Frequency Inverse Document Frequency (TF-IDF).

4.2. Baseline and ensemble classifiers

To evaluate the contribution of the proposed BMA, a comparison with the following baseline classifiers is proposed: Dictionary (DIC)

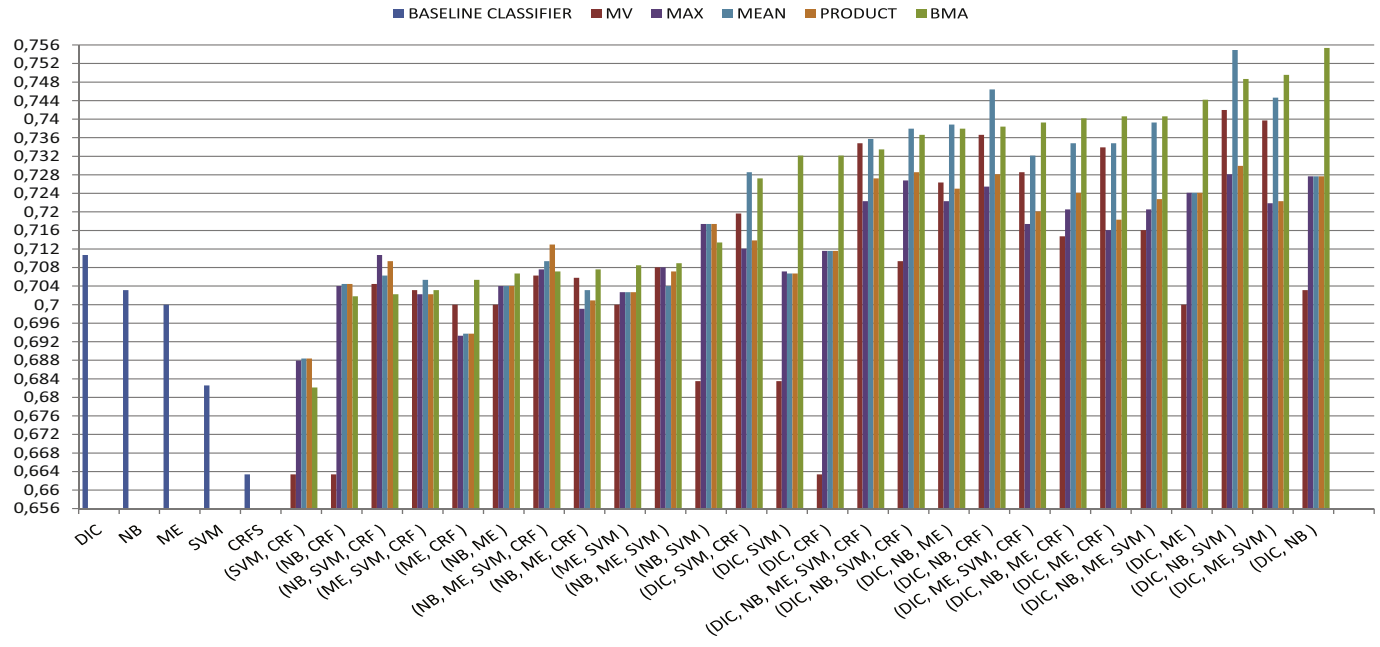

Fig. 6. Accuracy of baseline classifiers, SV and BMA on ProductData.

Table 2
Model selection on ProductData.

Step	DIC	NB	ME	SVM	CRF	ACC	Accuracy
1	2.1192	1.7118	1.6568	1.6244	**1.4389**	1.7102	0.7334
2	2.0972	1.6928	**1.6392**	1.6439	–	1.7603	0.7486
3	2.0903	1.9225	–	**1.6894**	–	1.9007	0.7495
4	2.1169	**2.0294**	–	–	–	**2.0731**	**0.7553**

Bold-faced numbers denote the contribution r_i^S of the worst classifier that will be consequently removed from the ensemble.

[19], Naïve Bayes (NB) [20], Support Vector Machines (SVM) [21], Maximum Entropy (ME) [22], and Conditional Random Fields (CRF) [23]. Concerning the traditional ensemble, the most widely used approaches have been considered:

- *Simple Voting* is the most popular ensemble system. The widely used technique is represented by *Majority Voting* (*MV*), which is characterized by a set of "experts" that classifies the sentence polarity by considering the vote of each classifier as equally important and determines the final polarity by selecting the most popular label prediction.
- *Bagging* [24] is another very popular ensemble technique also approached for polarity classification. The main goal of Bagging is to aggregate the multiple hypotheses generated by the same learner on different distributions of training data. Bagging assumes a dataset D and a learning system which trains a base classifier for each training set (i.e. bags) $b = 1, 2,\ldots, B$ sampled with replacement from D. The learning system is able to infer the label for each sentence of the testing set by aggregating over all the bags according to a *majority voting* decision rule. Considering that Bagging depends on a random sampling on the original dataset, 10 execution runs have been performed. Each run has enclosed 9 bags for inducing each classifier.

Simple Voting and Bagging can exploit combination rules based on the posterior probabilities to derive the final optimal label of a given sentence. The most popular decision rules that have been investigated into the experimental phase are *Maximum*, *Mean* and *Product* rules where the maximum, average and product of a posteriori probabilities of the classified sentence s among classifiers (or bags) are computed respectively.

5. Dataset and evaluation criteria

In order to evaluate and compare the proposed approach with the state of the art methodologies, several benchmarks have been considered. The first evaluation is based on *Review* datasets:

- *Sentence polarity dataset v1.0*, in the following *MovieData*. This dataset[1] [25] is composed of 10,662 positive and negative snippets of movie reviews extracted from Rotten Tomatoes.[2]
- *Finegrained Sentiment Dataset, Release 1*, in the following *ProductData*. This dataset[3] [26] relates to product reviews from Amazon.com. A reduction of instances has been performed to deal only with positive and negative opinions, resulting in a dataset composed of 1320 (≃58.84%) negative and 923 (≃41.16%) positive reviews.
- *Multi-Domain Sentiment Dataset*, in the following *ProductDataMD*. This dataset[4] [27] contains product reviews from Amazon.com. Reviews from categories "Music" and "Books" are studied separately.

The second type of evaluation is based on *Social* datasets collected from Twitter, i.e. *Gold Standard* benchmark [28]. The dataset has been distinguished in *Person* and *Movie* according to their main topic. Each set contains 1500 manually labeled Twitter data. A reduction of instances has been performed in order to deal only with positive and negative opinions: *Gold Standard Person* with 105 (≃26.44%) negative and 292 (≃73.56%) positive opinions, while *Gold Standard Movie* with 96 (≃18.6%) negative and 420 (≃81.4%) positive orientations. Concerning the evaluation criteria, a 10-fold cross validation has been adopted. The indices used for comparing the approaches are Precision (P), Recall (R) and F_1-measure [29], together with the overall Accuracy. Additionally, we have computed the Relative Improvement (RI) to test which Bagging combination rule improves or degrades the performance of the baseline classifiers:

$$\mathrm{RI} = \frac{ERR_{\mathrm{Bagging}} - ERR_{\mathrm{base}}}{ERR_{\mathrm{base}}} \tag{7}$$

where ERR_{base} and ERR_{Bagging} are the validation error rates for the baseline and Bagging classifiers, respectively. Considering that RI ranges in the interval $[-1, +\infty)$, a negative value indicates that the Bagging classifier has a decreasing error rate with respect to the baseline classifier.

