AMI&ERIC: How to Learn with Naive Bayes and Prior Knowledge: an Application to Sentiment Analysis

Mohamed Dermouche[1,2], **Leila Khouas**[1], **Julien Velcin**[2] and **Sabine Loudcher**[2]

[1]AMI Software R&D
1475 av. A. Einstein
34000 Montpellier, France
mde@amisw.com
lkh@amisw.com

[2]Université de Lyon, ERIC (Lyon 2)
5 av. P. Mendès-France
69676 Bron Cedex, France
julien.velcin@univ-lyon2.fr
sabine.loudcher@univ-lyon2.fr

Abstract

In this paper, we describe our system that participated in SemEval-2013, Task 2.B (sentiment analysis in Twitter). Our approach consists of adapting Naive Bayes probabilities in order to take into account prior knowledge (represented in the form of a sentiment lexicon). We propose two different methods to efficiently incorporate prior knowledge. We show that our approach outperforms the classical Naive Bayes method and shows competitive results with SVM while having less computational complexity.

1 Introduction

With the advent of Internet microblogging, social networks, like Twitter[1] and Facebook[2], have brought about a real revolution in our way of communicating. People share their opinions of everyday life without taboos or restrictions thanks to the anonymity offered by these tools, which makes them a valuable source of information rather rich of subjective data. These data can be mined using sentiment analysis as a means to understand people's feelings towards a political cause or what people are thinking about a product or a service. Recent works showed that Twitter sentiments can be correlated to box-office revenues (Asur and Huberman, 2010) or political polls (O'Connor et al., 2010).

Machine learning methods, like Naive Bayes (NB) and Support Vector Machines (SVM), have been widely used in sentiment analysis (Pang et al., 2002; Pak and Paroubek, 2010). One major problem with these methods, and in particular NB, is that the model is built only on the learning data which can lead to overfitting. In this paper, we describe our approach that participated in SemEval-2013, Task 2.B (sentiment analysis in Twitter) (Wilson et al., 2013). Our approach consists of learning with both NB and prior knowledge. We show that our approach outperforms the classical NB method and gives competitive results compared to SVM while having less computational complexity.

The remainder of this paper is organized as follows: prior works on sentiment analysis are discussed in Section 2. The proposed approach is detailed in Section 3. Then, experiments and results are given in Section 4 and 5.

2 Background

Sentiment analysis is a text mining task which deals with the feelings expressed explicitly or implicitly in a textual content. It concerns subjectivity analysis (subjective/objective), opinion mining (positive/negative/neutral), strength analysis, etc. Although the term "sentiment analysis" includes all these tasks, it often refers to opinion mining. Sentiment analysis methods can be categorized into machine learning, linguistic and hybrid methods.

Machine learning methods are usually supervised. A model is built based on a learning dataset composed of annotated texts and represented by a bag of words. The model is then deployed to classify new texts. Pang et al. (2002) use machine learning methods (NB, SVM and MaxEnt) to detect sentiments on movie reviews. Pak and Paroubek (2010) use NB to

[1] http://www.twitter.com/

[2] http://www.facebook.com/

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 364–368, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

perform sentiment analysis on Twitter data.

Linguistic methods use lexicons and manually-crafted rules to detect sentiments. Kennedy and Inkpen (2006) use syntactic analysis to capture language aspects like negation and contextual valence shifters. Other works (Turney and Littman, 2003; Kamps et al., 2004) propose to use a term similarity measure which can be statistical (e.g., Mutual Information, LSA) or semantic (e.g., WordNet-based).

Hybrid methods use both statistical and linguistic approaches. Esuli and Sebastiani (2011), which is the closest work to ours, propose to use annotated lexical resources to improve opinion extraction. The bag-of-word text representation is enriched by new tags (e.g. subjectivity, polarity). Then, an SVM-based system is used for opinion classification.

3 Our approach

NB is a machine learning method that builds a classification model based only on the learning data which makes it highly dependent on this data. For example, in a sentiment analysis task, if the term `actor` appears more frequently within a negative context than in a positive one, it will be classified as negative while actually it is not. Moreover, NB tends sometimes to predict the class of majority (observed on learning data) which increases classification errors on unbalanced data. Our approach consists of incorporating prior knowledge into the NB model to make it less dependent on learning data.

