Coping with the Document Frequency Bias in Sentiment Classification

Abdelhalim Rafrafi, Vincent Guigue, Patrick Gallinari
Computer Science Laboratory (LIP6), University Pierre et Marie Curie, Paris 6
4 place Jussieu, F-75005 Paris, France
{abdelhalim.rafrafi, vincent.guigue, patrick.gallinari}@lip6.fr

Abstract

In this article, we study the polarity detection problem using linear supervised classifiers. We show the interest of penalizing the document frequencies in the regularization process to increase the accuracy. We propose a systematic comparison of different loss and regularization functions on this particular task using the Amazon dataset. Then, we evaluate our models according to three criteria: accuracy, sparsity and subjectivity. The subjectivity is measured by projecting our dictionary and optimized weight vector on the SentiWordNet lexicon. This original approach highlights a bias in the selection of the relevant terms during the regularization procedure: frequent terms are overweighted compared to their intrinsic subjectivities. We show that this bias appears whatever the chosen loss or regularization and on all datasets: it is closely link to the gradient descent technique. Penalizing the document frequency during the learning step enables us to improve significantly our performances. A lot of sentimental markers appear rarely and thus, are unappreciated by statistical learning algorithms. Explicitly boosting their influences leads to increasing the accuracy in the sentiment classification task.

Introduction

Opinion mining (OM) has progressively emerged as a major application domain of text classification. Users being more and more used to provide opinions on websites, opinionated data represent a great opportunity for developing new applications targeting user modeling, e-reputation or recommendation for e-commerce sites. Different resources have been made available to the community, as for example corpora in the domains of e-commerce (Blitzer, Dredze, and Pereira 2007) or movie reviews (Pang, Lee, and Vaithyanathan 2002). The application field being wide and profitable, this explains the keen interest on the subject and the large number of references. An authoritative state of the art prior to 2008 is (Pang and Lee 2008). In this survey, Pang and Lee detail different tasks associated with OM, ranging from sentiment taxonomy to emotional quantification. For many tasks, an essential step is the development of accurate polarity classifiers. Further work (Blitzer, Dredze, and Pereira 2007; Ding, Liu, and Yu 2008; Pang, Lee, and Vaithyanathan 2002; Whitehead and Yaeger 2009) show that sentiment classification is complex and is still open to large improvements. In particular, extracting relevant terms and features is considerably more difficult than for thematic classification. Subjectivity information is more complex and less directly accessible than thematic information that directly relies on the lexical field.

Several works have then focused on the description enrichment and processing for improving polarity detection (Das and Chen 2001; Pang and Lee 2004; Matsumoto, Takamura, and Okumura 2005). They notably focus on negation coding, sentence level analysis, phrase structure coding and part-of-speech feature selection. However, the survey (Pang and Lee 2008) concludes that it is difficult to significantly take advantage from this enrichments (with respect to standard unigrams). (Mejova and Srinivasan 2011) propose a systematic analysis of different variable selection approaches wrt to the size of the datasets and the chosen representations; they point out the difficulty of establishing an universal procedure.

In this paper, we compare two learning formulations (respectively based on hinge loss and least squares) and we rely on the unifying elastic net regularization framework (Zou and Hastie 2005) to select the discriminative terms. The originality of our approach resides in the evaluation metrics: we obviously compute accuracies and sparseness but we also provide a subjectivity analysis of our models. We compare the optimized weights from our models with the subjectivity of the terms in the SentiWordNet lexicon (Esuli and Sebastiani 2006). This study highlights a clear frequency bias in the selection of discriminative terms: in all models, whatever the chosen representation (unigrams or bigrams) or datasets, the terms with high document frequencies are overweighted with respect to their intrinsic subjectivities. As a solution to this problem, we propose a specific regularization framework, which focuses on sentimental markers. This regularizer penalizes terms during training according to their document frequencies in a training set. Our framework remains scalable using a standard stochastic gradient descent and it offers a significant improvement of the accuracy in all cases. The a posteriori analysis of the resulting models shows the alignment between the new weight vectors and SentiWordNet term subjectivities (with respect to their frequencies). We also conclude that regularization highly contributes to the performance even if the resulting

models are never sparse.

Next section describes the models, the evaluation metrics and the datasets. We then present our results in three distinct sections: 1-we demonstrate the existence of the frequency bias and give an in-depth description. 2-we compare the accuracies of standard formulation with the performances of specific models that contend with the frequency bias. 3-we propose an analysis of the regularization role.

Formulations, Evaluation Criteria & Datasets

We consider the problem of sentiment classification, where documents can be either positive or negative (neutral class is removed as in (Pang, Lee, and Vaithyanathan 2002; Blitzer, Dredze, and Pereira 2007; Whitehead and Yaeger 2009)). $\mathbf{X}, \mathbf{Y}$ denote respectively a whole collection of documents and their associated labels. $\mathbf{x}_i$ denotes document i, x_{ij} its jth term and $y_i \in \{+1, -1\}$ its label. N denotes the document collection size and V the dictionary size. Term presence coding is preferred to frequency coding since it leads to higher performance for most reported cases in the literature (Pang and Lee 2008) and for all cases in our experiments: $\mathbf{x}_i \in \{0,1\}^V$. All our experiments are performed with linear classifiers: $f(\mathbf{x}_i) = \sum_{j=0}^{V} x_{ij} w_j, \mathbf{w} \in \mathbb{R}^V$ where w_j is the weight associated to term j. For linear models, $|w_j|$ is a measure of the jth term mean importance in solution f. The sign of f gives the class prediction.