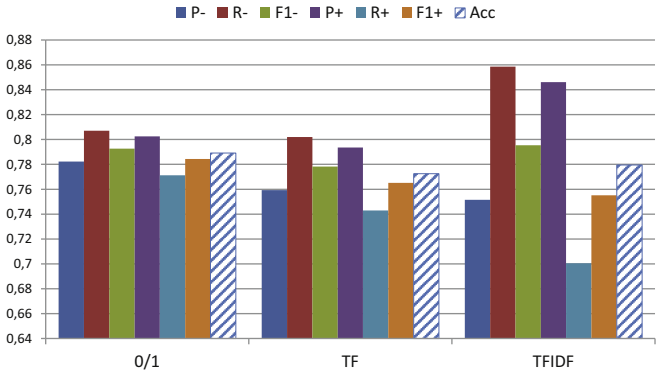

Fig. 7. Baseline models on ProductDataMD (Books) with different weighting schemas.

6. Computational results

6.1. MovieData

The first experimental investigation is aimed at determining the optimal weighting schema among Boolean, TF and TF-IDF by means of baseline classifiers (NB, SVM, ME, CRF). Fig. 1 shows the comparison between the three studied weighting schemas through the computation of Macro-averaging on MovieData. It suggests that TF-IDF is the worst performing weighting schema, while Boolean is slightly overperforming. In the following experimental results the Boolean weighting schema is assumed.

Bagging performance on MovieData, compared with the other approaches, is depicted in Fig. 2(a) (Simple Voting, denoted with SV-, and BMA bars are the rightmost blocks and are related to the best ensemble composition). A first interesting observation relates to the comparison between SV and Bagging: SV produces higher prediction accuracy than all the Bagging combination rules. Moreover, as highlighted in Fig. 2(b), Bagging is highly sensitive to the combination rule with respect to the classifier. While Bagging with SVM and CRF achieves high RI through the MEAN and PRODUCT combination rule, for NB and ME the base classifiers perform better than Bagging.

In order to compare baseline classifiers, SV and BMA, a summary of accuracy improvements is depicted in Fig. 3 (Bagging results are omitted because they are always worst than the other ensemble methods).

We can note that DIC obtains low performance on this dataset due to the dictionary composition: the opinion words belonging to the dictionary are concerned with products and do not fit the movie review of this dataset, leading to poor performance. Although DIC is the worst classifier, most of the promising ensembles contain the dictionary-based

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data/.
[2] http://www.rottentomatoes.com/.
[3] http://www.sics.se/people/oscar/datasets/.
[4] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/.

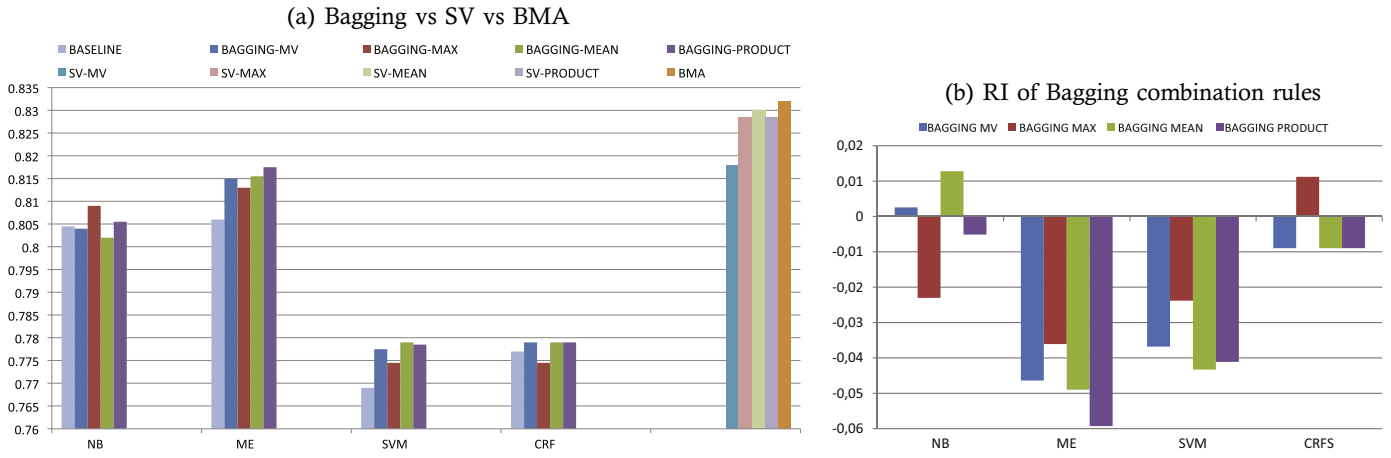

Fig. 8. Bagging performance on ProductData (Books).

classifier. This highlights that the best ensemble is not necessarily composed of the classifiers that individually obtain the highest accuracy: BMA is able to take into account the contribution of DIC by considering it's reliability and prediction, leading to robust and accurate polarity prediction. In particular, the optimal ensemble provided by BMA that includes DIC, NB, ME and SVM, outperforms SV achieving 79.41% of accuracy against 78.76% of MV, 74.13% by MAX, 79.31% by MEAN and 74.22% by PRODUCT. Concerning the proposed model selection strategy, the contribution of each classifier can be computed as shown in Table 1, where classifier contributions r_i^S, *Average Classifier Contribution* (denoted as ACC) and Accuracy are reported.

Starting from the initial set S = {DIC, NB, ME, SVM, CRF}, the contribution r_i^S is computed for each classifier. The model with the lowest contribution at the first step is CRF. Then, r_i^S is re-computed on the ensemble {S\CRF}, highlighting DIC as the classifier with the lowest contribution. At steps 3 and 4, the classifiers to be removed are SVM and NB respectively. It can be easy to note that the model selection strategy has radically reduced the search space and, thanks to the convergence criteria, the optimal BMA ensemble has been ensured at step 2.

6.2. ProductData

Regarding ProductData, Fig. 4 shows the comparison between the three studied weighting schemas through the computation of Macro-average on baseline classifiers. The comparison of the weighting schemas suggests that TF-IDF is again the worst performing weighting schema and Boolean is slightly over-performing.

Fig. 5 shows SV, Bagging and BMA accuracy on ProductData with the Boolean schema. While Bagging (MAX) with NB and SVM obtains higher performance than the corresponding baselines, for ME is the contrary.

The best Bagging for ProductData is evidently different from MovieData: while for MovieData the optimal Bagging was based on ME and PRODUCT combination rule, for ProductData is given by NB and MAX combination rule. This confirms that Bagging represents a weak ensemble technique.

