Determining Term Subjectivity and Term Orientation for Opinion Mining

Andrea Esuli[1] **and Fabrizio Sebastiani**[2]
(1) Istituto di Scienza e Tecnologie dell'Informazione – Consiglio Nazionale delle Ricerche
Via G Moruzzi, 1 – 56124 Pisa, Italy
andrea.esuli@isti.cnr.it
(2) Dipartimento di Matematica Pura e Applicata – Università di Padova
Via GB Belzoni, 7 – 35131 Padova, Italy
fabrizio.sebastiani@unipd.it

Abstract

Opinion mining is a recent subdiscipline of computational linguistics which is concerned not with the topic a document is about, but with the opinion it expresses. To aid the extraction of opinions from text, recent work has tackled the issue of determining the *orientation* of "subjective" terms contained in text, i.e. deciding whether a term that carries opinionated content has a positive or a negative connotation. This is believed to be of key importance for identifying the orientation of documents, i.e. determining whether a document expresses a positive or negative opinion about its subject matter.

We contend that the plain determination of the orientation of terms is not a realistic problem, since it starts from the non-realistic assumption that we already know whether a term is subjective or not; this would imply that a linguistic resource that marks terms as "subjective" or "objective" is available, which is usually not the case. In this paper we confront the task of deciding whether a given term has a positive connotation, or a negative connotation, *or has no subjective connotation at all*; this problem thus subsumes the problem of determining subjectivity *and* the problem of determining orientation. We tackle this problem by testing three different variants of a semi-supervised method previously proposed for orientation detection. Our results show that determining subjectivity *and* orientation is a much harder problem than determining orientation alone.

1 Introduction

Opinion mining is a recent subdiscipline of computational linguistics which is concerned not with the topic a document is about, but with the opinion it expresses. Opinion-driven content management has several important applications, such as determining critics' opinions about a given product by classifying online product reviews, or tracking the shifting attitudes of the general public toward a political candidate by mining online forums.

Within opinion mining, several subtasks can be identified, all of them having to do with tagging a given document according to expressed opinion:

1. *determining document subjectivity*, as in deciding whether a given text has a factual nature (i.e. describes a given situation or event, without expressing a positive or a negative opinion on it) or expresses an opinion on its subject matter. This amounts to performing binary text categorization under categories Objective and Subjective (Pang and Lee, 2004; Yu and Hatzivassiloglou, 2003);

2. *determining document orientation* (or *polarity*), as in deciding if a given Subjective text expresses a Positive or a Negative opinion on its subject matter (Pang and Lee, 2004; Turney, 2002);

3. *determining the strength of document orientation*, as in deciding e.g. whether the Positive opinion expressed by a text on its subject matter is Weakly Positive, Mildly Positive, or Strongly Positive (Wilson et al., 2004).

To aid these tasks, recent work (Esuli and Sebastiani, 2005; Hatzivassiloglou and McKeown, 1997; Kamps et al., 2004; Kim and Hovy, 2004; Takamura et al., 2005; Turney and Littman, 2003) has tackled the issue of identifying the orientation of subjective *terms* contained in text, i.e. determining whether a term that carries opinionated content has a positive or a negative connotation (e.g. deciding that — using Turney and Littman's (2003) examples — `honest` and `intrepid` have a positive connotation while `disturbing` and `superfluous` have a negative connotation).

This is believed to be of key importance for identifying the orientation of documents, since it is by considering the combined contribution of these terms that one may hope to solve Tasks 1, 2 and 3 above. The conceptually simplest approach to this latter problem is probably Turney's (2002), who has obtained interesting results on Task 2 by considering the algebraic sum of the orientations of terms as representative of the orientation of the document they belong to; but more sophisticated approaches are also possible (Hatzivassiloglou and Wiebe, 2000; Riloff et al., 2003; Wilson et al., 2004).

