Argumentative insights from an opinion classification task on a French corpus

Marc Vincent[1] and Grégoire Winterstein[2]

[1] Université Paris 5, UMR S775
marc.r.vincent@gmail.com,
http://www.dsi.unifi.it/~vincent/
[2] Aix Marseille Université, Laboratoire Parole et Langage
and Nanyang Technological University**
gregoire.winterstein@linguist.univ-paris-diderot.fr,
http://gregoire.winterstein.free.fr

Abstract. This work deals with sentiment analysis on a corpus of French product reviews. We first introduce the corpus and how it was built. Then we present the results of two classification tasks that aimed at automatically detecting positive, negative and neutral reviews by using various machine learning techniques. We focus on methods that make use of feature selection techniques. This is done in order to facilitate the interpretation of the models produced so as to get some insights on the relative importance of linguistic items for marking sentiment and opinion. We develop this topic by looking at the output of the selection processes on various classes of lexical items and providing an explanation of the selection in argumentative terms.

Sentiment analysis and opinion mining cover a wide range of techniques and tasks that are oriented towards the classification and extraction of the opinions and sentiments that can be found in a text (see e.g. Pang & Lee (2008) for an extensive review). Interestingly, these tasks are sufficiently different from those of information extraction that they deserve specific approaches. For example, sentiments are seldom expressed overtly in a text, and a keyword based approach for sentiment detection is not very effective, even though it yields some results for information extraction, cf. Cambria & Hussain (2012). One classical task in opinion mining is that of opinion classification. Given a text, the aim of the task is to assign it a label from a pre-determined set (e.g. *positive*, *negative* or *neutral/balanced*). Successful attempts usually involve the use of machine learning techniques, see e.g. Pang et al. (2002) and Pang & Lee (2008).

In this work we pursue two objectives. First, we deal with the task of opinion classification on a corpus of French texts extracted from the web (Sect. 1). We begin by presenting the results of a binary classification task which aims at setting apart positive and negative reviews. In a second experiment, the classification is ternary with the introduction a middle class of neutral (or balanced)

** This research was supported in part by the Erasmus Mundus Action 2 program MULTI of the European Union, grant agreement number 2010-5094-7.

reviews. To enhance the performances of our classifiers we use dimension reduction techniques and show that they have a positive impact.

In the second part of the paper, we try to interpret the output of these selection algorithms from a linguistic point of view (Sect. 2). To carry this out, we look at the elements that "survive" the selection process, and show that there are some similarities between the elements selected in various lexical classes such as coordinating conjunctions, prepositions and adverbs.

1 Opinion classification

1.1 Corpus

The corpus used for the opinion classification task is based on the automatic extraction of 14 000 product reviews taken from three websites that allow their users to post their opinion online. Along with the textual content of the reviews, the score attributed by the users to the product was also extracted. All three websites use a 5 point scale to measure the product quality (1 being the lowest grade and 5 the highest). The origin and number of reviews per grade is given in table 1.

Product type	Source	N. Reviews (per grade)
Hotels	`tripadvisor.fr`	1000
Movies	`allocine.fr`	1000
Books	`amazon.fr`	800

Table 1. Contents of the corpus (total number of reviews: 14 000)

Besides the contents of the reviews and the grade (or score) attributed by the author, we also extracted other information for future use (only from the Amazon and TripAdvisor reviews):

- A one sentence summary of the review (as written by the author of the review).
- A measure of usefulness of the review, indicated by the number of users who judged the review useful.

The TripAdvisor part of the corpus also offers some scores on specific attributes of the hotels reviewed such as the cleanliness of the rooms, the service etc. The complete corpus is available upon request to the authors.

For each grade, the diversity of products and authors was maximized, i.e. one given class of notation contains as many different products and authors as possible. This is to ensure the generality of the models produced by the learning algorithms.

1.2 Classification method

We tried two different opinion classification tasks on this corpora:

1. A binary classification task to differentiate positive and negative reviews. For this task only reviews with a score of 1 (*negative*) and 5 (*positive*) were considered.
2. A ternary classification task with three possible levels: *positive* (scores 4 and 5), *negative* (scores 1 and 2) and *neutral* (score 3).