The outperforming ensemble is obtained by BMA leading to a small set of experts. Our approach, based on DIC and NB, is able to achieve 75.53% of accuracy against 70.31% by MV, 72.76% by MAX, 72.76% by

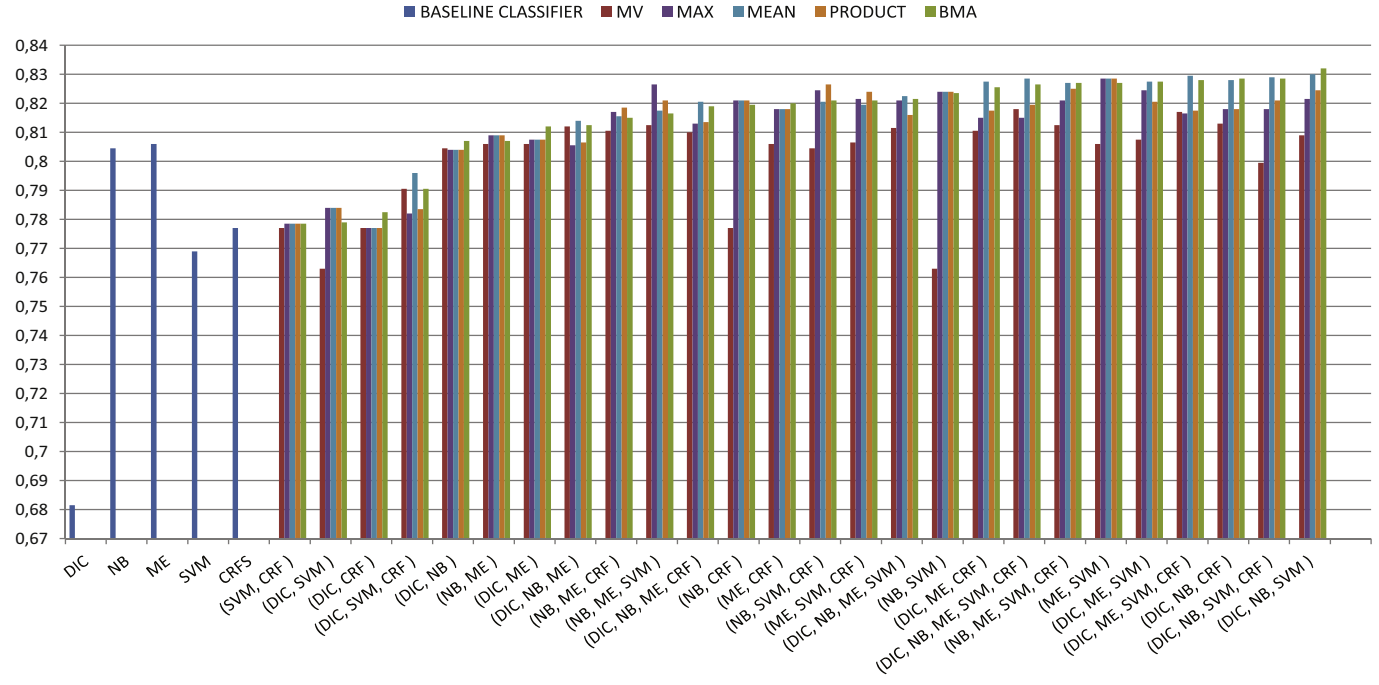

Fig. 9. Accuracy of baseline classifiers, SV and BMA on ProductDataMD (Books).

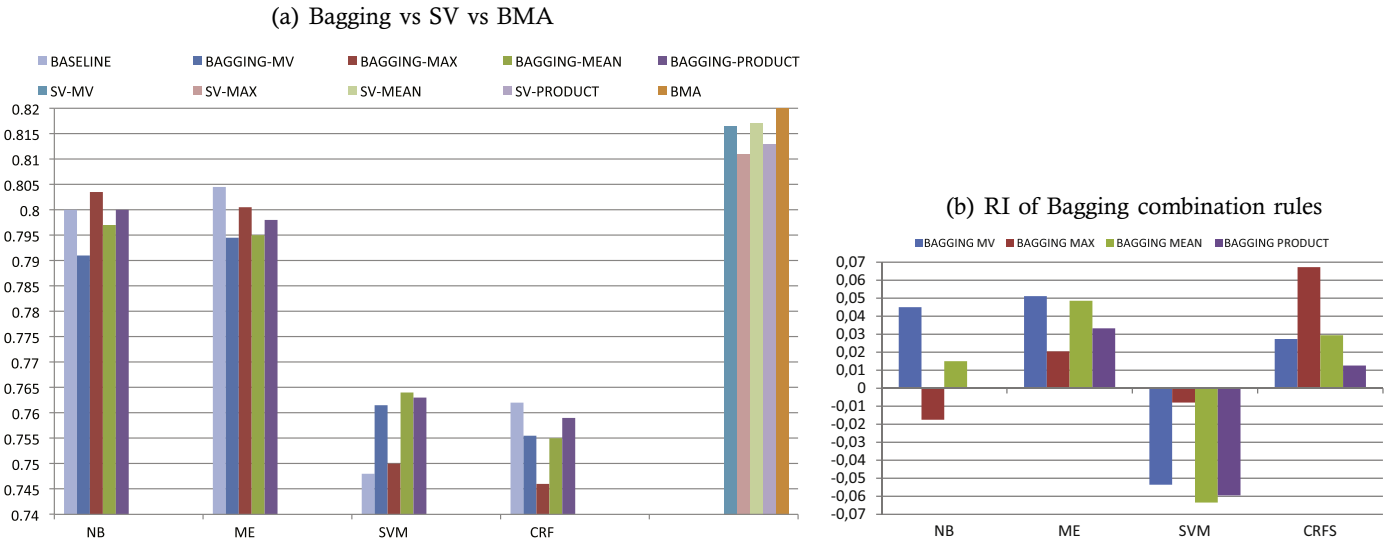

Fig. 10. Bagging performance on ProductDataMD (Music).

MEAN and 72.76% by PRODUCT obtained by Bagging compositions based on the same experts (all ensemble performance is depicted in Fig. 6). Also in this case, optimal ensemble can be easily identified through the model selection strategy. Starting from the initial ensemble S = {DIC, NB, ME, SVM, CRF}, the classifiers are sorted with respect to their contribution by computing Eq. (5). As shown in Table 2, the classifier with the lowest contribution at the first step is CRF. Then, r_i^S is recomputed on the ensemble {S\CRF}, highlighting ME as the classifier with the lowest contribution. At steps 3 and 4, the worst classifiers to be removed from the ensemble are SVM and NB respectively. Once more, the model selection strategy ensures the best BMA ensemble: the greedy search reveals the optimal ensemble at the last step when a local optimum of ACC is achieved.

6.3. ProductDataMD (Books)

Regarding ProductDataMD (Books), Fig. 7 shows the comparison between the three studied weighting schemas through the computation of Macro-averaging among performance of classifiers. Also in this case, Macro-averaging analysis confirms that the Boolean weighting schema is slightly over-performing in terms of accuracy.