Unified Learning Framework

Learning a classifier can usually be formulated as the following optimization problem:

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}} C(\mathbf{X}, \mathbf{Y}) + \lambda_1 \Omega_{\mathcal{L}_1}(f) + \lambda_2 \Omega_{\mathcal{L}_2}(f) \quad (1)$$

where C denotes a cost function, quantifying the error made by f on the training data and Ω denotes a regularization term, which prevents overfitting (elastic net includes both $\mathcal{L}_1$ and $\mathcal{L}_2$ penalizations), $\mathbf{w}^\star$ is the optimal solution, and λ are the regularization tradeoff. We consider the following loss and regularization functions:

$$\text{Hinge loss:} \quad C_h(\mathbf{X}, \mathbf{Y}) = \sum_{i=0}^{N} (1 - y_i f(\mathbf{x}_i))_+ \quad (2)$$

$$\text{Least squares:} \quad C_{ls}(\mathbf{X}, \mathbf{Y}) = \sum_{i=0}^{N} (y_i - f(\mathbf{x}_i))^2 \quad (3)$$

$$\text{Regul. } \mathcal{L}_2\text{:} \quad \Omega_{(\mathcal{L}_2)}(f) = \sum_{j=1}^{V} w_j^2 \quad (4)$$

$$\text{Regul. } \mathcal{L}_1\text{:} \quad \Omega_{(\mathcal{L}_1)}(f) = \sum_{j=1}^{V} |w_j| \quad (5)$$

This framework unifies notably: linear SVM (Boser, Guyon, and Vapnik 1992), $L1$-SVM (Bradley and Mangasarian 1998), LASSO (Tibshirani 1996), Spline (Tikhonov 1963) and Elastic net (Zou and Hastie 2005).

These baselines are representative of typical families of classifiers with different behaviors. Hinge loss approximates the classification error and focuses on ambiguous documents close to the frontier, whereas least squares is a regression criteria minimizing mean document distance to the frontier. L_2-regularization helps to prevent overfitting while preserving good regression performance. When gradient descent is used for learning, the weight vector $\mathbf{w}$ is updated according to $\mathbf{w} \leftarrow \mathbf{w} - 2\epsilon\mathbf{w}$ (taking into account the L_2 regularization only), and never becomes zero. On the opposite, L_1-regularization acts as a sparsifier. During gradient descent, the L_1 term update is done according to: $\mathbf{w} \leftarrow \mathbf{w} - \epsilon \text{ sign}(\mathbf{w})$. If the sign of w_j changes, the weight is set to 0 (cf. (Friedman et al. 2007)). Each step makes w_j moves towards 0 and each small enough coefficient is set to 0. Note that both regularizers act upon the weights uniformly. Let us now consider the update rule for one of these classifiers, the Spline case for example. The gradient for weight w_j and example x_i is proportional to $(y_i - f(\mathbf{x}_i))x_{ij} - \lambda w_j$. The regularizer decreases all weights uniformly and the weight modification is proportional to x_{ij}, which in our case is 1 if the term is present. As a consequence, weights corresponding to non-frequent terms (x_{ij} = 0 most often) will mainly be affected by the regularization term and decrease towards 0, while weights for frequent terms (x_{ij} = 1 quite often) will see their value increase. This observation holds for any considered classifier. We now proceed with this observation in the following section.

Document Frequency Regularization (DFR)

Our analysis of the classifiers behavior (cf. next sections), suggests that classical term selection or term weighting schemes do not operate properly for sentiment classification. As a consequence, both relevant and irrelevant but frequent terms will influence the classification decision. This is a direct consequence of the usual classification formulation. Based on this observation, we propose to introduce a prior on the regularizer, which will improve the term selection or weighting using a criterion drawn directly from the data. This prior penalizes weights according to their document frequencies: the more frequent a term is, the more penalized it is. This will allow for a better compromise between the regularizer and the classification loss influence and will help to select relevant terms. The formulation of the new regularizer is:

$$\Omega(f) = \sum_{j=1}^{V} \nu_j \Omega_j(f), \; \nu_j = \frac{\#\{\mathbf{x}_i | x_{ij} \neq 0\}}{\#\mathbf{X}} \in [0,1] \quad (6)$$

$\Omega_j(f)$ denotes the component of Ω related to term j. ν_j corresponds to the document frequency of word j in the learning set.

Computing the gradient in this new formulation is straightforward: $\frac{\partial \Omega(f)}{\partial w_j} = \nu_j \frac{\partial \Omega_j(f)}{\partial w_j}$. Comparing with the previous formulation, it is clear that this formulation will help important but rare terms to influence the classifier decision. Low frequency terms will be less impacted and will contribute more to the solution.