The set of features used in each task was determined in identical ways:[3]

- Each review was first lemmatized using the state of the art POS tagger and lemmatizer `MElt` by Denis & Sagot (2012).
- Only the lemmas that were successfully recognized by the tagger were used to produce a bag of words representation of the reviews.
- In order to minimize the domain sensitivity of the models produced, all items tagged as proper nouns were removed from the feature set.
- The lemmas that appeared less than 10 times were also ignored because they were deemed too specific.

1.3 Results

The binary classification task was carried out by using three different techniques:

1. *Support Vector Machines* (using `SVM`^{light}, Joachims (1999)).
2. Logistic regression with *elastic net* regularization (cf. Zou & Hastie (2005)).
3. *SVM* on a reduced feature set obtained with the output of the elastic net regularization.

For both tasks the performances were estimated by 10-fold cross validation. The parameters for each approach (i.e the gaussian kernel size and the `c` coefficient for SVM, and (α,λ) for the elastic net regularization) were optimized inside each fold by a subsequent 5-fold cross validation.

The results of each approach are given in table 2. As can be seen, the regularization with elastic net not only greatly reduces the number of initial features (by more than half) but also helps to improve the performance of the classifiers.

Given the results of the binary task, we focused on logistic regression with elastic net regularization for the ternary classification task.

The approach we used is to learn a multinomial logistic regression model with an *elastic net* penalty, meaning that three binary classifiers were produced concurrently, each classifying one class (its associated positive class) against the other two and so that the output class probabilities sum to one. The final classifier is a combination of these three, it predicts the class which is associated

[3] Some approaches use the presence of negation as a feature. This was experimented with, but it did not improve the results and it added a great number of features which slowed down the learning. Therefore it was abandoned.

	N. features	Precision	Recall	F-value
SVM	2829	88.18%	89.54%	88.84
Logistic reg. + *elastic net* sel.	1219	**88.78%**	**91.61%**	**90.16**
SVM + *elastic net* sel.	1219	88.22%	90.32%	89.25

Table 2. Binary classification task: results

to the binary classifier that outputs the highest probability. By analogy to the *one-vs-rest* multiclass approach (cf. Bishop (2006)), we will refer to the performances obtained on each binary classification subtasks as *one-vs-rest* classifier perfomances.

Models were fitted using the `glmnet` package by Friedman et al. (2010) available for the R environment (R Development Core Team (2011))). The results are given in table 3.

	N. features	Precision	Recall	F1	F1 μ
1,2 vs. 3 vs. 4,5	2082.5	63.56	61.35	60.11	69.94
1,2 vs. *rest*	1147.3	71.52	80.57	75.77	-
3 vs. *rest*	645.2	46.80	18.39	26.35	-
4,5 vs. *rest*	1026.6	72.37	85.09	78.20	-

Table 3. Ternary classification task: results. Precision, recall and F1-score are macro-averaged over classes. The micro averaged F1-score is also given (F1 μ). All measures are averaged over the 10 final test folds.

1.4 Discussion

The results of table 2 show the great benefit in using feature selection techniques both for reasons of dimension reduction and for the improvement of the final performance. The results prove superior to the baselines usually reported for English (e.g. by Pang et al. (2002), who report a $F1$ score of about 83 on a similar task).

The results of the ternary task appear poorer. This is essentially due to the poor performance of the *3 vs. rest* classifier who sports a very low recall. To explain these poor performances we had a closer look at the reviews scored 3 by the users and manually re-labeled them. This manual reclassification was done on a subset of 1 667 reviews (mainly for reasons of time) and it led to the reclassification of about 24% of the reviews. This means that on average, one review out of four that was labeled "neutral" because of its grade of 3 was manually reassigned either to the "positive" or "negative" class due to its content. The table 4 gives the number of reviews with score 3 whose labels were manually checked, for each of part of the corpus.