To be efficiently used, prior knowledge must be represented in a structured form. We choose, here, to represent it by a sentiment lexicon (a set of positive and negative terms). Several lexicons have already been developed to address sentiment analysis issues. Some of them are publicly available like the MPQA subjectivity lexicon (Wilson et al., 2005), Liu's opinion lexicon (Ding et al., 2008), SentiWordNet (Esuli and Sebastiani, 2006). We believe that such knowledge can be quite useful if used correctly and efficiently by machine learning methods.

In the following, we settle for a 2-way classification task (positive vs. negative). Texts are represented by a vector space model (Salton et al., 1975) and terms are weighted according to their presence/absence in the text because previous works (Pang et al., 2002; Pak and Paroubek, 2010) showed that Boolean model performs better than other weighting schemes in sentiment analysis. We denote by w and $\bar{w}$ the presence, respectively absence, modality of a word w. A "term" stands, here, for any type of text features (smileys, n-grams).

3.1 Sentiment lexicon

We represent the prior knowledge by a 2-class sentiment lexicon: a list of subjective terms (words, n-grams and smileys) manually annotated with two scores: positive ($score_{c_+}$) and negative ($score_{c_-}$). Each term has a score of 1 on a class polarity (we call it right class) and 0 on the other one (wrong class). For example, the word `good` has $score_{c_+} = 1$ and $score_{c_-} = 0$. Then, c_+ is the right class of the word `good` and c_- is the wrong class.

3.2 NB method

NB is based on calculating class-wise term probabilities on a learning dataset D where each text $d \in D$ is annotated with a class $c \in \{c_+, c_-\}$. In the learning step, probability values $p(w|c)$ are estimated from D as follows:

$$p(w|c) = \frac{1}{nb(c)} \cdot nb(w,c) \qquad (1)$$

Where $nb(c)$ denotes the number of texts of class c and $nb(w,c)$ is the number of texts of class c that contain the term w.

Once these probabilities are calculated for each couple (w,c), the model can be used to classify new texts. We choose to assign a new text d to the class that maximizes the probability $p(c|d)$. Using Bayes' theorem and independence assumption between term distributions, this probability is calculated as follows (the denominator can be dropped because it is not dependent on the class c):

$$p(c|d) = \frac{p(c) \cdot \prod_{w \in d} p(w|c)}{p(d)} \qquad (2)$$

3.3 Incorporating prior knowledge

Prior knowledge is incorporated by adapting NB formulas. We propose two different methods to do this: Add & Remove and Transfer. These methods differ in the way to calculate the class-wise term probabilities $p(w|c)$ but use the same classification rule: $class(d) = \arg\max_{c \in \{c_+, c_-\}} p(c|d)$.

Add & Remove. This method consists of artificially adding some occurrences of term w to the right class and removing some occurrences from the wrong class. The lexicon is used to determine for each term its right and wrong classes. To ensure that probability values do not exceed 1, we introduce $nb(\bar{w}, c)$, the number of texts of class c that do not contain the term w, which is also equal to the maximum number of occurrences of w that can be added to the class c. Thus, the number of added occurrences is a ratio α_c of this maximum ($0 \leq \alpha_c \leq 1$). Likewise, if c was the wrong class of w, the number of removed occurrences from the class c is a ratio β_c of the maximum number that can be removed from the class c, $nb(w, c)$, with $0 \leq \beta_c \leq 1$. Formally, term probabilities are calculated as follows:

$$p(w|c) = \frac{1}{nb(c)} \cdot [nb(w, c) + \alpha_c \cdot score_c(w) \cdot nb(\bar{w}, c) - \beta_c \cdot score_{\bar{c}}(w) \cdot nb(w, c)] \quad (3)$$

Transfer. This method consists of transferring some occurrences of a term w from the wrong class to the right class. The number of transferred occurrences is such that the final probability is not greater than 1 and the number of transferred occurrences is not greater than the actual number of occurrences in the wrong class. To meet these constraints, we introduce $max(w, c)$: the maximum number of occurrences of w that can be transferred to the class c from the other class $\bar{c}$. This number must not be greater than both the number of texts from $\bar{c}$ containing w and the number of texts from c not containing w.