Our formulation can be seen as a variant of Confident Weighted Models (Dredze, Kulesza, and Crammer 2010; Crammer, Kulesza, and Dredze 2009), however their approach focuses on adaptive weighting of influent factors whereas, our approach is not adaptive, the penalization is defined once for all, according to the document frequency.

Evaluation Criteria

Accuracy Rate of well classified samples in test. All accuracies are computed using a 10-fold cross validation scheme (90% of the datasets being used to learn the model).

Sparseness Number of non-zero coefficients w_j in the models. This criterion is closely linked with the regularization mechanism. We will also study the values of $\Omega_{(\mathcal{L}_1)}$ and $\Omega_{(\mathcal{L}_2)}$.

SentiWordNet Subjectivity For each term j in our dictionary, we look for its subjectivity in the SentiWordNet lexicon. As a consequence, we get V subjectivity scores $subj_j^{SWN} \in [0,1]$ cf. (Esuli and Sebastiani 2006)[1]. Obviously, the scores of terms that are not found in SentiWordNet are set to 0. This subjectivity is defined for a dictionary (namely a couple dataset/description); it does not depend on the models.

Model Subjectivity The absolute values of weights w_j can be seen as another measure of term subjectivity. Indeed, given that $x_{ij} \in \{0,1\}$, the weight associated to the term j directly influences the final score: the bigger it is (in absolute value), the more subjectivity it introduces. As a matter of fact, a w_j with a great absolute value moves the $f(\mathbf{x}_i)$ away from 0. In the following, we compare our model subjectivity with the SentiWordNet subjectivity.

Datasets & Features

We perform experiments on 4 classical Amazon subsets (Books, DVD, Electronics and Kitchen) (Blitzer, Dredze, and Pereira 2007). The Amazon dataset includes 2 closely related subsets (books and DVD), 1 subset on general product sales (electronics) and 1 *eccentric* subset (kitchen). The vocabulary size is different for all datasets, statistics are given in table 1.

We use only two descriptions: unigrams (U) and unigrams + bigrams (UB). This choice is closely linked to the use of SentiWordNet, which mainly indexes unigrams and bigrams. As already said, we use binary term encoding since it performs better than frequency term coding in all our experiments[2].

Neither stemming nor lemmatization is performed. We use a Part-Of-Speech (POS) filters to keep the following tags: JJ JJR JJS RB RBR RBS NN NNS VB VBD VBG VBN VBP VBZ MD. This corresponds roughly to adjectives, nouns and verbs as in (Turney 2002). Rare words appearing only one time are suppressed.

[1] SentiWordNet offers a subjectivity measure for 117000 terms, which reflect the human perception of their sentimental content. This measure is homogeneous to a probability, namely included in $[0,1]$

[2] Presence coding, which is known to be efficient for the sentiment classification task (Pang and Lee 2008) can be seen as frequency penalization. All terms that appear more than once in a document have less influence using this coding.

Datasets	nb docs (N)	Review length	Vocabulary (V)	
			Uni.	Uni.+Bi.
Books	2000	240	10536	45750
DVD	2000	235	10392	48955
Electronics	2000	154	5611	30101
Kitchen	2000	133	5314	26156

Table 1: Description of the 4 datasets. The vocabulary size depends on the description. Review lengths are averaged.

Gradient Descent Solver & Settings

In our experiments, parameter learning will be solved using mini-batch gradient descent inspired from (Bottou and LeCun 2004). This is a compromise between stochastic and batch gradient procedure: it is more computationally efficient than a batch approach and more robust than a stochastic approach. Moreover the implementation is scalable, robust and the complexity only depends on the mini-batch size, whatever the size of the whole dataset is. In order to preserve scalability in the Document Frequency Regularization frameworks, the ν_j are estimated on the mini-batch.

We set a maximum of 200 iterations (an iteration corresponds to N samples seen), we use mini-batches of size 50 and ϵ is set to 0.001. We also use a decay policy: ϵ is multiplied by 99% at each iteration. We add an early stopping criterion: when no more error is made on the learning set, the algorithm is stopped.

$\mathcal{L}_1$ regularization tradeoff ranges from 0 to 0.005 in 8 logarithmic step ($0,\ 5e-5,\ 1e-4,\ 2e-4,\ 5e-4,\ 1e-3,\ 2e-3,\ 5e-3$). $\mathcal{L}_2$ regularization tradeoff ranges from 0 to 0.5 in 8 logarithmic step ($0,\ 5e-3,\ 1e-2,\ 2e-2,\ 5e-2,\ 1e-1,\ 2e-1,\ 5e-1$).

Frequency Bias

In this section, we propose to compare graphically two subjectivity measures with respect to the document frequencies of words. We aim at showing that the weights of linear models are bigger for frequent terms independently from their intrinsic subjectivities. We also show that our dedicated regularization framework enables us to cope with this phenomenon.

Demonstration of the frequency bias

First, we compute the term distribution over the frequencies (Zipf law): Fig. 1 shows the percentage of terms for each number of occurrences[3]. Words that appear twice represent respectively 33% and 50% of the dictionary for unigrams and bigrams on the dataset Books (all datasets have close behaviors). We can conclude that rare words should play a great role in the decision.