Manual tag	`Allocine`	`Amazon`	`TripAdvisor`	**Total**
Neutral	458	418	394	1270
Negative	15	36	61	112
Positive	97	135	53	285

Table 4. Manual classification of a subset of reviews with a score of 3

It is worth noting that depending on the origins of the reviews, the relabeling is different. Reviews from Allocine and Amazon were mainly done from the neutral towards the positive, with much fewer towards the negative. In contrast, reviews from TripAdvisor are roughly equally divided between the positive and negative.

F1	1,2 vs. *rest*	3 vs. *rest*	45 vs. *rest*	12 vs. 3 vs. 45
AMZ	73.39	19.27	76.13	56.26
TA	83.62	35.50	84.41	67.84
AC	69.81	21.43	73.90	55.05

Table 5. Performances by corpus measured by the F1 score (macro averaged for the ternary classification problem). AMZ: Amazon reviews, TA: TripAdvisor reviews, AC: Allocine reviews

Decomposing the classification performances of our classifiers with respect to the source of the test data provides some further insights into the meaning of these relabeling statistics. As showed in table 5, the three corpora are not equal in that matter, with TripAdvisor reviews being clearly better classified than reviews of the other two sources. Although the differences in lexicons used to characterize cultural products and travel accommodations may be responsible, this also correlates intriguingly well with the previous observation that while "mistaken" labels of Amazon and Allocine are biased toward the positive class, those same labels are fairly well balanced between positive and negative class for TripAdvisor. This provides further motivation to investigate the effect of potential mislabeling.

To assess the usefulness of our recoding, we ran a quick comparison between the results of the ternary classification task using a maximum entropy algorithm (using the `megam` software, Daumé III (2004)). The choice was mainly due to the fact that this algorithm is fast, even if less efficient than the other techniques we used. The results comparing the overall performance before and after the manual relabeling are given in table 6.

Although the improvement is small, it seems that manually reclassifying the reviews with a score of 3 has a positive effect on the classification task. We therefore ran a classification similar as the previous one on the relabeled corpus. The results are however not up to the expectations

	Precision	Recall	F1
Original corpus	64.03	48.45	55.16
Partially manually relabeled corpus	64.52	49.74	56.00

Table 6. Ternary classification task: original vs. partially relabelled corpus (Max. Ent. algorithm)

	N. features	Precision	Recall	F1	F1 μ
1,2 vs. 3 vs. 4,5	2136	64.67	61.05	59.94	72.19
1,2 vs. *rest*	1171.1	73.26	81.35	77.07	-
3 vs. *rest*	632.5	46.51	15.23	22.84	-
4,5 vs. *rest*	1083	74.23	86.56	79.90	-

Table 7. Ternary classification task after reclassification of reviews with a score 3: results. Precision, recall and F1-score are macro-averaged over classes. The micro averaged F1-score is also given (F1 μ) All measures are averaged overed the 10 final test folds.

As can be seen from table 7 by looking at macro scores, the prediction performance of the three classes altogether is almost identical to the one before the relabeling. In detail, looking at binary classifiers it is more difficult to predict the partially re-labelled third class while the prediction of the two other classes improves, although by a smaller margin. Looking at the micro-F1 scores gives the other side of the story, i.e. since these latter classes are more populated, the modest improvement in their prediction performances is sufficient to increase the number of instances that are correctly classified. Still, our original goal to improve the classification of the middle class is not satisfied and we lose in recall compared to what we had before reencoding (while keeping the same precision).

However, it must be remembered that all these observations have to be relativized by the fact that only 1 667 of the 2 800 reviews scored 3 have been manually checked. This means that about 300 reviews are still incorrectly labeled as "neutral". The completion of the manual check should therefore further help to improve the classifiers. Another factor to take into account is that the classes are fairly imbalanced while the models produced aim at producing an overall best fit, therefore favorishing these classes. It is possible that giving a reasonably bigger weight to the middle class examples would help improve the classification of its instances.

