10.12871/clicit2014212

Subjectivity, Polarity And Irony Detection: A Multi-Layer Approach

Elisabetta Fersini, Enza Messina, Federico Alberto Pozzi
DISCo, University of Milano-Bicocca
Viale Sarca 336
20126 - Milan
{fersini,messina,federico.pozzi}@disco.unimib.it

Abstract

English. In the literature, subjectivity, polarity and irony detection have been often considered as independent tasks. However, since there are multiple ties between them, they should be jointly addressed. In this paper we propose a hierarchical system, where the classifiers of each layer are built upon an ensemble approach known as Bayesian Model Averaging.

Italiano. *In letteratura, le classificazioni di soggettività, polarità e ironia sono state spesso affrontate come task indipendenti. Tuttavia, dal momento che esistono tra loro diversi legami impliciti, tali task dovrebbero essere affrontati congiuntamente. In questo lavoro proponiamo un sistema gerarchico, dove i classificatori di ogni layer sono costruiti ricorrendo ad un approccio di ensemble learning noto come Bayesian Model Averaging.*

1 Introduction

Among the computational approaches for distinguishing subjective vs objective messages, ironic vs not ironic and different classes of polarities, we can point out two main research directions: the first one focuses on machine learning algorithms for automatic recognition (Pang et al., 2002; Chen et al., 2008; Ye et al., 2009; Perea-Ortega et al., 2013; Pozzi et al., 2013c; Pozzi et al., 2013a), while the second one is aimed at the identification of linguistic and metalinguistic features useful for automatic detection (Carvalho et al., 2009; Filatova, 2012; Pozzi et al., 2013b; Davidov et al., 2010; Reyes et al., 2013). As far is concerned with the machine learning perspective, although some approaches are widely used in sentiment analysis, they suffer from two main limitations that the proposed paper intends to overcome. First, all the issues related to sentiment analysis are usually approached by focusing on specific tasks separately, i.e. subjectivity, polarity and irony are tackled independently on each other. In a real context all these issues should be addressed by a single model able to distinguish at first if a message is either subjective or objective, to subsequently address polarity and irony detection and deal with the potential relationships that could exists between them. Second, within the sentiment analysis research field there is no agreement on which machine learning 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. In this paper we present a system based on a multi-layer Bayesian ensemble learning that tries to overcome the above mentioned limitations. The focus is therefore intentionally on learning strategies instead of on linguistic aspects to investigate the potential of multiple and interconnected layers of ensembles on real word Italian Twitter data.

2 Description of the system

2.1 Hierarchical Bayesian Model Averaging

In the literature, *subjectivity*, *polarity* and *irony* detection have been often considered as independent tasks. However, since there are multiple ties between them, they should be jointly addressed. Different works have usually treated subjectivity and polarity classification as two-stage binary classification process, where the first level distinguishes subjective and objective (neutral) statements, and the second level then further distinguishes subjectivity into: subjective-positive / subjective-negative (Refaee and Rieser, 2014; Baugh, 2013). The results proposed in (Wilson et

al., 2009) support the validity of this process, indicating that the ability to recognize neutral classes in the first place can greatly improve the performance in distinguishing between positive and negative utterances at a later time. However, as briefly introduced, also irony can give its contribution in improving the classification performance. An ironic message involves a shift in evaluative valence, which can be treated in two ways: it could be a shift from a literally positive to an intended negative meaning, or a shift from a literally negative to an intended positive evaluation.

According to the above mentioned considerations, we propose a hierarchical framework able to jointly address subjectivity, polarity and irony detection. An overview of the working system, named *Hierarchical Bayesian Model Averaging* (H-BMA), is presented in Figure 1.

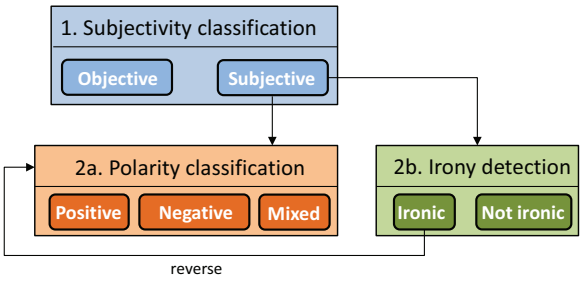

Figure 1: Hierarchical BMA.