Fig. 8 shows Bagging, SV and BMA performance achieved on ProductDataMD (Books). In Fig. 8(a) we can easily highlight that, also for this dataset, SV achieves better accuracy than any Bagging combination rule. As discussed for other datasets, Bagging does not guarantee robust performance: ME and SVM achieve high performance through the PRODUCT and MEAN combination rules, while NB and CRF provide poor results through the MV, MEAN and MAX combination rules.

In Fig. 9, it can be easy to note that the outperforming ensemble is obtained by BMA with DIC, NB and SVM (83.20% of accuracy). Other approaches obtain accuracy ranging from 80.90% of MV, 82.15% of MAX, 83% of MEAN and 82.45% of PRODUCT following a Bagging paradigm.

The model selection computation ensures, also for ProductDataMD (Books), the composition of the best BMA ensemble during the backward elimination.

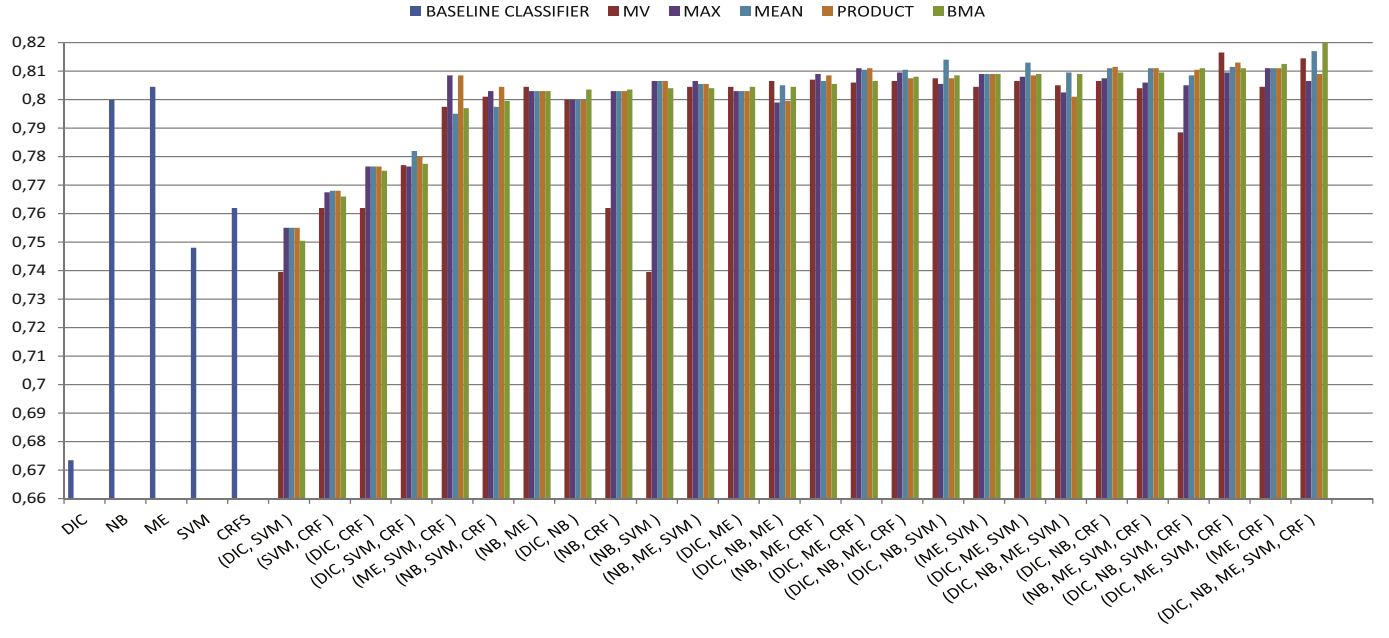

Fig. 11. Accuracy of baseline classifiers, SV and BMA on ProductDataMD (Music).

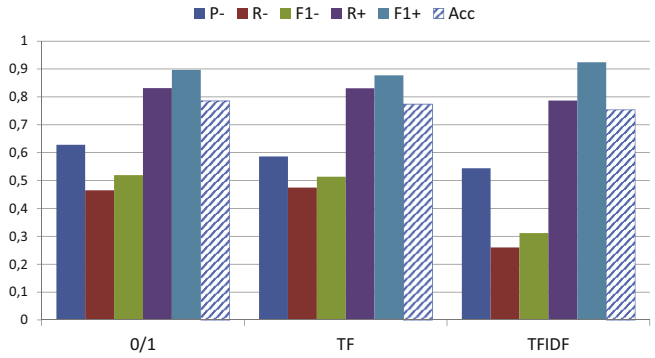

Fig. 12. Comparison of different weighting schemas on Gold Standard Person.

6.4. ProductDataMD (Music)

Regarding ProductDataMD (Music), Macro-averaging analysis confirms that Boolean weighting outperforms the other schemas. Fig. 10 shows Bagging versus the other approaches on ProductDataMD (Music).

In this case, Bagging-MEAN achieves high performance through SVM, while for NB, ME and CRF the base classifiers perform better than Bagging. Concerning the combination rules enclosed in Bagging, the results about error relative improvement show low-quality performance. Simple Voting is confirmed to be a promising approach to be compared with the proposed BMA. In order to compare baseline classifiers, SV and BMA, a summary of accuracy improvements is depicted in Fig. 11 (also in this case Bagging is omitted because it is always outperformed by the other ensembles).

Concerning ProductDataMD (Music) the accuracy measure reveals that the outperforming ensemble, obtained by the BMA paradigm, comprises all the baseline classifiers (NB, ME, SVM, CRF and DIC). The proposed approach is able to achieve 82% of accuracy against the SV systems (81.45% by MV, 80.65% by MAX, 81.7% by MEAN and 80.9% by PRODUCT). Also in this case, the model selection strategy ensures the composition of the best BMA ensemble during the greedy search, radically reducing the space search.

6.5. Gold Standard Person

Regarding Gold Standard Person, Fig. 12 shows the comparison between the three studied weighting schemas through the computation of Macro-averaging among performance of classifiers.

Macro-averaging analysis confirms that Boolean representation denotes a suitable weighting schema also in the context of social media. Fig. 13 shows Bagging versus the other approaches on Gold Standard Person. In this case, Bagging achieves high performance through NB considering all the combination rules, while for SVM only the MEAN and PRODUCT rules achieve the highest results through Bagging. For ME and CRF, the base classifiers perform better than Bagging. Concerning the combination rules enclosed in Bagging, the results about error relative improvement show low-quality performance. Simple Voting is confirmed as a promising approach to be compared with the proposed BMA.

In order to compare baseline classifiers, SV and BMA, a summary of accuracy improvements is depicted in Fig. 14 (also in this case Bagging is omitted because it is always outperformed by the other ensembles). Concerning Gold Standard Person, the accuracy measure reveals that the outperforming ensemble, obtained by the BMA paradigm, comprises only DIC and NB. The proposed approach is able to achieve 87.13% of accuracy against the SV systems (83.84% by MV, 85.64% by MAX, MEAN and PRODUCT).