On a final note, it should however be noted that even with the poor performance of one of the classifiers, the global classifier achieves results that can be compared with the usual baselines for French on this particular task (cf. the results of the DEFT'07 challenge by Grouin & al. (2007) who report a $F1$ of 60.3 on a corpus that is comparable, although less general in the range of topics covered).

2 Interpreting the models

The reviews that form the corpus used here express more than just opinions: they are also *argumentative* because they (usually) provide rationales to back up the opinions expressed by their authors. Therefore, the study of these reviews can be of some interest for the study of the way people use argumentative connectives and schemes to convey a positive or a negative opinion. One of the upshots of the elastic net regularization is a reduction of the feature space that retains only those features that are relevant for the classification task. Therefore, by studying the features that come out of this feature selection phase, one can try to get an idea of the argumentative strategies employed by the authors, or at least use them as a way to profile classes of expressive items (in the same vein as has been done by Constant et al. (2009) on the topic of expressive items, although with a different methodology).

We also compared the output of the selection derived from the elastic net with another selection method based on *bootstrapping*. This section first introduces the technique of bootstrapping and then presents the output of the selection processes by distinguishing between elements belonging to open categories and those belonging to closed categories.

2.1 Bootstrapping

Following the general method of the non parametric bootstrap (cf. Efron (1979)), 200 bootstrap samples were generated by sampling with replacement from the original dataset. Sampling was done for each stratum separately to ensure that bootstrap samples had the same number of examples from each stratum as in the original dataset. Logistic regression models were learnt from each of these samples, using an elastic net penalization with α and λ parameters chosen using a 3 folds cross-validation. Distribution of each regression coefficient β_i with respect to the bootstrap samples is used to qualify the robustness of the corresponding feature f_i.

The results of the bootstrap offer a way to test the robustness of a feature: if the feature gets consistently selected over the samples, this means that its contribution is general. Therefore, to determine the general relevance and impact of a feature, we first begin by looking at the percentages of bootstrap samples where its coefficient is non null. If this percentage is high enough, we look at two values:

1. the value of the coefficient coming from the elastic net regularization
2. the average of the coefficients coming from each of the bootstrap samples

It is expected that these two values are rather similar, but for reasons of completeness we report both of them in the following tables.

2.2 Closed categories

In this section we focus on two specific closed categories: coordinating conjunctions on one hand and prepositions on the other. The elements in those classes

are few in number and usually very frequent. We are thus mainly interested in knowing which of these elements are the most relevant for the classification task.

Coordinating Conjunctions Coordinating conjunctions are obvious discourse connectives, and as such it is interesting to check which of those prove to be the most relevant for sentiment analysis.

We begin by looking at the binary classification task. Table 8 shows for each conjunction: the proportion of bootstrap samples where its coefficient was not null, the average of its bootstrap coefficient, its elastic net coefficient and the number of occurrences of the conjunction in the corpus.

Conjunction	Proportion	Bootstrap avg.	Elastic Net	N. occ.
et	0.97	0.157	0.139	4284
ou	0.27	0.019	0.0	864
donc	0.15	-0.049	0.0	11
sinon	0.30	-0.052	0.0	44
voire	0.23	-0.069	0.0	33
soit	0.4	-0.095	-0.028	69
car	0.73	-0.103	-0.076	549
puis	0.57	-0.129	-0.075	120
mais	1	-0.335	-0.245	1889
ni	0.99	-0.511	-0.464	169
or	0.83	-0.528	-0.693	21

Table 8. Coordinating conjunctions: coefficients selection (Binary task).

Only four conjunctions seem to have a significant contribution here, i.e. get selected in more than 75% of the bootstrap samples and have non-null coefficients. On the positive side there is the conjunction *et* (≈*and*), while on the negative are *mais* (≈*but*), the correlative *ni* (≈*neither/nor*) and the adversative *or* (≈*yet/as it turns out*).[4]

The presence of the negative *ni* is expected to be correlated with negative reviews as it has an intrinsically negative meaning.