Since subjectivity classification is usually the most performing task in Sentiment Analysis, the first level distinguishes subjective and objective statements (neutral is supposed to be objective), and the second level then distinguishes subjectivity into: subjective-positive / subjective-negative / subjective-mixed (a sentence which is subjective, positive and negative at the same time). Jointly with polarity classification, irony detection is also performed. If a given sentence is detected as ironic, then its positive or negative polarity is reversed. On the other side, if the sentence is ironic but its polarity has been classified as mixed, then it is switched to negative. Thus a message s, identified as mixed by the polarity classification layer and ironic (denoted as *iro*) by the irony detection layer, is finally labelled as negative ($-$) due to the conditional distribution

$$P(s = \text{-} \mid s = \text{iro}) >> P(s = \text{+} \mid s = \text{iro}) \quad (1)$$

In the literature, *subjectivity*, *polarity* and *irony* detection have been often addressed applying the most varied machine learning approaches. As outlined in the Introduction, there is no agreement on which methodology is better than others. The uncertainty about which model represents the optimal one in different context has been overcome in this work by introducing Bayesian Model Averaging (Pozzi et al., 2013a), a novel ensemble learning approach able to exploit the potentials of several learners when predicting the labels for each task (subjectivity, irony and polarity) of the hierarchical framework.

2.2 Bayesian Model Averaging

The most important limitation of traditional ensemble approaches is that the models to be included in the *set of experts* have uniform distributed weights regardless their reliability. However, the uncertainty left by data and models can be filtered by considering the Bayesian paradigm. In particular, through Bayesian Model Averaging (BMA) all possible models in the hypothesis space could be used when making predictions, considering their marginal prediction capabilities and their reliability. Given a dataset $\mathcal{D}$ and a set C of classifiers, the approach assigns to a message s the label $l(s)$ that maximizes:

$$P(l(s) \mid C, \mathcal{D}) = \sum_{i \in C} P(l(s) \mid i, \mathcal{D}) P(i \mid \mathcal{D}) \quad (2)$$

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

$$P(i \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid i) P(i)}{\sum_{j \in C} P(\mathcal{D} \mid j) P(j)} \quad (3)$$

where $P(i)$ is the prior probability of i and $P(\mathcal{D} \mid i)$ is the model likelihood. In eq. 3, $P(i)$ and $\sum_{j \in C} P(\mathcal{D} \mid j) P(j)$ are assumed to be a constant and therefore can be omitted. Therefore, BMA assigns the label $l^{BMA}(s)$ to s according to the following decision rule:

$$\begin{aligned} l^{BMA}(s) &= \arg\max_{l(s)} P(l(m)|C, \mathcal{D}) \\ &= \sum_{i \in C} P(l(s)|i, \mathcal{D}) P(i|\mathcal{D}) \\ &= \sum_{i \in C} P(l(s)|i, \mathcal{D}) P(\mathcal{D}|i) P(i) \\ &= \sum_{i \in C} P(l(s)|i, \mathcal{D}) P(\mathcal{D}|i) \end{aligned} \quad (4)$$

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

$$P(\mathcal{D} \mid i) \approx \frac{1}{\iota} \sum_{\iota=1}^{\phi} \frac{2 \times P_{i\iota}(\mathcal{D}) \times R_{i\iota}(\mathcal{D})}{P_{i\iota}(\mathcal{D}) + R_{i\iota}(\mathcal{D})} \quad (5)$$

where $P_{i\iota}(\mathcal{D})$ and $R_{i\iota}(\mathcal{D})$ denotes precision and recall obtained by classifier i in fold ι.

In this way we tune the probabilistic claim of each classifier in the ensemble according to its ability to fit the training data. This approach allows the uncertainty of each classifier to be taken into account, avoiding over-confident inferences.

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 (Partalas et al., 2010)). To the best of our knowledge this measures are not suitable for a Bayesian Ensemble, because they assume uniform weight distribution of classifiers. In this study, we used 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:

3. 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.

4. 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, let compute $P(i = 1 \mid j = 0)$ as the number of instances correctly classified by i over the number of instances incorrectly classified by j (case 1) and $P(i = 1 \mid j = 1)$ the number of instances correctly classified both by i over the number of instances correctly classified by j (case 2). Analogously, let $P(i = 0 \mid j = 1)$ be the number of instances misclassified by i over the number of instances correctly classified by j (case 3) and $P(i = 0 \mid j = 0)$ the number of instances misclassified by i over the number of instances misclassified also by j (case 4).

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

$$r_i^S = \frac{\sum\limits_{j \in \{S \setminus i\}} \sum\limits_{q \in \{0,1\}} P(i = 1 \mid j = q) P(j = q)}{\sum\limits_{j \in \{S \setminus i\}} \sum\limits_{q \in \{0,1\}} P(i = 0 \mid j = q) P(j = q)} \quad (6)$$

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 (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 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 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 maxima and therefore is accepted as optimal ensemble combination. More formally, an ensemble S is accepted as optimal composition if the following condition is satisfied:

$$\frac{ACC(S)}{|S|} \geq \frac{ACC(S \setminus x)}{|S-1|} \tag{7}$$

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 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 $|S| = N, N-1, \ldots, 1$.