6.6. Gold Standard Movie

Regarding Gold Standard Movie, analogous results have been obtained for the optimal weighting schema. The comparison between Bagging and the other approaches on Gold Standard Movie is shown in Fig. 15.

In this case, Bagging achieves high performance through NB and ME with all the considered combination rules, while for CRF the base classifiers perform better than Bagging. Using SVM, Bagging-MV achieves the highest performance. Concerning the combination rules enclosed in Bagging, the results about error relative improvement show high-quality performance, except for CRF. A summary of accuracy improvements is depicted in Fig. 16, where the outperforming ensemble obtained by BMA comprises DIC, SVM and CRF. The proposed approach is able to achieve 88.43% of accuracy against the SV systems (87.25% by MV, 85.68% by MAX, 86.86% by MEAN and 85.68% by PRODUCT). The model selection strategy, applied to both Gold Standard Movie and Person, ensures again the composition of the optimal BMA ensemble through the greedy search.

In conclusion, Fig. 17 shows that BMA with the proposed model selection strategy, ensures a significant performance improvement with regard to the studied baseline classifiers and SV. BMA outperforms not only any Bagging composition obtained with the considered combination rules, but also the best baseline classifiers and Simple Voting (i.e. SV-MEAN).

Moreover, BMA works better both on well-formed documents (reviews) and social network messages (tweets). It is important to

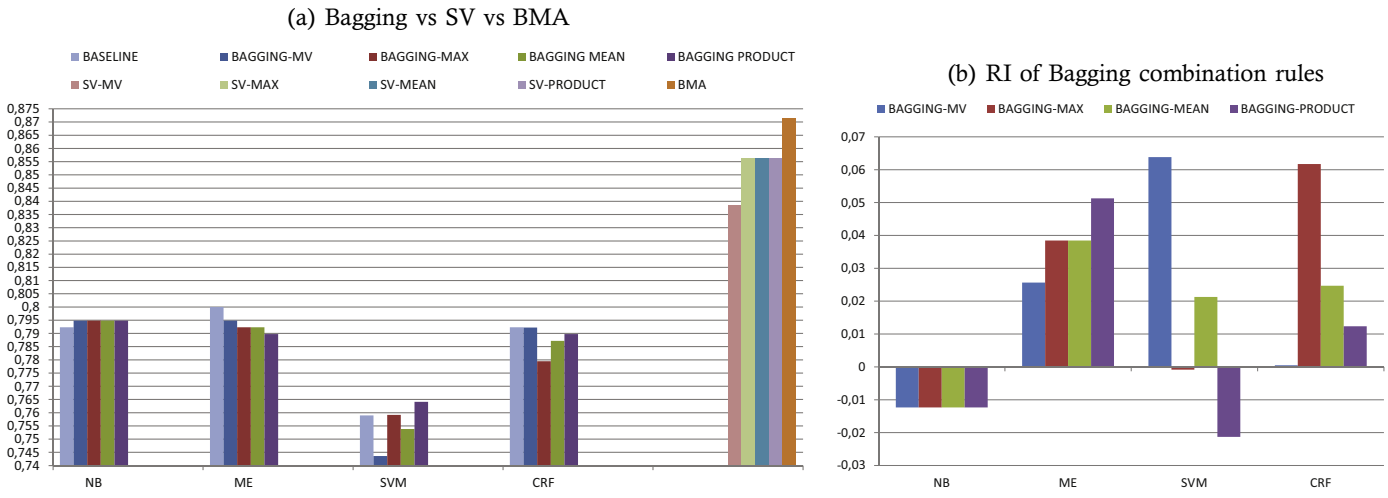

Fig. 13. Bagging performance on Gold Standard Person.

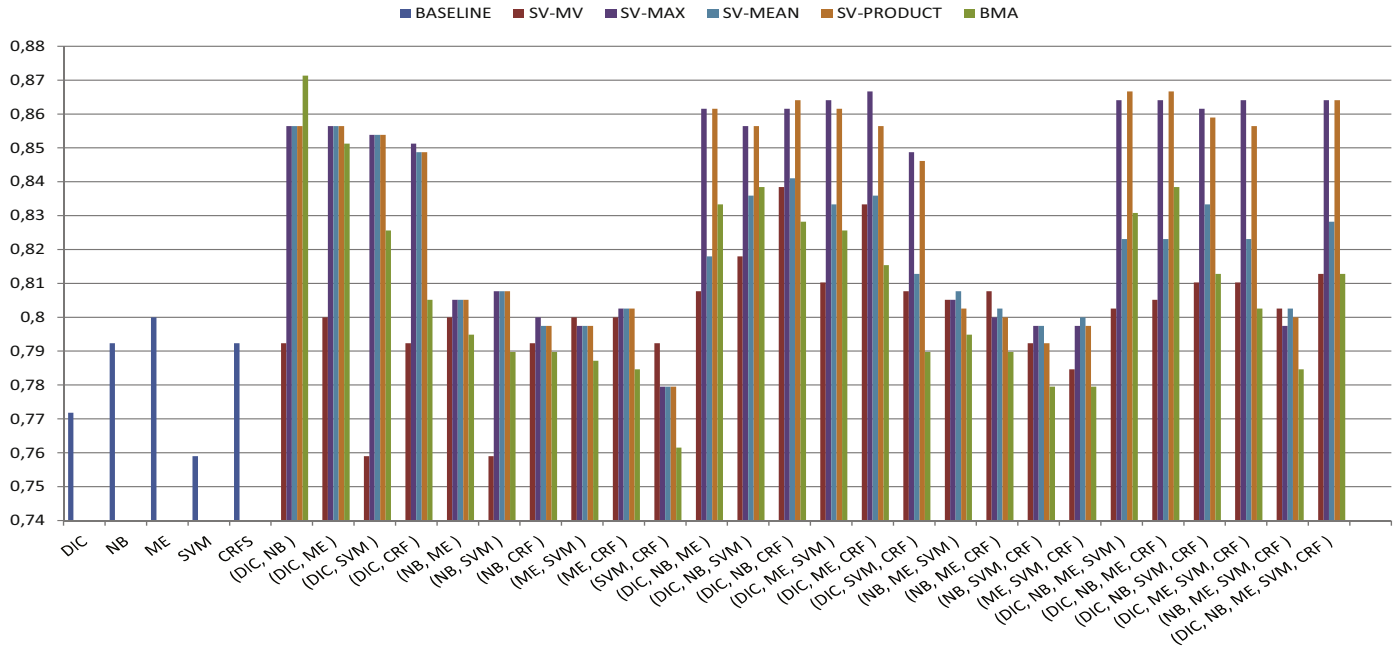

Fig. 14. Accuracy of baseline classifiers, SV and BMA on Gold Standard Person.

remark that the proposed approach radically reduces the search space for composing the optimal ensemble, ensuring outperforming performance than other approaches.