Implicit in most works dealing with term orientation is the assumption that, for many languages for which one would like to perform opinion mining, there is no available lexical resource where terms are tagged as having either a Positive or a Negative connotation, and that in the absence of such a resource the only available route is to generate such a resource automatically.

However, we think this approach lacks realism, since it is also true that, for the very same languages, there is no available lexical resource where terms are tagged as having either a Subjective or an Objective connotation. Thus, the availability of an algorithm that tags Subjective terms as being either Positive or Negative is of little help, since determining if a term is Subjective is itself non-trivial.

In this paper we confront the task of determining whether a given term has a Positive connotation (e.g. `honest`, `intrepid`), or a Negative connotation (e.g. `disturbing`, `superfluous`), or has instead no Subjective connotation at all (e.g. `white`, `triangular`); this problem thus subsumes the problem of deciding between Subjective and Objective *and* the problem of deciding between Positive and Negative. We tackle this problem by testing three different variants of the semi-supervised method for orientation detection proposed in (Esuli and Sebastiani, 2005). Our results show that determining subjectivity *and* orientation is a much harder problem than determining orientation alone.

1.1 Outline of the paper

The rest of the paper is structured as follows. Section 2 reviews related work dealing with term orientation and/or subjectivity detection. Section 3 briefly reviews the semi-supervised method for orientation detection presented in (Esuli and Sebastiani, 2005). Section 4 describes in detail three different variants of it we propose for determining, at the same time, subjectivity *and* orientation, and describes the general setup of our experiments. In Section 5 we discuss the results we have obtained. Section 6 concludes.

2 Related work

2.1 Determining term orientation

Most previous works dealing with the properties of terms within an opinion mining perspective have focused on determining term orientation.

Hatzivassiloglou and McKeown (1997) attempt to predict the orientation of subjective *adjectives* by analysing pairs of adjectives (conjoined by `and`, `or`, `but`, `either-or`, or `neither-nor`) extracted from a large unlabelled document set. The underlying intuition is that the act of conjoining adjectives is subject to linguistic constraints on the orientation of the adjectives involved; e.g. `and` usually conjoins adjectives of equal orientation, while `but` conjoins adjectives of opposite orientation. The authors generate a graph where terms are nodes connected by "equal-orientation" or "opposite-orientation" edges, depending on the conjunctions extracted from the document set. A clustering algorithm then partitions the graph into a Positive cluster and a Negative cluster, based on a relation of similarity induced by the edges.

Turney and Littman (2003) determine term orientation by bootstrapping from two small sets of subjective "seed" terms (with the seed set for Positive containing terms such as `good` and `nice`, and the seed set for Negative containing terms such as `bad` and `nasty`). Their method is based on computing the *pointwise mutual information* (PMI) of the target term t with each seed term t_i as a measure of their semantic association. Given a target term t, its orientation value $O(t)$ (where positive value means positive orientation, and higher absolute value means stronger orientation) is given by the sum of the weights of its semantic association with the seed positive terms minus the sum of the weights of its semantic association with the seed negative terms. For computing PMI, term frequencies and co-occurrence frequencies are measured by querying a document set by means of the AltaVista search engine[1] with a "t" query, a "t_i" query, and a "t NEAR t_i" query, and using the number of matching documents returned by the search engine as estimates of the probabilities needed for the computation of PMI.

Kamps et al. (2004) consider instead the graph defined on adjectives by the WordNet[2] synonymy relation, and determine the orientation of a target

[1] `http://www.altavista.com/`

[2] `http://wordnet.princeton.edu/`

adjective t contained in the graph by comparing the lengths of (i) the shortest path between t and the seed term `good`, and (ii) the shortest path between t and the seed term `bad`: if the former is shorter than the latter, than t is deemed to be Positive, otherwise it is deemed to be Negative.