Then, for each document frequency, we compute the average subjectivity according to SentiWordNet and according to our models (cf. previous section for the metrics). On Fig. 2, SentiWordNet metric shows clearly that the subjectivity distribution over the frequencies is approximatively constant

[3] All words appearing more than 30 times are gathered in the last bar of the histogram

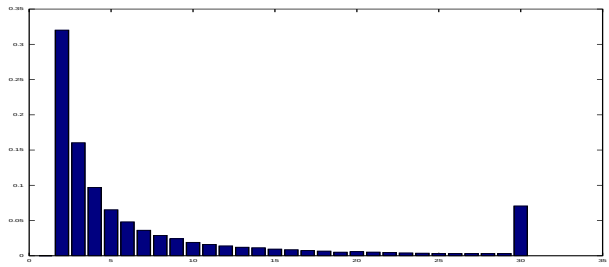
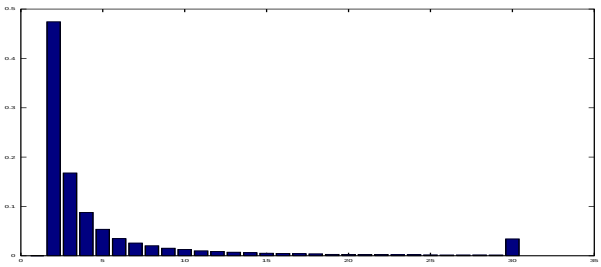

Figure 1: Term distributions over frequencies for unigrams (first plot) and bigrams (second plot) on the dataset *Books*. With unigrams, terms that appear twice represent 33% of the total; with bigrams, they represent over 50% of the dictionary. All words that appear more than 30 times are gathered in the last bar of the histogram.

for unigrams. On the contrary, standard linear models (based on hinge loss and least squares) lead to weight vectors, the absolute values of which increase with the frequencies. This observation holds whatever the dataset or model we consider: this is what we call the frequency bias.

With the bigram description (Fig. 3), our conclusion is less clear. The SentiWordNet subjectivity also increases with the frequency. This phenomenon explanation resides in the combinatorial process of the bigram generation: a lot of rare bigrams are not referenced in SentiWordNet and thus, the subjectivity curve is penalized for low frequency terms. However, we still observe the frequency bias: all standard weight vectors show some picks for terms that appear more than 25 times whereas the subjectivity curve does not offer the same behavior.

Our Document Frequency Regularization framework (DFR) explicitly penalizes high frequencies. As a consequence, it enables us to build models that are closer to the SentiWordNet subjectivity curves. We will show in the next section that the penalization of frequent terms also enables us to improve the accuracies of all our models.

Qualitative analysis of the DFR framework on the dataset books

We propose to sort our dictionary with respect to the weight vector of our models. The greatest weights correspond to words that contribute to positive sentiment whereas lowest weights correspond to words that contribute to negative sentiment. We study here only the dataset books with SVM algorithm (hinge loss and $\lambda_1 = 0$), however, our general observations hold for other models and datasets.

Table 2 confirms that our dedicated regularization framework boosts the weights of low frequency terms: top 100 words from DFR-SVM have a much lower frequency than top 100 words from classical SVM.

Table 3 proposes the top 15 most influent sentimental markers for classical SVM + unigrams, DFR-SVM + unigrams and DFR-SVM + bigrams. We can draw several conclusions from this table: the DFR framework enables us to get rid of frequent terms that are not related to sentiment (e.g. *let, don't, still, also...*) while strong sentimental markers are preserved (e.g. *disappointing, boring,...*). As far as unigrams are concerned, it is difficult to draw another general conclusion; The lists of top words are not comparable and we can not discuss their semantic relevances.

DFR SVM		Classical SVM	
aver. nb. occ. +	aver. nb. occ. −	aver. nb. occ. +	aver. nb. occ. −
11.85	15.04	24.46	249.17

Table 2: Average number occurrences of top 100 words in the learning set with DFR-SVM and SVM on books (Unigrams)

The DFR framework also enables us to extract very relevant bigrams. Bag of words is certainly not an adequate representation for sentiment classification. Enriched representations have been proposed by different authors, but higher dimensionality requires efficient selection and complexity control methods. For this, one needs efficient regularization strategies, and DFR might represent an interesting solution. In the table, terms selected with bigrams are more relevant than those selected with unigrams. Bigrams make it possible to use efficiently quantifiers, punctuation and even negation to make our decision.

Accuracies

Table 4 offers a comparison between all our models with respect to the two representations that we used (unigrams and bigrams). It clearly shows that combining rich term encoding (UB) together with an efficient regularization scheme allows us to bypass the baseline performances on all datasets. For a given standard classifier, using enriched features wrt unigrams enables us to gain between 0.3% and 2.3% of accuracy. Combining these feature representations with DFR offers improvements ranging from 2.2% to 4.8% when compared with the unigrams + classical framework. This series of experiments shows the interest of the DFR framework in term of performances. The DFR version of an algorithm systematically overcomes the standard version (with the same representation). This systematic advantage is all the more significant that the four datasets are very different: their vocabulary sizes and their sentimental markers are known to change a lot.