Mais and *or* can be grouped together: they both are adversative, i.e. they introduce a sentence that is opposed in one way or another to the left argument of the connective. From the argumentative point of view, it is considered that these items connect opposed arguments (cf. Anscombre & Ducrot (1977), Winterstein (2012)).

On the other hand *et* has been described as a connective that conjoins two arguments that argue for the same goal and are (at least) partly independent arguments for this goal (cf. Jayez & Winterstein (2013)).

Therefore, it seems that negative reviews tend to involve opposed arguments more often than positive reviews (as marked by the significance of adversative

[4] We ignore the borderline case of *car* (≈ *because/since*).

connectives for these reviews). On the other hand, positive reviews involve sequences of arguments that target the same goal, but are independent (as marked by *et*).

One way to interpret this is to consider that positive and negative reviews involve different argumentative strategies. Arguing positively requires more effort to convince. A successful positive argumentation will have more chance of being persuasive if it gives several independent arguments in favor of its conclusion. On the other hand, in order to argue negatively, a single negative argument appears to be enough, even if it is put in perspective with a positive one.

This interpretation is further confirmed if one looks at the results of a Naive Bayes approach to the classification task. While such an approach does not give results as good as those reported in table 2, it can easily be used to detect bigrams that are correlated to positive or negative reviews. Among the ten bigrams whose significance for the classification is the highest one can find the bigram *point positif* ("positive point"). Contrary to what could be expected, this bigram is a strong indicator of a *negative* review. This is in line with our previous observation on the use of *but*: mentioning a positive point usually entails also mentioning a negative one. In case of conflicting arguments, it is expected that the negative one will win.

We now turn to the ternary task. Tables 9 and 10 present the same information as table 8 but for the ternary task. Table 9 presents the results for the classifier of the positive class (scores 4 and 5) against the rest, whilst table 10 is for the classifier of the negative class (scores 1 and 2) against the rest. Features for which the elastic net coefficient is null have been omitted from the tables. The middle classifier is ignored because of its poor performances (cf. the discussion on table 7 above).

Conjunction	Proportion	Bootstrap avg.	Elastic Net	N. occ.
et	1	0.108	0.121	10693
car	1	-0.155	-0.075	1539
puis	0.975	-0.207	-0.049	329
mais	1	-0.367	-0.203	6003
ni	0.98	-0.228	-0.003	407

Table 9. Coordinating conjunctions: coefficients selection (Ternary task, positive classifier).

For the positive classifier, we see that the positive role of the conjunction *et* remains: our hypothesis about the preference to use additive argumentative strategies in positive reviews appears confirmed. However the case of *mais* has to be somehow refined. The adversative is still an indicator of a non-positive review as seen in table 9, but it is no longer an indicator of a negative one. Therefore we can still consider that positive reviews tend to eschew balancing positive and negative arguments, but we can no longer assume that this is the

Conjunction	Proportion	Bootstrap avg.	Elastic Net	N. occ.
or	0.99	0.647	0.664	39
voire	0.95	0.344	0.135	89
puis	0.98	0.224	0.135	329
ni	0.1	0.247	0.094	407
comme	0.92	-0.73	-0.328	12

Table 10. Coordinating conjunctions: coefficients selection (Ternary task, negative classifier).

hallmark of negative reviews. This appears quite sensible: middle reviews should form the prototypical case of balanced arguments and thus are good candidates for involving the use of adversative markers. However negative reviews still use adversative strategies: the adversative connective *or* is the strongest indicator for the negative class amongst all conjunctions.

In the end, the study of the output of the feature selection processes on the case of conjunctions outlines the fact that positive, balanced and negative reviews do not use the same argumentative schemes. Further investigation of the reviews, for example at the sentence level and by using a polarity lexicon such as Senticnet (cf. Cambria & Hussain (2012)), should help to strengthen these claims.

Prepositions We look here at prepositions in the same perspective as the conjunctions. To keep the presentation short, we only present the results of the binary task in table 11 where we only mention those for which both selection methods produced non-null coefficients.