$$max(w, c) = \min\{nb(w, \bar{c}), nb(\bar{w}, c)\} \quad (4)$$

Finally, the number of occurrences actually transferred is a ratio α_c of $max(w, c)$ with $0 \leq \alpha_c \leq 1$. Term probabilities are estimated as follows:

$$p(w|c) = \frac{1}{nb(c)} \cdot [nb(w, c) + \alpha_c \cdot score_c(w) \cdot max(w, c) - \alpha_c \cdot score_{\bar{c}}(w) \cdot max(w, \bar{c})] \quad (5)$$

Both methods, Add & Remove and Transfer, consist of removing occurrences from the wrong class and adding occurrences to the right class with the difference that in Transfer, the number of added occurrences is exactly the number of removed ones.

4 Experiment

4.1 Sentiment lexicon

For SemEval-2013 contest (Wilson et al., 2013), we have developed our own lexicon based on Liu's opinion lexicon (Ding et al., 2008) and enriched with some "microblogging style" terms (e.g., `luv, xox, gd`) manually collected on the Urban Dictionary[3]. The whole lexicon contains 7720 English terms (words, 2-grams, 3-grams and smileys) where 2475 are positive and 5245 negative.

4.2 Dataset and preprocessing

To evaluate the proposed approach, we use SemEval-2013 datasets: TW (tweets obtained by merging learn and development data) and SMS, in addition to MR (English movie reviews of Pang and Lee (2004)). Concerning SMS, the classification is performed using the model learned on tweets (TW) in order to assess how it generalizes on SMS data. Note that our approach is adapted to binary classification but can be used for 3-way classification (which is the case of TW and SMS). We do this by adapting only positive and negative probabilities, neutral ones remain unchanged.

Texts are preprocessed by removing stopwords, numerics, punctuation and terms that occur only once (to reduce vocabulary size and data sparseness). Texts are then stemmed using Porter stemmer (Porter, 1997). We also remove URLs and Twitter keywords (`via, RT`) from tweets.

4.3 Tools

As we compare our approach to SVM method, we have used SVM$^{\text{multiclass}}$ (Crammer and Singer, 2002). For a compromise between processing time and performance, we set the trade-off parameter c to 4 on MR dataset and 20 on TW and SMS (based on empirical results).

5 Results and discussion

In addition to the two proposed methods: Add & Remove (A&R) and Transfer (TRA), texts are classified using NB and SVM with two kernels: linear (SVM-L) and polynomial of degree 2 (SVM-P). All the scores given below correspond to the average

[3] http://www.urbandictionary.com/

F-score of positive and negative classes, even for 3-way classification. This measure is also used in SemEval-2013 result evaluation and ranking (Wilson et al., 2013).

5.1 General results

General results are obtained only with unigrams and smileys. Figure 1 presents the results obtained on the different datasets on both 2-way (left) and 3-way (right) classifications. For 2-way classification, neutral texts are ignored and the model is evaluated using a 5-fold cross validation. For 3-way classification, the model is evaluated on the provided test data. Compared with NB, our approach performs better on all datasets. It also outperforms SVM, that achieves poor results, except on MR.

Method	2-class		3-class	
	TW	MR	TW	SMS
NB	74.07	73.06	59.43	48.80
SVM-L	49.79	74.56	37.56	32.13
SVM-P	49.74	**84.64**	37.56	32.13
A&R	**76.05**	80.57	**60.57**	49.42
TRA	76.00	75.53	60.27	**51.35**

Figure 1: General results (unigrams and smileys)

Parameter effect. To examine the effect of parameters, we perform a 2-way classification on TW and MR datasets using 5-fold cross validation (Figure 2). We take, for A&R method, $\beta_{c_+} = \beta_{c_-} = 0$ and for both methods, $\alpha_{c_+} = \alpha_{c_-}$ (denoted α). This configuration does not necessarily give the best scores. However, empirical tests showed that scores are not significantly lower than the best ones. We choose this configuration for simplicity (only one parameter to tune).