With our settings, least squares slightly overcome hinge

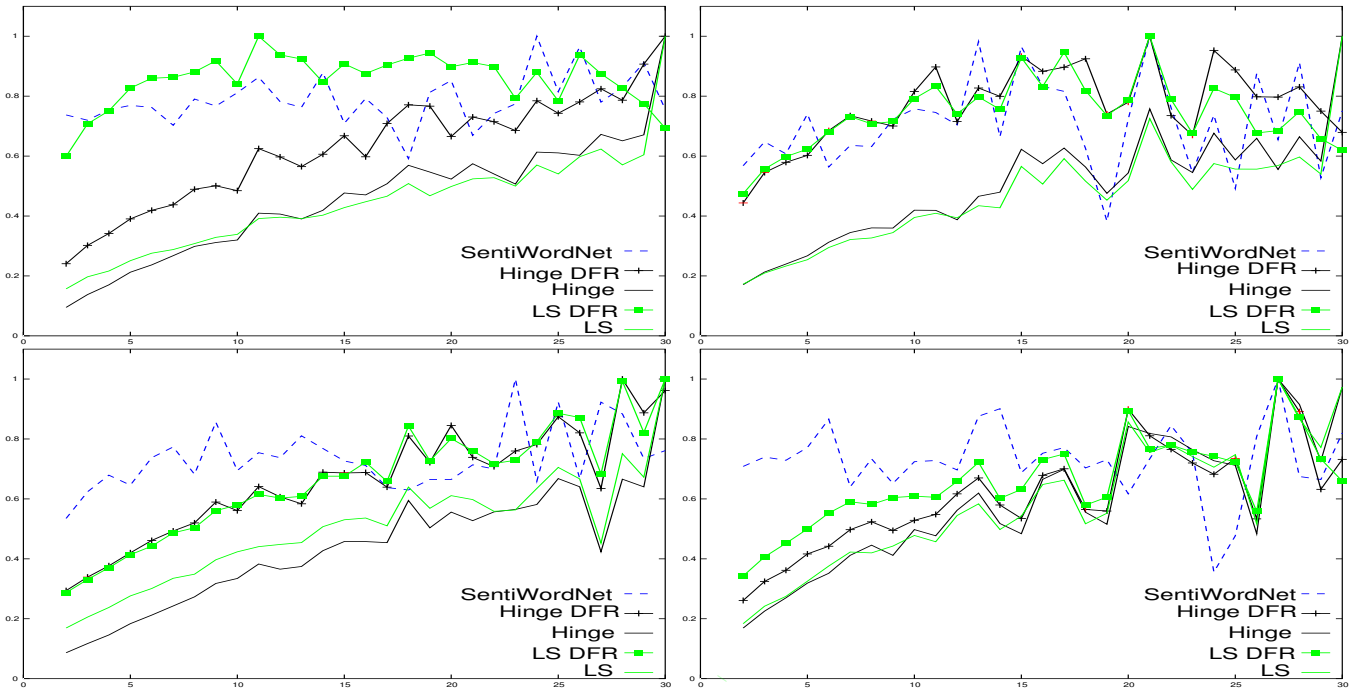

Figure 2: Subjectivities (SentiWordNet & Models) wrt to the word frequencies using unigram description. Datasets from top left to bottom right: books, electronics, DVD, kitchen. For each dataset, we consider the weight vector of the best model in term of accuracy (evaluated in cross-validation). The DFR subjectivity curves (with squares and plus) are systematically closer to the SentiWordNet curve. In the legends, LS stands for least squares and hinge for hinge loss.

loss on all datasets: it is probably due to the high dimensionality of the data. Indeed, we have much more variables than samples (cf. Table 1); it makes our problem very noisy. In this kind of context, least squares, that update models on a correlation criterion, are known to be (slightly) more efficient than L_1 cost functions such as hinge loss.

Regularization

In this section, we discuss the interest of the regularization process as well as the statistical properties of our optimal solutions. We first study the sparseness of our solutions and then we discuss the performances of our models with respect to the regularization.

Sparseness

As we already said, our two regularization processes have not the same behaviors: the $\mathcal{L}_2$ regularization decreases the w_j coefficients without vanishing whereas $\mathcal{L}_1$ regularization acts as a sparsifier. This result is illustrated on Fig. 4: the number of non-zero coefficients decreases quickly when the $\mathcal{L}_1$ regularization tradeoff increases. On the contrary, we observe no sparseness over the whole range of $\mathcal{L}_2$ regularization tradeoff: there is always 100% of non-zeros coefficients when $\lambda_1 = 0$. Combining the two regularizations increases the sparsity: $\mathcal{L}_2$ process weakens the coefficients and $\mathcal{L}_1$ set them to 0.

Our best results are never sparse, we always get more than 99% of non-zero coefficients in our optimal solutions (optimal tradeoffs belongs to the red part of Fig.4). In spite of the proved interest of the regularization (cf. following subsections), we fail at completely vanishing useless coefficients.