Preposition	Proportion	Bootstrap avg.	Elastic Net	N. occ.
avec	0.97	0.165	0.139	1804
chez	0.58	0.131	0.115	167
selon	0.42	0.101	0.005	74
en	0.88	0.087	0.051	2537
pour	0.67	0.051	0.025	2736
sous	0.39	-0.053	-0.050	210
de	0.76	-0.083	-0.075	4879
jusque	0.5	-0.086	-0.031	193
envers	0.3	-0.089	-0.098	30
sans	1	-0.340	-0.264	1035
malgré	0.94	-0.344	-0.281	208
sauf	0.96	-0.511	-0.416	106

Table 11. Prepositions: selection coefficients (Binary task)

The main point we wish to underline here is that the selection is consistent with that of the coordinating conjunctions. On one hand the additive preposition *avec* (≈*with*) is an indicator of positive reviews, like the additive conjunction *et*. This remains true in the ternary classification task for the positive classifier.

Regarding the prepositions that have a negative impact, the case of the adversative *malgré* (≈*in spite of*) appears similar to the adversative conjunctions of the previous sections. It is a marker of negative review in the binary task, and in the ternary task it is a mark of a non-positive review, but it does not specifically mark negative reviews.

Finally, the case of *sauf* (≈*except*) and *sans* (≈*without*) also prove to be interesting. These two prepositions have the same profile as the adversative elements. So far these elements have not been described in these terms, but the results presented here suggest that these elements might also be appropriately be described in argumentative terms as carrying an adversative value. Roughly both these prepositions are exceptive, i.e. they indicate that an element is not included in some predication. If the excepted element is important, then it is expected that the use of these prepositions carries an argumentative reversal effect similar to what the exclusive adverb *only* conveys in some contexts.

2.3 Open categories

We briefly focus here on the case of elements belonging to open categories, i.e. on lemmas that are either verbs, nouns, adjectives or adverbs.

Class distribution First, we look at how the relative importance of each of the four open categories is affected by the selection process. For this we look at the number of items in each class before and after the selection process. The numbers in table 12 correspond to the binary task. The number of items after selection correspond to the number of items which have non null coefficients in more than 75% of the bootstrap samples and for which the elastic net coefficient is not null.

Category	Before selection	Proportion	After Selection	Proportion
Adjective	543	20.03%	148	35.41%
Adverb	158	5.83%	28	6.7%
Noun	1398	51.57%	164	39.23%
Verb	612	22.57%	78	18.66%
Total	2711	100%	418	100%

Table 12. Open categories: number of items before and after selection.

The differences in the distribution of the categories before and after the selection are quite significant ($\chi^2 = 57.71, p\text{-value} \ll 1.0^{-10}$). They show that

adverbs and adjectives are more represented after the selection process, whereas nouns and verbs see a strong decrease in their frequencies. This strongly supports the opinion that adjectives and adverbs are the most likely elements to convey sentiment in a text, as has been claimed previously (e.g. by Turney (2002) or Benamara et al. (2007)).

Adverbs For reasons of space and relevance, we only develop the case of adverbs here. Table 13 gives the list of the 28 adverbs that were selected.

Adverb	Proportion	Bootstrap avg.	Elastic Net	N. Occ.
magnifiquement	0.99	1.44	1.05	13
agréablement	0.92	0.71	0.70	21
bientôt	0.97	0.78	0.69	34
absolument	1.00	0.62	0.42	253
très	1.00	0.51	0.37	2247
vivement	0.96	0.43	0.37	118
bien	1.00	0.37	0.29	1543
toujours	0.95	0.25	0.21	406
aussi	0.99	0.25	0.21	588
peu	0.85	-0.15	-0.09	710
même	0.93	-0.18	-0.16	749
là	0.87	-0.22	-0.17	358
mieux	0.95	-0.31	-0.24	319
totalement	0.87	-0.32	-0.26	130
bref	0.96	-0.30	-0.27	212
alors	0.98	-0.31	-0.28	381
sûrement	0.87	-0.46	-0.34	53
ne	1.00	-0.40	-0.35	2915
franchement	0.89	-0.33	-0.36	129
vite	0.97	-0.48	-0.40	128
non	1.00	-0.54	-0.40	421
pourtant	0.99	-0.53	-0.41	182
pas	1.00	-0.66	-0.49	2791
plutôt	0.99	-0.59	-0.50	211
trop	1.00	-0.82	-0.62	465
strictement	0.90	-1.03	-0.73	16
heureusement	1.00	-1.08	-0.88	102
mal	1.00	-1.21	-0.93	308