7. Computational complexity analysis

In order to demonstrate the efficiency of the proposed approach compared to the other ensemble techniques, a computational complexity analysis has been performed. Concerning SV, it is well known that it belongs to the NP-Hard complexity class due to the exhaustive search (combinatorial problem) to find the optimal ensemble composition. Bagging, as a bootstrapping technique, is a computationally intensive procedure mainly affected by the training phase. The resampling step, together with the learning of a given classifier for a given number of bags, leads to a time expensive approach. On the contrary, BMA results in a more efficient paradigm characterized by reduced time costs. To better grasp the computational complexity of the considered ensemble techniques, we can distinguish between three phases: training, search of the optimal ensemble and inference. While the computational complexity of SV and BMA depends on the number N of classifiers, Bagging takes also into account the number of replacements R and the number of bags B. Assuming that learning and inference on a given classifier take $\mathcal{O}(1)$, a comparison of the time complexity is reported in Table 3.

Concerning the training phase, although Bagging is characterized by a linear time complexity as well as the two other approaches, it results to be the most computationally intensive technique in practice. Indeed, while BMA and SV can be solved in $\mathcal{O}(N)$, Bagging results in a higher computational complexity equal to $\mathcal{O}(B(N+R))$ due to B and R that are by definition greater than one. Regarding the search of the optimal ensemble composition, SV is the most time consuming approach characterized by an exponential time complexity of $\mathcal{O}\left(2^N\right)$ compared to the linear BMA and Bagging. Indeed, while SV must search the optimal ensemble over a hypothesis space of $\sum_{p=1}^{N} \frac{N!}{p!(N-p)!} = 2^{N-1}$, Bagging has to search over N possible candidates (all the weak classifiers are candidate to be bootstrapped) and BMA over $N-1$ candidate sub-ensembles. In the inference phase, all the ensemble learning approaches result to be linear in time complexity. However,

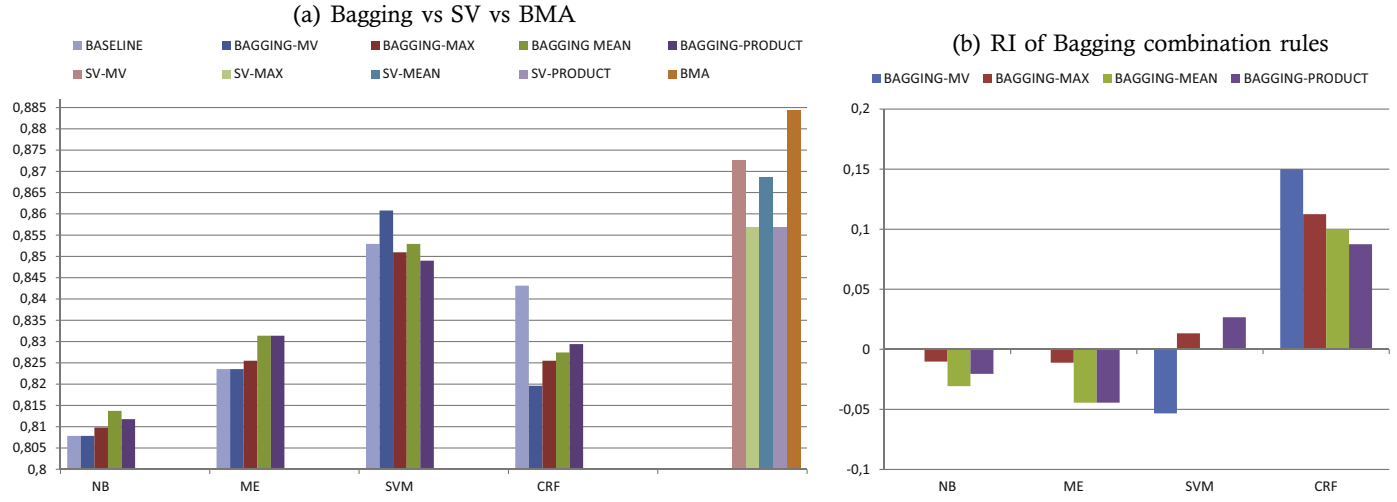

Fig. 15. Bagging performance on Gold Standard Movie.

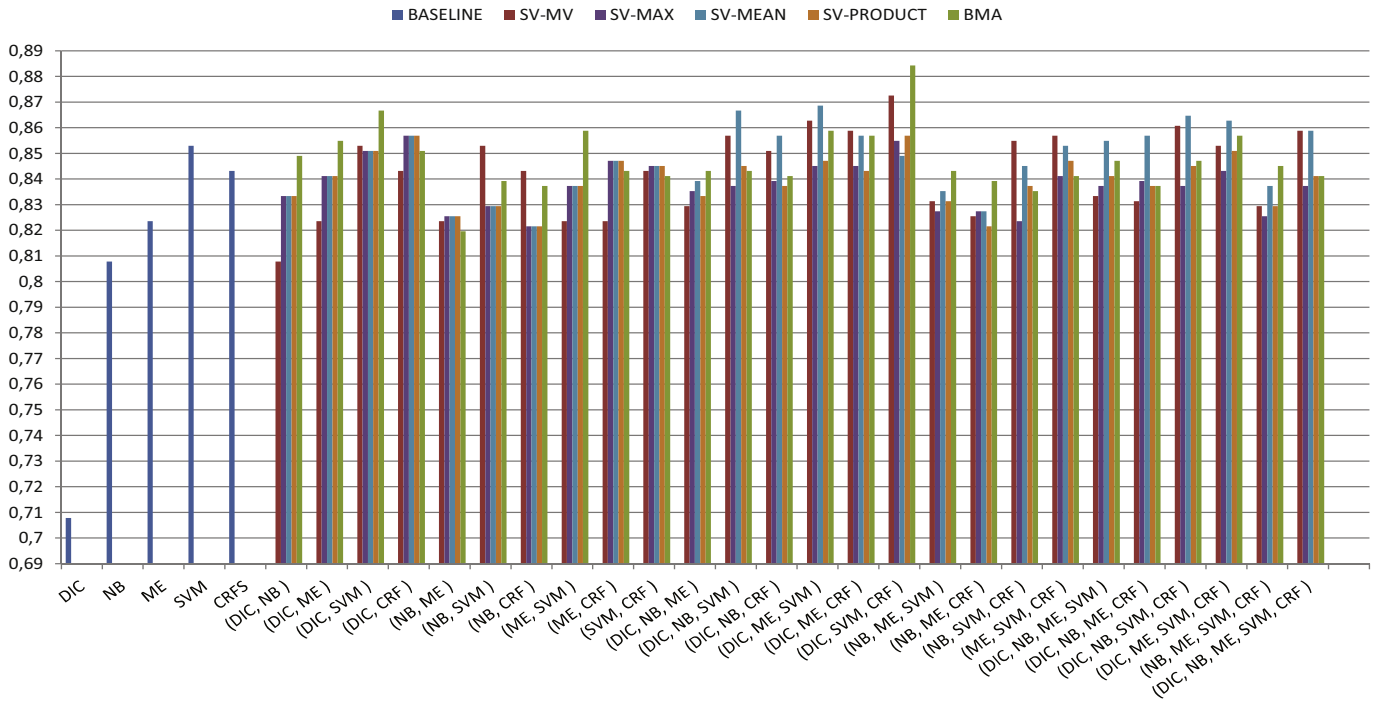

Fig. 16. Accuracy of baseline classifiers, SV and BMA on Gold Standard Movie.