Takamura et al. (2005) determine term orientation (for Japanese) according to a "spin model", i.e. a physical model of a set of electrons each endowed with one between two possible spin directions, and where electrons propagate their spin direction to neighbouring electrons until the system reaches a stable configuration. The authors equate terms with electrons and term orientation to spin direction. They build a neighbourhood matrix connecting each pair of terms if one appears in the gloss of the other, and iteratively apply the spin model on the matrix until a "minimum energy" configuration is reached. The orientation assigned to a term then corresponds to the spin direction assigned to electrons.

The system of Kim and Hovy (2004) tackles orientation detection by attributing, to each term, a positivity score *and* a negativity score; interestingly, terms may thus be deemed to have both a positive and a negative correlation, maybe with different degrees, and some terms may be deemed to carry a stronger positive (or negative) orientation than others. Their system starts from a set of positive and negative seed terms, and expands the positive (resp. negative) seed set by adding to it the synonyms of positive (resp. negative) seed terms and the antonyms of negative (resp. positive) seed terms. The system classifies then a target term t into either Positive or Negative by means of two alternative learning-free methods based on the probabilities that synonyms of t also appear in the respective expanded seed sets. A problem with this method is that it can classify only terms that share some synonyms with the expanded seed sets. Kim and Hovy also report an evaluation of human inter-coder agreement. We compare this evaluation with our results in Section 5.

The approach we have proposed for determining term orientation (Esuli and Sebastiani, 2005) is described in more detail in Section 3, since it will be extensively used in this paper.

All these works evaluate the performance of the proposed algorithms by checking them against precompiled sets of Positive and Negative terms, i.e. checking how good the algorithms are at classifying a term known to be subjective into either Positive or Negative. When tested on the same benchmarks, the methods of (Esuli and Sebastiani, 2005; Turney and Littman, 2003) have performed with comparable accuracies (however, the method of (Esuli and Sebastiani, 2005) is much more efficient than the one of (Turney and Littman, 2003)), and have outperformed the method of (Hatzivassiloglou and McKeown, 1997) by a wide margin and the one by (Kamps et al., 2004) by a very wide margin. The methods described in (Hatzivassiloglou and McKeown, 1997) is also limited by the fact that it can only decide the orientation of *adjectives*, while the method of (Kamps et al., 2004) is further limited in that it can only work on adjectives that are present in WordNet. The methods of (Kim and Hovy, 2004; Takamura et al., 2005) are instead difficult to compare with the other ones since they were not evaluated on publicly available datasets.

2.2 Determining term subjectivity

Riloff et al. (2003) develop a method to determine whether a term has a Subjective or an Objective connotation, based on bootstrapping algorithms. The method identifies patterns for the extraction of subjective *nouns* from text, bootstrapping from a seed set of 20 terms that the authors judge to be strongly subjective and have found to have high frequency in the text collection from which the subjective nouns must be extracted. The results of this method are not easy to compare with the ones we present in this paper because of the different evaluation methodologies. While we adopt the evaluation methodology used in all of the papers reviewed so far (i.e. checking how good our system is at replicating an existing, independently motivated lexical resource), the authors do not test their method on an independently identified set of labelled terms, but on the set of terms that the algorithm itself extracts. This evaluation methodology only allows to test precision, and not accuracy *tout court*, since no quantification can be made of false negatives (i.e. the subjective terms that the algorithm should have spotted but has not spotted). In Section 5 this will prevent us from drawing comparisons between this method and our own.

Baroni and Vegnaduzzo (2004) apply the PMI method, first used by Turney and Littman (2003) to determine term orientation, to determine term subjectivity. Their method uses a small set S_s of 35 adjectives, marked as subjective by human judges, to assign a subjectivity score to each adjective to be classified. Therefore, their method, unlike our own, does not *classify* terms (i.e. take firm classification decisions), but *ranks* them according to a subjectivity score, on which they evaluate precision at various level of recall.

3 Determining term subjectivity and term orientation by semi-supervised learning

The method we use in this paper for determining term subjectivity and term orientation is a variant of the method proposed in (Esuli and Sebastiani, 2005) for determining term orientation alone.