Regularization levels in Classical *vs* DFR Frameworks

The regularization operates differently in the classical framework as in the DFR one. The Fig. 5 shows the $\mathcal{L}_2$ ($\sum_j w_j^2$) and $\mathcal{L}_1$ ($\sum_j |w_j|$) criteria for all the models and datasets. We notice that in all cases, DFR optimal models are less regularized (namely, have higher criteria) than classical models. We can conclude roughly that DFR uses more terms than classical models: the DFR process does not eliminate the frequent words; it prevents the elimination of rare terms. Finally, the improvement of the sentiment classification accuracy is closely linked to the contribution of terms that were previously eliminated by the standard approaches. The DFR process acts as a relaxation of the regularization constraints.

We also see that both regularizations of the optimal models are higher with least squares than with hinge loss in all cases.

Accuracies with Bi-Regularization *vs* Mono-Regularization

First, we notice that regularized models perform always better than non-regularized ones in our experiments. In table 4, the given accuracies overcome non-regularized approaches by 0.5% to 2%.

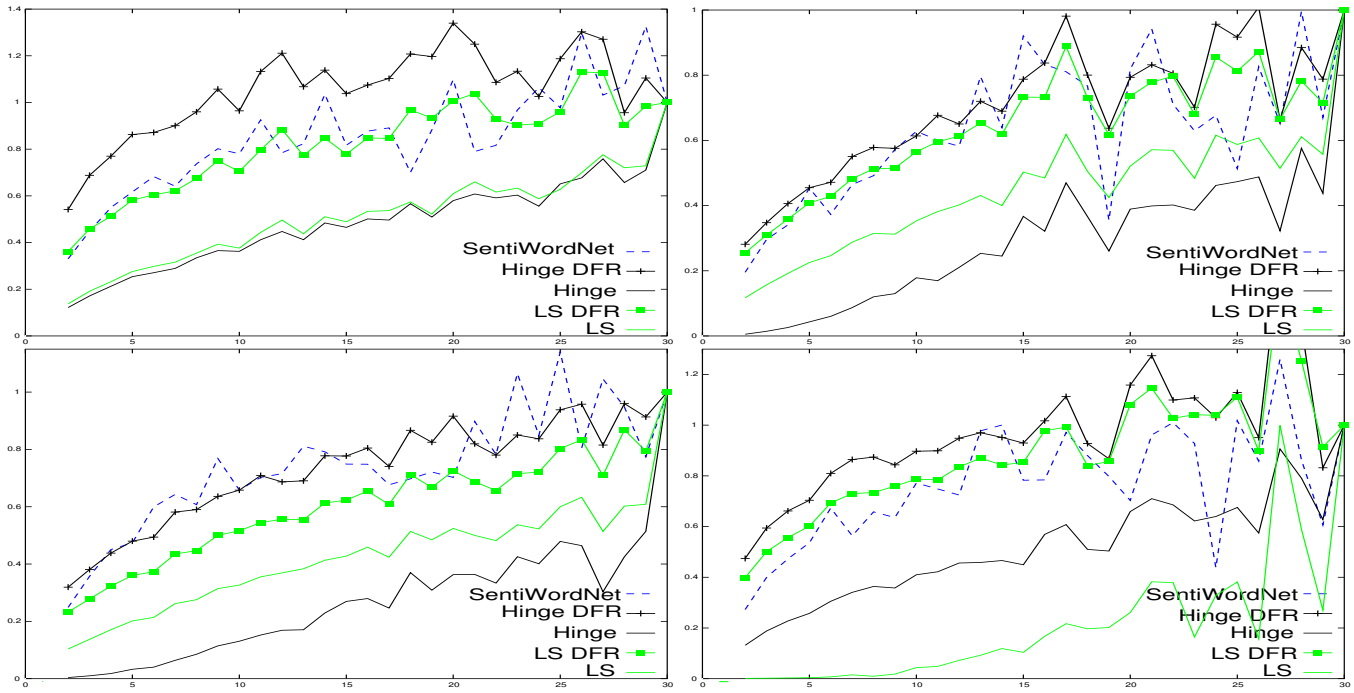

Figure 3: Subjectivities (SentiWordNet & Models) wrt to the word frequencies using **bigram** description. This curves show a slightly different behaviors from the unigrams (cf Fig. 2). Datasets from top left to bottom right: books, electronics, DVD, kitchen. For each dataset, we consider the weight vector of the best model in term of accuracy (evaluated in cross-validation).

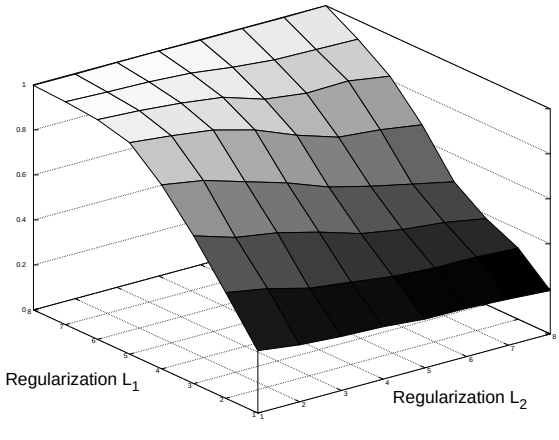

Figure 4: Percentage of non-zero coefficients with respect to $\mathcal{L}_2$ (axis x) and $\mathcal{L}_1$ (axis y) regularization processes on the dataset books (bigrams). x and y axis use logarithmic scale ranging respectively from 0 to 0.5 and from 0 to 0.005.