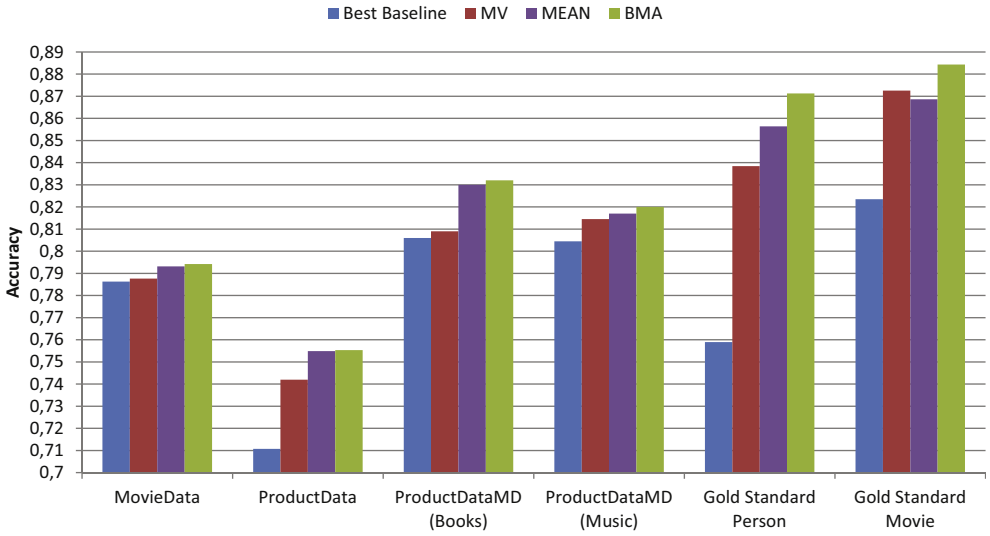

Fig. 17. Summary of accuracy comparison.

SV and BMA are more efficient when the number of bags B is greater than the number of classifiers N enclosed in the ensemble, i.e. $B > N$. From a theoretical point of view, for an increasing number of classifiers to enclose in an ensemble, BMA results to be the most efficient approach. In order to show the efficiency of the proposed approach, a further analysis has been performed. In Fig. 18 (*Review* benchmarks) and 19 (*Social* datasets) the time reductions of BMA in respect of the other approaches are reported considering the three phases, i.e. training, search of the optimal ensemble and inference.[5] If we focus on Fig. 18, we can easily observe that BMA ensures a valuable time reduction for all *Review* datasets.

When dealing with large datasets, as for example MovieData, the gain becomes more evident: while for the entire process BMA takes 8.41 min, Bagging performs in 78.22 min and SV in 8.43 min. On *Social* datasets (Fig. 19), the time reductions of BMA are lower than *Review* benchmarks. This is due both to a smaller set of instances and the short nature of text. However, BMA guarantees a time reduction: on Gold Standard Person, it takes 1151 ms against 1400 of SV and 7194

Table 3
Computational complexity of ensemble learning techniques.

	Training	Search	Inference
SV	$\mathcal{O}(N)$	$\mathcal{O}\left(2^N\right)$	$\mathcal{O}(N)$
Bagging	$\mathcal{O}(B(N+R))$	$\mathcal{O}(N)$	$\mathcal{O}(B)$
BMA	$\mathcal{O}(N)$	$\mathcal{O}(N)$	$\mathcal{O}(N)$

[5] Time performance has been measured on a Desktop PC with a Windows 7 64-bit Operating System, Pentium Quad Core i7 3.10 GHz Processor and 8 GB RAM.

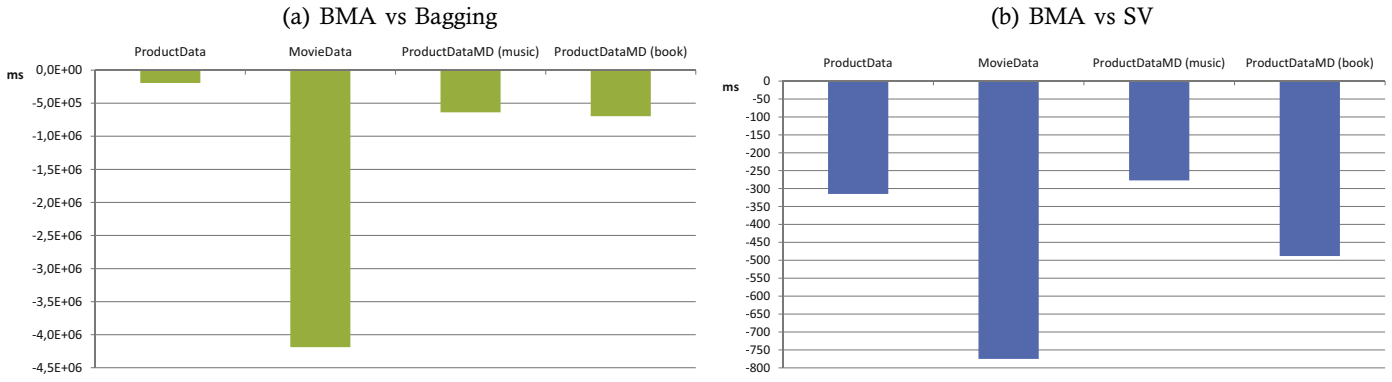

Fig. 18. Efficiency comparison (in terms of milliseconds) on *Review* benchmarks.

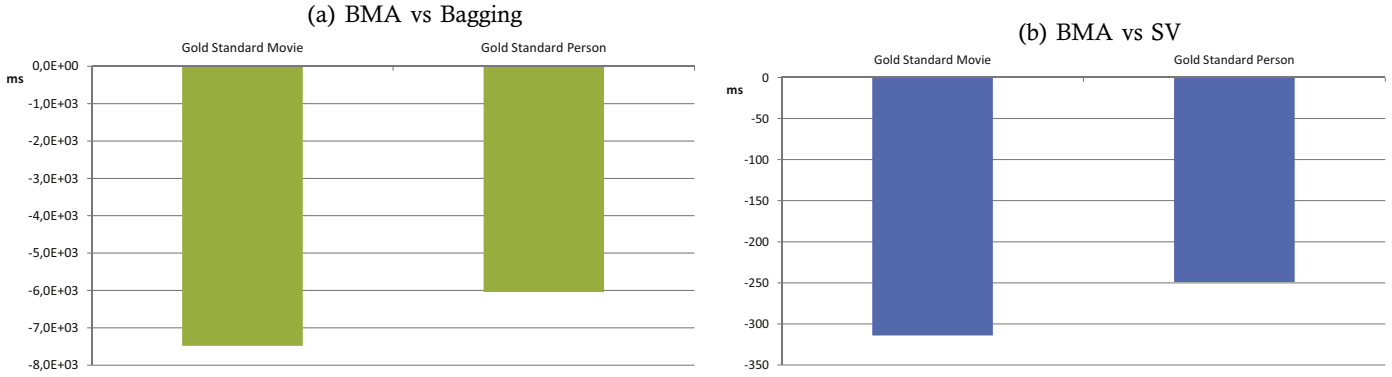

Fig. 19. Efficiency comparison (in terms of milliseconds) on *Social* benchmarks.

of Bagging, while on Gold Standard Movie, it takes 1353 ms against 1667 of SV and 8835 of Bagging.