This latter method relies on training, in a semi-supervised way, a binary classifier that labels terms as either Positive or Negative. A *semi-supervised* method is a learning process whereby only a small subset $L \subset Tr$ of the training data Tr are human-labelled. In origin the training data in $U = Tr - L$ are instead unlabelled; it is the process itself that labels them, automatically, by using L (with the possible addition of other publicly available resources) as input. The method of (Esuli and Sebastiani, 2005) starts from two small seed (i.e. training) sets L_p and L_n of known Positive and Negative terms, respectively, and expands them into the two final training sets $Tr_p \supset L_p$ and $Tr_n \supset L_n$ by adding them new sets of terms U_p and U_n found by navigating the WordNet graph along the synonymy and antonymy relations[3]. This process is based on the hypothesis that synonymy and antonymy, in addition to defining a relation of meaning, also define a relation of orientation, i.e. that two synonyms typically have the same orientation and two antonyms typically have opposite orientation. The method is iterative, generating two sets Tr_p^k and Tr_n^k at each iteration k, where $Tr_p^k \supset Tr_p^{k-1} \supset \ldots \supset Tr_p^1 = L_p$ and $Tr_n^k \supset Tr_n^{k-1} \supset \ldots \supset Tr_n^1 = L_n$. At iteration k, Tr_p^k is obtained by adding to Tr_p^{k-1} all synonyms of terms in Tr_p^{k-1} and all antonyms of terms in Tr_n^{k-1}; similarly, Tr_n^k is obtained by adding to Tr_n^{k-1} all synonyms of terms in Tr_n^{k-1} and all antonyms of terms in Tr_p^{k-1}. If a total of K iterations are performed, then $Tr = Tr_p^K \cup Tr_n^K$.

The second main feature of the method presented in (Esuli and Sebastiani, 2005) is that terms are given vectorial representations based on their WordNet *glosses* (i.e. textual definitions). For each term t_i in $Tr \cup Te$ (Te being the test set, i.e. the set of terms to be classified), a textual representation of t_i is generated by collating all the glosses of t_i as found in WordNet[4]. Each such representation is converted into vectorial form by standard text indexing techniques (in (Esuli and Sebastiani, 2005) and in the present work, stop words are removed and the remaining words are weighted by cosine-normalized $tfidf$; no stemming is performed)[5]. This representation method is based on the assumption that terms with a similar orientation tend to have "similar" glosses: for instance, that the glosses of `honest` and `intrepid` will both contain appreciative expressions, while the glosses of `disturbing` and `superfluous` will both contain derogative expressions. Note that this method allows to classify *any* term, independently of its POS, provided there is a gloss for it in the lexical resource.

Once the vectorial representations for all terms in $Tr \cup Te$ have been generated, those for the terms in Tr are fed to a supervised learner, which thus generates a binary classifier. This latter, once fed with the vectorial representations of the terms in Te, classifies each of them as either Positive or Negative.

4 Experiments

In this paper we extend the method of (Esuli and Sebastiani, 2005) to the determination of term subjectivity *and* term orientation altogether.

4.1 Test sets

The benchmark (i.e. test set) we use for our experiments is the General Inquirer (GI) lexicon (Stone et al., 1966). This is a lexicon of terms labelled according to a large set of categories[6], each one denoting the presence of a specific trait in the term. The two main categories, and the ones we will be concerned with, are Positive/Negative, which contain 1,915/2,291 terms having a positive/negative orientation (in what follows we will also refer to the category Subjective, which we define as the union of the two categories Positive and Negative). In opinion mining research the GI was first used by Turney and Littman (2003), who reduced the list of terms to 1,614/1,982 entries af-

[3] Several other WordNet lexical relations, and several combinations of them, are tested in (Esuli and Sebastiani, 2005). In the present paper we only use the best-performing such combination, as described in detail in Section 4.2. The version of WordNet used here and in (Esuli and Sebastiani, 2005) is 2.0.

[4] In general a term t_i may have more than one gloss, since it may have more than one sense; dictionaries normally associate one gloss to each sense.