Table 13. Selected adverbs (Binary task)

Negation A first feature to be noted is that these results are consistent with those found about English by Potts (2011) concerning the "negativity" of negation. Potts underlines that negation is more than just a logical switch for truth-values, but also seems to carry an intrinsic negative tone. He shows how this is

confirmed by the fact that the distribution of negative markers is not homogeneous across notations: elements like *not* appear more often in negative reviews than in positive ones. According to Potts, this is explained by the fact that a negative sentence is usually less informative than a positive one, and is thus more likely to be used as a rebuttal rather than as an informative statement.

We find the same situation for French in the reviews of our corpus: the markers of negation *ne* and *pas* both appear as strong indicators of negative reviews, selected in all the bootstrap samples, with relatively strong coefficients.

From the methodological point of view our approach slightly differs from the one of Potts. In both cases, we adopt a reader-oriented perspective: given a lexical item, we evaluate which kind of opinion is the most likely (i.e. positive or negative), so the general goal is the same for Potts and us.

However, Potts typically starts by selecting some elements which he assumes have a specific profile and uses the data as a way to confirm his hypotheses (e.g. as was done with negation). In a related work and using similar data, Constant et al. (2009) use the profiles of known elements as a way to discover other elements which share the same profile, aiming at automatically discovering classes of expressive elements.

Our approach is different in that we do not make any preliminary assumption on the profile of lexical items. The learning and selection processes automatically provide us with classes of elements which behave similarly regarding the task at hand. One drawback is that the classes have to be coarser than the ones one can obtain by Potts's approach. This comes as a consequence of the fact that predicting notations beyond the positive/negative dual case is difficult (cf. the discussion on the ternary classification task). This means that elements that are not characteristic of either the positive or negative notation class will be harder to detect since this implies dealing with three or more classes of notation.

Nevertheless, the fact that our approach and that of Potts give similar results for negation gives credence to both as ways to get some pragmatic insights by relying on large corpora and meta-textual data.

Other elements Apart from the case of negation other features of table 13 appear striking:

- Positivity intensifying adverbs such as *magnifiquement* (≈*beautifully*) or *agréablement* (≈ *pleasantly*) are strong positive indicators. This is expected since those elements are non-controversially positive.
- *Heureusement* (≈ *fortunately*) might appear as a counter-example because of its apparent positive undertone. However, the use of this adverb usually marks a counter expectation akin to an adversative reading. This is again consistent with our observations on conjunctions and prepositions. The presence of *pourtant* (≈*yet*) as a negative indicator is also coherent.
- Purely intensifying adverbs have mixed profiles:
  - *très* (≈ *very*) and *absolument* (≈ *absolutely*) are positive indicators.
  - *totalement* (≈ *totally*) is negative.

Initially, one could have thought that those intensifying adverbs have no polarity bias since their essential meaning is to indicate a high degree of the property they modify, without further constraints on the kind of property it can act on (in the same way that one could initially expect negation to have no specific orientation on its own). However, it seems that speakers have preferences for using some adverbs for intensifying positive properties and others for negative properties. We leave the question of why this happens to future work.

– Finally, the adverb *aussi* ($\approx$ *too*) is often described as being additive (e.g. by Winterstein & Zeevat (2012)) and is shown here to be a positive marker. This is consistent with the previous hypothesis about positive reviews involving multiple parallel arguments since *aussi* indicates that the speaker is using two sentences that are related and argumentatively co-oriented.