The double regularization process usually called Elastic Net enables us to overcome the bests results obtained with mono-regularized models as shown on Fig. 6. In table 4, approximatively 75% of the scores come from bi-regularization (the remaining 25% are obtained with one null λ: sometimes λ_1, sometimes λ_2). The gains related to double regularization range from 0 to 1%.

This figure illustrates another classical phenomenon: the $\mathcal{L}_1$ regularization is more sensitive than $\mathcal{L}_2$ regularization. As soon as λ_1 becomes greater, the sparsity increases rapidly and the accuracy falls. Finding the optimal λ_2 value is easier, the parameter is less sensitive.

Conclusion

We have demonstrated the existence of a frequency bias in the optimization process of sentimental classifiers: frequent terms are overweighted even if they are not subjective. On the contrary, a lot of rare terms are eliminated during the regularization process whereas they contain valuable piece of information. This phenomenon occurs with all models and on all datasets.

First, we show that rare terms contain as much subjectivity as frequent ones according to SentiWordNet. Then, we introduce an explicit penalization of the frequent terms to prevent the overweighting. Finally, we obtain a new regularization framework that makes an efficient use of rare terms: the performances of all models benefit from DFR (on all datasets). Although the presence coding is well adapted to sentiment classification, it could be interest-

Classical Model		DFR Framework			
Unigrams		Unigrams		Bigrams	
top −	top +	top −	top +	top −	top +
waste	excellent	disappointing	summer	not_recommend	a_must
unfortunately	introduction	useless	terrific	best_answer	I_sure
boring	wonderful	valuation	bible	save_your	really_enjoyed
worst	enjoyed	outdated	displayed	too_much	read_from
nothing	informative	shallow	editions	skip_this	excellent_book
disappointing	amazing	poorly	refreshing	reference	wow
bad	still	wasted	concise	disappointing	loved_this
terrible	favorite	unrealistic	profession	shallow	gift
money	heart	norm	bike	unless_you	good_reference
don't	stories	hype	shines	way_too	enjoyed_this
better	great	incorrect	coping	was_looking	very_pleased
poorly	familiar	burn	blended	nothing_new	it_helped
let	also	boring	humorous	your_money	terrific
disappointed	thorough	york	lighten	very_disappointing	great_!
instead	definitely	hated	amazed	first_trip	all_ages

Table 3: Compared top 15 words (positive and negative) with classical SVM + unigrams, DFR-SVM + unigrams and DFR-SVM + bigrams on the dataset Books.

		DFR				Baselines			
	Loss	Hinge		Least squares		Hinge		Least squares	
	Features	U	UB	U	UB	U	UB	U	UB
Datasets	Books	83.1	86.0	83.5	**86.9**	82.1	84.3	82.9	83.7
	DVD	82.8	83.9	83.6	**84.4**	82.2	82.9	83.3	83.8
	Electronics	84.4	88.4	86.1	**88.7**	84.0	86.3	85.2	86.3
	Kitchen	86.4	**87.5**	87.0	**87.5**	85.2	86.5	85.7	86.3

Table 4: Best accuracies obtained with respect to the features and the loss function. All accuracies are computed using 10 folds cross-validation. Overall best performances are bolded for each dataset.

ing to (re-)test some alternatives like term-presence-inverse-document-frequency.

In order to study systematically both $\mathcal{L}_1$ and $\mathcal{L}_2$ regularization processes, we relied on the Elastic Net framework: this is original for the sentiment classification problem and we show that this approach is interesting in term of accuracy. Given the dimensionality of the representations, the regularization has a role to play: our main contribution is to make the regularization process more efficient on this task. It enables us to take advantage of larger description and we will try to use richer descriptors in our future experiments. Indeed, this study clearly shows the weaknesses of unigram representations: the fact that unigrams are sometimes preferred to advanced representations is due to computational and optimization problems rather than to objective choices. Finally, our regularization process is efficient but not sparse; we still have to improve this formulation to be able to eliminate useless features.

Acknowledgments This work was supported by the DOXA project and the Cap Digital association (French business cluster for digital content and services in Paris and the Ile de France region).

References

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*.

Boser, B.; Guyon, I.; and Vapnik, V. 1992. An training algorithm for optimal margin classifiers. In *Workshop on Computational Learning Theory*, 144–152.

Bottou, L., and LeCun, Y. 2004. Large scale online learning. In *NIPS*. MIT Press.

Bradley, P., and Mangasarian, O. 1998. Feature selection via concave minimization and support vector machines. In *ICML*, 82–90.

Crammer, K.; Kulesza, A.; and Dredze, M. 2009. Adaptive regularization of weight vectors. In *NIPS*.

Das, S., and Chen, M. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Asia Pacific Finance Association Annual Conference*.