[5] Several combinations of subparts of a WordNet gloss are tested as textual representations of terms in (Esuli and Sebastiani, 2005). Of all those combinations, in the present paper we always use the DGS$\neg$ combination, since this is the one that has been shown to perform best in (Esuli and Sebastiani, 2005). DGS$\neg$ corresponds to using the entire gloss and performing *negation propagation* on its text, i.e. replacing all the terms that occur after a negation in a sentence with negated versions of the term (see (Esuli and Sebastiani, 2005) for details).

[6] The definitions of all such categories are available at `http://www.webuse.umd.edu:9090/`

ter removing 17 terms appearing in both categories (e.g. deal) and reducing all the multiple entries of the same term in a category, caused by multiple senses, to a single entry. Likewise, we take all the 7,582 GI terms that are not labelled as either Positive or Negative, as being (implicitly) labelled as Objective, and reduce them to 5,009 terms after combining multiple entries of the same term, caused by multiple senses, to a single entry.

The effectiveness of our classifiers will thus be evaluated in terms of their ability to assign the total 8,605 GI terms to the correct category among Positive, Negative, and Objective[7].

4.2 Seed sets and training sets

Similarly to (Esuli and Sebastiani, 2005), our training set is obtained by expanding initial seed sets by means of WordNet lexical relations. The main difference is that our training set is now the union of *three* sets of training terms $Tr = Tr_p^K \cup Tr_n^K \cup Tr_o^K$ obtained by expanding, through K iterations, three seed sets Tr_p^1, Tr_n^1, Tr_o^1, one for each of the categories Positive, Negative, and Objective, respectively.

Concerning categories Positive and Negative, we have used the seed sets, expansion policy, and number of iterations, that have performed best in the experiments of (Esuli and Sebastiani, 2005), i.e. the seed sets $Tr_p^1 = \{\texttt{good}\}$ and $Tr_n^1 = \{\texttt{bad}\}$ expanded by using the union of synonymy and indirect antonymy, restricting the relations only to terms with the same POS of the original terms (i.e. adjectives), for a total of $K = 4$ iterations. The final expanded sets contain 6,053 Positive terms and 6,874 Negative terms.

Concerning the category Objective, the process we have followed is similar, but with a few key differences. These are motivated by the fact that the Objective category coincides with the complement of the union of Positive and Negative; therefore, Objective terms are more varied and diverse in meaning than the terms in the other two categories. To obtain a representative expanded set Tr_o^K, we have chosen the seed set $Tr_o^1 = \{\texttt{entity}\}$ and we have expanded it by using, along with synonymy and antonymy, the WordNet relation of hyponymy (e.g. vehicle / car), and without imposing the restriction that the two related terms must have the same POS. These choices are strictly related to each other: the term entity is the root term of the largest generalization hierarchy in WordNet, with more than 40,000 terms (Devitt and Vogel, 2004), thus allowing to reach a very large number of terms by using the hyponymy relation[8]. Moreover, it seems reasonable to assume that terms that refer to *entities* are likely to have an "objective" nature, and that hyponyms (and also synonyms and antonyms) of an objective term are also objective. Note that, at each iteration k, a given term t is added to Tr_o^k only if it does not already belong to either Tr_p or Tr_n. We experiment with two different choices for the Tr_o set, corresponding to the sets generated in $K = 3$ and $K = 4$ iterations, respectively; this yields sets Tr_o^3 and Tr_o^4 consisting of 8,353 and 33,870 training terms, respectively.

4.3 Learning approaches and evaluation measures

We experiment with three "philosophically" different learning approaches to the problem of distinguishing between Positive, Negative, and Objective terms.

Approach I is a two-stage method which consists in learning two binary classifiers: the first classifier places terms into either Subjective or Objective, while the second classifier places terms that have been classified as Subjective by the first classifier into either Positive or Negative. In the training phase, the terms in $Tr_p^K \cup Tr_n^K$ are used as training examples of category Subjective.