3 Conclusion

The work reported here underlined the importance and usefulness of feature selection techniques when tackling a problem like opinion classification. Not only do these techniques improve the performances of the classifiers, they also offer some insight on the way the classifiers work and on which elements have the same profile regarding the task at hand.

In future work, we intend to try to further enhance the ternary classifier. First, we will complete the manual check of the reviews scored 3 to get a final evaluation of the usefulness of this relabeling. Another improvement direction is in the detection of irony which appears rather common, especially in negative reviews. Finally, an analysis of the reviews at the sentence level, by using a polarity lexicon, is also a potential solution to improve the performances. However, such a resource is not readily available yet for French and needs to be constructed beforehand.

Another direction of research is testing the general character of the argumentative strategies we have characterized. First, we intend to determine whether the same conclusions can be reached on other languages, notably English for which resources of the type we need already abound. Another fruitful comparison is to compare our results with insights gathered on reviews of a different kind. For example, reviews of scientific papers might exhibit different profiles. There we would expect negative reviews to be more thorough and involve parallel, independent negative arguments. This is because, at least intuitively, a negative review should be strongly motivated and usually cannot be reduced to a single negative point.

Finally, regarding the interpretation of the models, a further investigation of the ternary models should be carried out once they have been improved, especially regarding the middle classifier.

References

Jean-Claude Anscombre, Oswald Ducrot (1977). "Deux *mais* en français". In: *Lingua 43*, pp. 23–40.

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, V.S. Subrahmanian (2007). "Sentiment analysis: Adjectives and adverbs are better than adjectives alone". In: *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Christopher M. Bishop (2006). *Pattern Recognition and Machine Learning*. Berlin: Springer.

Erik Cambria, Amir Hussain (2012). *Sentic Computing. Techniques, Tools, and Applications*. Berlin: Springer.

Noah Constant, Christopher Davis, Christopher Potts, Florian Schwarz (2009). "The Pragmatics of Expressive Content: Evidence from Large Corpora". In: *Sprache und Datenverarbeitung 33*, 1–2, pp. 5–21.

Hal Daumé III (2004). "Notes on CG and LM-BFGS Optimization of Logistic Regression". Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`.

Pascal Denis, Benoît Sagot (2012). "Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging". In: *Language Resources and Evaluation 46*, pp. 721–746.

Bradley Efron (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics 7*, pp. 1–26.

Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software 33*, 1, pp. 1–22. URL http://www.jstatsoft.org/v33/i01/.

C. Grouin, al. (2007). "Présentation de l'édition 2007 du Défi fouille de textes (DEFT07)". In: *Actes de l'atelier de clôture du 3ème Défi Fouille de Textes (DEFT07)*. pp. 1–8.

Jacques Jayez, Grégoire Winterstein (2013). "Additivity and Probability". In: *Lingua 132*, pp. 85–102.

Thornsten Joachims (1999). "Making large-Scale SVM Learning Practical". In: Bernhard Schölkopf, Christopher J. C. B Burges, Alexander J. Smola (eds.), *Advances in Kernel Methods - Support Vector Learning*, MIT Press, pp. 41–56.

Bo Pang, Lillian Lee (2008). "Opinion mining and sentiment analysis". In: *Foundations and Trends in Information Retrieval 2*, 1–2, pp. 1–135.

Bo Pang, Lillian Lee, Shivakumar Vaithyanathan (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 79–86.

Christopher Potts (2011). "On the Negativity of Negation". In: Nan Li, David Lutz (eds.), *Semantics and Linguistic Theory (SALT) 20*. eLanguage, pp. 636–659.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Peter Turney (2002). "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews". In: *Proceedings of the ACL*.

Grégoire Winterstein (2012). "What *but*-sentences argue for: a modern argumentative analysis of *but*". In: *Lingua 122*, 15, pp. 1864–1885.

Grégoire Winterstein, Henk Zeevat (2012). "Empirical Constraints on Accounts of *too*". In: *Lingua 122*, 15, pp. 1787–1800.

Hui Zou, Trevor Hastie (2005). "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society Series B*, pp. 301–320.