8. Conclusion

In this paper, a novel ensemble approach for sentiment classification purposes has been introduced. The experimental results show that the proposed solution is particularly effective and efficient, thanks to its ability to define a strategic combination of different classifiers through an accurate and computationally efficient heuristic. However, an increasing number of classifiers to be enclosed in the ensemble together with large dataset open to deeper considerations in terms of complexity. The selection of the initial ensemble should consider the different complexities of each single learner and inference algorithm, leading to a reasonable trade-off between their contribution in terms of accuracy and the related computational time. A further ongoing research is related to the development of a hierarchical ensemble framework where the discrimination between "objective" and "subjective" is firstly addressed. The polarity classification of subjective expressions is then performed considering a wider range of labels.

References

[1] M.J. Silva, P. Carvalho, L. Sarmento, E. de Oliveira, P. Magalhaes, The design of optimism, an opinion mining system for Portuguese politics, New Trends Artifical Intelligence (2009) 12–15.
[2] Y.-M. Li, T.-Y. Li, Deriving market intelligence from microblogs, Decision Support Systems 55 (1) (2013) 206–217.
[3] A. Ortigosa, J.M. Martín, R.M. Carro, Sentiment analysis in Facebook and its application to e-learning, Computers in Human Behavior 31 (2014) 527–541.
[4] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundation and Trends in Information Retrieval 2 (2008) 1–135.
[5] F.A. Pozzi, E. Fersini, E. Messina, D. Blanc, Enhance polarity classification on social media through sentiment-based feature expansion, Proceedings of the 14th Workshop "From Objects to Agents" — 13th Conference of the Italian Association for Artificial Intelligence, 2013, pp. 78–84.
[6] F.A. Pozzi, D. Maccagnola, E. Fersini, E. Messina, Enhance user-level sentiment analysis on microblogs with approval relations, Proceedings of the 13th Conference of the Italian Association for Artificial Intelligence, 2013, pp. 133–144.
[7] H. Zhang, Z. Yu, M. Xu, Y. Shi, Feature-level sentiment analysis for Chinese product reviews, Proceedings of the 3rd International Conference on Computer Research and Development, 2011, pp. 135–140.
[8] P.D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 417–424.
[9] W. Peng, D.H. Park, Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization, Proceedings of the 5th International Conference on Weblogs and Social Media, 2011, pp. 273–280.
[10] D. Rao, D. Ravichandran, Semi-supervised polarity lexicon induction, Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 675–682.
[11] G. Wang, J. Sun, J. Ma, K. Xu, J. Gu, Sentiment classification: the contribution of ensemble learning, Decision Support Systems 57 (2014) 77–93.
[12] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, Proceedings of the Association of Computational Linguistics Conference on Empirical Methods in Natural Language Processing, 2002, pp. 79–86.
[13] A. Hassan, A. Abbasi, D. Zeng, Twitter sentiment analysis: a bootstrap ensemble framework, Proceedings of the International Conference on Social Computing, 2013, pp. 357–364.
[14] R. Xia, C. Zong, S. Li, Ensemble of feature sets and classification algorithms for sentiment classification, Information Sciences 181 (6) (2011) 1138–1152.
[15] F.A. Pozzi, E. Fersini, E. Messina, Bayesian model averaging and model selection for polarity classification, Proceedings of the 18th International Conference on Applications of Natural Language to Information Systems, 2013, pp. 189–200.
[16] I. Partalas, G. Tsoumakas, I. Vlahavas, An ensemble uncertainty aware measure for directed hill climbing ensemble pruning, Machine Learning 81 (3) (2010) 257–282.
[17] G. Martínez-Muñoz, A. Suárez, Pruning in ordered bagging ensembles, Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 609–616.
[18] R. Caruana, A. Niculescu-Mizil, G. Crew, A. Ksikes, Ensemble selection from libraries of models, Proceedings of the 21st International Conference on Machine Learning, 2004, pp. 18–25.
[19] M. Hu, B. Liu, Mining and summarizing customer reviews, Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.
[20] A. McCallum, K. Nigam, A comparison of event models for Naive Bayes text classification, Proceedings of the Workshop on Learning for Text Categorization — 15th National Conference on Artificial Intelligence, 1998, pp. 41–48.

[21] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.
[22] A. McCallum, C. Pal, G. Druck, X. Wang, Multi-conditional learning: generative/discriminative training for clustering and classification, Proceedings of the 21st National Conference on Artificial intelligence, 2006, pp. 433–439.
[23] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, Proceedings of the 18th International Conference on Machine Learning, 2001, pp. 282–289.
[24] J.R. Quinlan, Bagging, boosting, and c4.5, Proceedings of the 13th National Conference on Artificial Intelligence, 1996, pp. 725–730.
[25] B. Pang, L. Lee, Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales, Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005, pp. 115–124.
[26] O. Täckström, R. McDonald, Semi-supervised latent variable models for sentence-level sentiment analysis, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011, pp. 569–574.
[27] J. Blitzer, M. Dredze, F. Pereira, Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, 2007, pp. 187–205.
[28] C. Lu, W. Wenbo, N. Meenakshi, W. Shaojun, S. Amit P, Extracting diverse sentiment expressions with target-dependent polarity from Twitter, Proceedings of the 5th International Conference on Weblogs and Social Media, 2012, pp. 50–57.
[29] M.J. Zaki, W. Meira Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014.

Elisabetta Fersini is currently a postdoctoral research fellow at the University of Milano-Bicocca (Department of Informatics, Systems and Communication — DISCo). Her research activity is mainly focused on statistical relational learning with particular interests in supervised and unsupervised classification. The research activity finds application to Decision Support Systems, Web/Text mining, e-Justice and Bioinformatics.

Enza Messina is a Professor in Operations Research at the University of Milano-Bicocca (Department of Informatics, Systems and Communications), where she founded the research Laboratory MIND (www.mind.disco.unimib.it). She holds a PhD in Computational Mathematics and Operations Research from the University of Milano. Her research activity is mainly focused on the development of models and methods for decision-making under uncertainty and statistical relational models for data analysis.

Federico Alberto Pozzi is currently a PhD Candidate in Computer Science at the University of Milano-Bicocca (Department of Informatics, Systems and Communications). His research activity, mainly aimed at the development of models and methods for sentiment analysis, finds application in Decision Support Systems and Computational Finance.