Approach II is again based on learning two binary classifiers. Here, one of them must discriminate between terms that belong to the Positive category and ones that belong to its complement (**not** Positive), while the other must discriminate between terms that belong to the Negative category and ones that belong to its complement (**not** Negative). Terms that have been classified *both* into Positive by the former classifier and into (**not** Negative) by the latter are deemed to be positive, and terms that have been classified *both* into (**not** Positive) by the former classifier and into Negative by the latter are deemed to be negative. The terms that have been classified (i) into both (**not** Positive) and (**not** Negative), or (ii) into both Positive and Negative, are taken to be Objective. In the training phase of Approach II, the terms in $Tr_n^K \cup Tr_o^K$ are used as training examples of category (**not** Positive), and the terms in $Tr_p^K \cup Tr_o^K$ are used as training examples of category (**not** Negative).

Approach III consists instead in viewing Positive, Negative, and Objective as three categories

[7] We make this labelled term set available for download at http://patty.isti.cnr.it/~esuli/software/SentiGI.tgz.

[8] The synonymy relation connects instead only 10,992 terms at most (Kamps et al., 2004).

with equal status, and in learning a ternary classifier that classifies each term into exactly one among the three categories.

There are several differences among these three approaches. A first difference, of a conceptual nature, is that only Approaches I and III view Objective as a category, or concept, in its own right, while Approach II views objectivity as a nonexistent entity, i.e. as the "absence of subjectivity" (in fact, in Approach II the training examples of Objective are only used as training examples of the *complements* of Positive and Negative). A second difference is that Approaches I and II are based on standard binary classification technology, while Approach III requires "multiclass" (i.e. 1-of-m) classification. As a consequence, while for the former we use well-known learners for binary classification (the naive Bayesian learner using the multinomial model (McCallum and Nigam, 1998), support vector machines using linear kernels (Joachims, 1998), the Rocchio learner, and its PrTFIDF probabilistic version (Joachims, 1997)), for Approach III we use their multiclass versions[9].

Before running our learners we make a pass of feature selection, with the intent of retaining only those features that are good at discriminating our categories, while discarding those which are not. Feature selection is implemented by scoring each feature f_k (i.e. each term that occurs in the glosses of at least one training term) by means of the *mutual information* (MI) function, defined as

$$MI(f_k) = \sum_{\substack{c\in\{c_1,\ldots,c_m\},\\ f\in\{f_k,\overline{f}_k\}}} \Pr(f,c)\cdot\log\frac{\Pr(f,c)}{\Pr(f)\Pr(c)} \quad (1)$$

and discarding the $x\%$ features f_k that minimize it. We will call $x\%$ the *reduction factor*. Note that the set $\{c_1,\ldots,c_m\}$ from Equation 1 is interpreted differently in Approaches I to III, and always consistently with who the categories at stake are.

Since the task we aim to solve is manifold, we will evaluate our classifiers according to two evaluation measures:

- *SO-accuracy*, i.e. the accuracy of a classifier in separating Subjective from Objective, i.e. in deciding term subjectivity alone;
- *PNO-accuracy*, the accuracy of a classifier in discriminating among Positive, Negative, and Objective, i.e. in deciding both term orientation and subjectivity.

[9]The naive Bayesian, Rocchio, and PrTFIDF learners we have used are from Andrew McCallum's *Bow* package (`http://www-2.cs.cmu.edu/~mccallum/bow/`), while the SVMs learner we have used is Thorsten Joachims' SVM^{light} (`http://svmlight.joachims.org/`), version 6.01. Both packages allow the respective learners to be run in "multiclass" fashion.

Table 1: Average and best accuracy values over the four dimensions analysed in the experiments.