3 Results

In order to derive the feature space used for learning, a vector space model has been adopted. Each sentence s is represented as a vector composed of terms for which a corresponding weight w can be computed as Boolean (0/1). No additional information, such as linguistic cues, has been provided to the learning approaches investigated in this paper. The proposed Hierarchical Bayesian Model Averaging (H-BMA) has been compared with traditional Bayesian Model Averaging (BMA) and the baseline provided by Sentipolc 2014 organizers (Basile et al., 2014). The classifiers enclosed in H-BMA and BMA for addressing the three tasks are: Decision Tree (DT) (Quinlan, 1993), Support Vector Machines (SVM) (Vapnik and Vapnik, 1998), Multinomial Naive Bayes (MNB) (Langley et al., 1992) and K-Nearest Neighbors (KNN) (Aha et al., 1991). The indices used for comparing the approaches are Precision, Recall and F_1-measure.

	Baseline	**BMA**	**H-BMA***
Subjectivity	0.4005	0.6173	0.6173
Polarity	0.3718	0.4907	0.5253
Irony	0.4441	0.5253	0.5261

Table 1: Comparison of F_1-measure

The results reported in Table 1 show the F_1-measure performance on the three tasks*. The optimal ensemble composition of both BMA and H-BMA has been obtained according the greedy backward elimination strategy that lead to ensemble composed of DT, SVM and MNB (for all the three tasks). It can be easily noted that addressing Subjectivity, Polarity and Irony detection with H-BMA, where tasks are modelled as interdependent, the performance tend to improve with respect to the other approaches where the issues are tackled independently.

4 Discussion

In this paper, a novel system for jointly modelling subjectivity, polarity and irony detection has been introduced. The experimental results show the potential of the proposed model to address interdependent tasks with no additional information derived by linguistic cues. 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 each ensemble in all the layers 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 linguistic aspects that could be taken into account during the learning phase of the models in the ensembles. Specific linguistic cues able to characterise subjectivity, polarity and irony could lead to more accurate learning and prediction.

*Official results provided to Sentipolc 2014 organizers (Basile et al., 2014) lead to the following F_1-measure performance: Subjectivity 0.5901, Polarity 0.5341 and Irony 0.4771. The results reported in Table 1 differ from the ones reported in the official ranking because of a mistake in sending the correct predictions.

References

David W Aha, Dennis Kibler, and Marc K Albert. 1991. Instance-based learning algorithms. *Machine learning*, 6(1):37–66.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa, Italy.

Wesley Baugh. 2013. bwbaugh : Hierarchical sentiment analysis with partial self-training. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 539–542. Association for Computational Linguistics.

Paula Carvalho, Luís Sarmento, Mário J Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM.

Bo Chen, Hui He, and Jun Guo. 2008. Constructing maximum entropy language models for movie review subjectivity analysis. *Journal of Computer Science and Technology*, 23(2):231–239.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *LREC*, pages 392–398.

Pat Langley, Wayne Iba, and, and Kevin Thompson. 1992. An analysis of bayesian classifiers. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, pages 223–228. AAAI Press.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Ioannis Partalas, Grigorios Tsoumakas, and Ioannis Vlahavas. 2010. An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. *Machine Learning*, 81(3):257–282.

José M Perea-Ortega, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, and L Alfonso Ureña-López. 2013. Combining supervised and unsupervised polarity classification for non-english reviews. In *Computational Linguistics and Intelligent Text Processing*, pages 63–74. Springer.

Federico Alberto Pozzi, Elisabetta Fersini, and Enza Messina. 2013a. Bayesian model averaging and model selection for polarity classification. In *Proceedings of the 18th International Conference on Applications of Natural Language to Information Systems*, pages 189–200.

Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Daniele Blanc. 2013b. Enhance polarity classification on social media through sentiment-based feature expansion. In *Proceedings of the 14th Workshop "From Objects to Agents" co-located with the 13th Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), Torino, Italy, December 2-3, 2013.*, pages 78–84.

Federico Alberto Pozzi, Daniele Maccagnola, Elisabetta Fersini, and Enza Messina. 2013c. Enhance user-level sentiment analysis on microblogs with approval relations. In *AI* IA 2013: Advances in Artificial Intelligence*, pages 133–144. Springer.

John Ross Quinlan. 1993. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann.

Eshrag Refaee and Verena Rieser. 2014. Subjectivity and sentiment analysis of arabic twitter feeds with limited resources. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, LREC14, pages 16–21.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268.

Vladimir Naumovich Vapnik and Vlamimir Vapnik. 1998. *Statistical learning theory*, volume 2. Wiley New York.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Comput. Linguist.*, 35(3):399–433.

Qiang Ye, Ziqiong Zhang, and Rob Law. 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3):6527–6535.