Figure 2 shows that best scores are achieved with different values of α depending on the used method (A&R, TRA) and the data. Therefore, parameters must be fine-tuned for each dataset separately.

5.2 SemEval-2013 results

For SemEval-2013 contest, we have enriched text representation by 2-grams and 3-grams and used A&R method with: $\alpha_{c_+} = \alpha_{c_-} = 0.003$, $\beta_{c_+} = 0.04$ and $\beta_{c_-} = 0.02$. All of these parameters have been fine-tuned using the development data. We have also made an Information Gain-based feature selection (Mitchell, 1997). Only the best 2000 terms are kept to which we have added terms of the lexicon. Under these conditions, our approach achieved the scores 62.55% on tweets (ranked 6th/35) and 53.63% on SMS (ranked 9th/28).

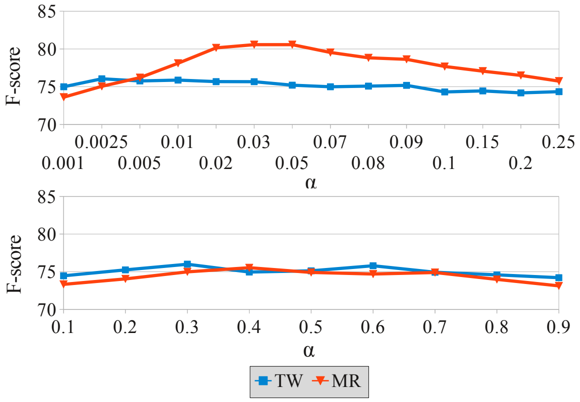

Figure 2: Effect of the parameter α on a 2-way classification using methods: A&R (top) and TRA (bottom)

Dataset	Class	Pecision	Recall	F-score
TW	positive	62.12	74.49	67.75
	negative	46.23	75.54	57.36
	neutral	76.74	44.27	56.15
SMS	positive	39.59	78.86	52.72
	negative	45.64	67.77	54.55
	neutral	90.93	39.82	55.38

Figure 3: SemEval-2013 results (A&R method)

Regarding F-score of each class (Figure 3), our approach gave better results on the negative class (under-represented in the learning data) than NB (49.09% on TW and 47.63% on SMS).

6 Conclusion

In this paper, we have presented a novel approach to sentiment analysis by incorporating prior knowledge into NB model. We showed that our approach outperforms NB and gives competitive results with SVM while better handling unbalanced data.

As a future work, further processing may be required on Twitter data. Tweets, in contrast to traditional text genres, show many specificities (short size, high misspelling rate, informal text, etc.). Moreover, tweets rely on an underlying structure (re-tweets, hashtags) that may be quite useful to build more accurate analysis tools.

References

Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'10)*, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)*, pages 231–240, New York, NY, USA, 2008. ACM.

Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, pages 417–422, Genova, IT, 2006.

Andrea Esuli and Fabrizio Sebastiani. Enhancing opinion extraction by automatically annotated lexical resources. In *Proceedings of the 4th conference on Human language technology: challenges for computer science and linguistics (LTC'09)*, pages 500–511, Poznan, Poland, 2011. Springer-Verlag.

Vasileios Hatzivassiloglou and Kathleen R Mckeown. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the eighth conference of the European chapter of the Association for Computational Linguistics (EACL'97)*, pages 174–181, Madrid, Spain, 1997. ACL.

Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. Using WordNet to measure semantic orientations of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-04)*, pages 1115–1118, Lisbon, PT, 2004.

Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, May 2006.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Washington, DC, USA, 2010.

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1320–1326, Valletta, Malta, 2010. ELRA.

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

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*, pages 271–278, Barcelona, Catalonia, Spain, 2004. ACL.

Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

Martin F. Porter. An algorithm for suffix stripping. In *Readings in information retrieval*, number 3, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

Gerard Salton, Andrew K. C. Wong, and Chung S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 347–354, Vancouver, British Columbia, Canada, 2005. ACL.

Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. SemEval-2013 Task 2: Sentiment Analysis in Twitter. *In Proceedings of the International Workshop on Semantic Evaluation (SemEval'13)*, Atlanta, Georgia, USA, 2013. ACL.