Ding, X.; Liu, B.; and Yu, P. S. 2008. A holistic lexicon-based approach to opinion mining. In *WSDM: International conference on Web search and web data mining*, 231–240.

Dredze, M.; Kulesza, A.; and Crammer, K. 2010. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning Jour.* 79(1–2):123–149.

Esuli, A., and Sebastiani, F. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Conf. on Language Resources and Evaluation*, 417–422.

Friedman, J.; Hastie, T.; Hoefling, H.; and Tibshirani, R.

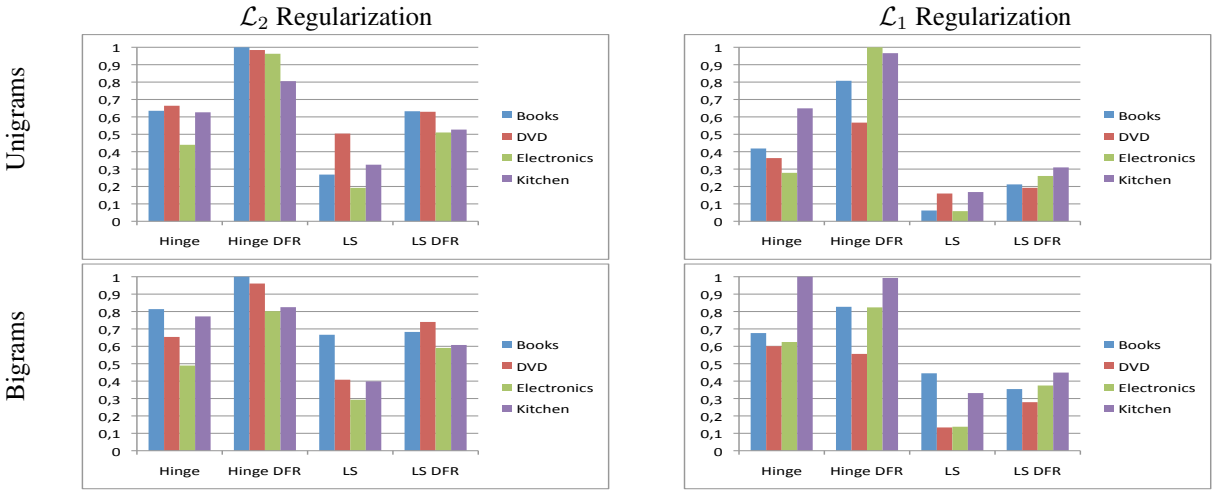

Figure 5: $\mathcal{L}_2$ and $\mathcal{L}_1$ regularization criteria values for optimal unigram and bigram models (cf equations (4) and (5)). As previously, LS stands for least squares and Hinge for hinge loss.

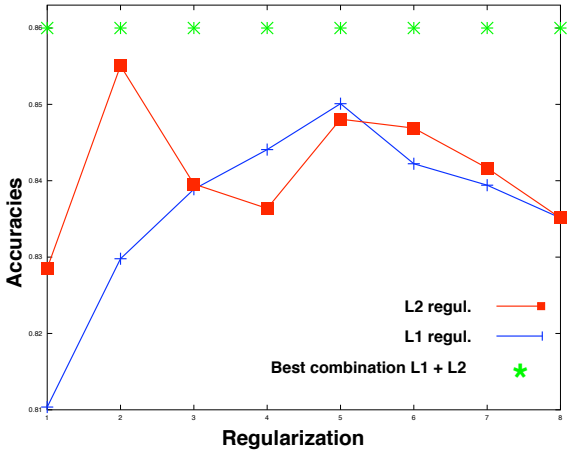

Figure 6: The curve with squares describes the evolution of the performance wrt $\mathcal{L}_2$ regularization tradeoff (with $\lambda_1 = 0$): the maximum accuracy reaches 85.6%. The curve with plus describes the evolution of the performance wrt $\mathcal{L}_1$ regularization tradeoff (with $\lambda_2 = 0$): the maximum accuracy reaches 84.9%. Blue and red curves converge at the last point where $\lambda_1 = \lambda_2 = 0$. The green stars give the performance of the best model, which combines both regularizations (86.0% accuracy). All accuracies are computed using 10 folds cross-validation on the dataset books with hinge loss and UB coding.

2007. Pathwise coordinate optimization. *Annals of Applied Statistics* 1(2):302–332.

Matsumoto, S.; Takamura, H.; and Okumura, M. 2005. Sentiment classification using word sub-sequences and dependency sub-tree. In *PAKDD*.

Mejova, Y., and Srinivasan, P. 2011. Exploring feature definition and selection for sentiment classifiers. In *AAAI ICWSM*.

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*, 271–278.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Information Retrieval* 2:1–135.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *ACL-Empirical Methods in NLP*, volume 10, 79–86.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal Royal Statistics* 58(1):267–288.

Tikhonov, A. 1963. Regularization of incorrectly posed problems. *Soviet Math. Dokl.* 4(6):1624–1627.

Turney, P.D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL*, 417–424

Whitehead, M., and Yaeger, L. 2009. Building a general purpose cross-domain sentiment mining model. In *IEEE World Congress on Computer Science and Information Engineering*, 472–476.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67:301–320.