Dimension	SO-accuracy		PNO-accuracy	
	Avg (σ)	Best	Avg (σ)	Best
Approach				
I	.635 (.020)	.668	.595 (.029)	.635
II	**.636** (.033)	**.676**	**.614** (.037)	**.660**
III	.635 (.036)	.674	.600 (.039)	.648
Learner				
NB	**.653** (.014)	.674	**.619** (.022)	.647
SVMs	.627 (.033)	.671	.601 (.037)	.658
Rocchio	.624 (.030)	.654	.585 (.033)	.616
PrTFIDF	.637 (.031)	**.676**	.606 (.042)	**.660**
TSR				
0%	.649 (.025)	**.676**	.619 (.027)	**.660**
50%	**.650** (.022)	.670	**.622** (.022)	.657
80%	.646 (.023)	.674	.621 (.021)	.647
90%	.642 (.024)	.667	.616 (.024)	.651
95%	.635 (.027)	.671	.606 (.031)	.658
99%	.612 (.036)	.661	.570 (.049)	.647
Tr_o^K *set*				
Tr_o^3	**.645** (.006)	**.676**	.608 (.007)	.658
Tr_o^4	.633 (.013)	.674	**.610** (.018)	**.660**

5 Results

We present results obtained from running every combination of (i) the three approaches to classification described in Section 4.3, (ii) the four learners mentioned in the same section, (iii) five different reduction factors for feature selection (0%, 50%, 90%, 95%, 99%), and (iv) the two different training sets (Tr_o^3 and Tr_o^4) for Objective mentioned in Section 4.2. We discuss each of these four dimensions of the problem individually, for each one reporting results averaged across all the experiments we have run (see Table 1).

The first and most important observation is that, with respect to a pure term orientation task, accuracy drops significantly. In fact, the best SO-accuracy and the best PNO-accuracy results obtained across the 120 different experiments are .676 and .660, respectively (these were obtained by using Approach II with the PrTFIDF learner and no feature selection, with $Tr_o = Tr_o^3$ for the .676 SO-accuracy result and $Tr_o = Tr_o^4$ for the .660 PNO-accuracy result); this contrasts sharply with the accuracy obtained in (Esuli and Sebastiani, 2005) on discriminating Positive from Negative (where the best run obtained .830 accuracy), *on the same benchmarks and essentially the same algorithms*. This suggests that good performance at orientation detection (as e.g. in (Esuli and Sebastiani, 2005; Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003)) may not be a

Table 2: Human inter-coder agreement values reported by Kim and Hovy (2004).

Agreement measure	Adjectives (462) Hum1 vs Hum2	Verbs (502) Hum2 vs Hum3
Strict	.762	.623
Lenient	.890	.851

guarantee of good performance at subjectivity detection, quite evidently a harder (and, as we have suggested, more realistic) task.

This hypothesis is confirmed by an experiment performed by Kim and Hovy (2004) on testing the agreement of two human coders at tagging words with the Positive, Negative, and Objective labels. The authors define two measures of such agreement: *strict* agreement, equivalent to our PNO-accuracy, and *lenient* agreement, which measures the accuracy at telling Negative against the rest. For any experiment, strict agreement values are then going to be, by definition, lower or equal than the corresponding lenient ones. The authors use two sets of 462 adjectives and 502 verbs, respectively, randomly extracted from the basic English word list of the TOEFL test. The inter-coder agreement results (see Table 2) show a deterioration in agreement (from lenient to strict) of 16.77% for adjectives and 36.42% for verbs. Following this, we evaluated our best experiment according to these measures, and obtained a "strict" accuracy value of .660 and a "lenient" accuracy value of .821, with a relative deterioration of 24.39%, in line with Kim and Hovy's observation[10]. This confirms that determining subjectivity and orientation is a much harder task than determining orientation alone.

The second important observation is that there is very little variance in the results: across all 120 experiments, average SO-accuracy and PNO-accuracy results were .635 (with standard deviation $\sigma = .030$) and .603 ($\sigma = .036$), a mere 6.06% and 8.64% deterioration from the best results reported above. This seems to indicate that the levels of performance obtained may be hard to improve upon, especially if working in a similar framework.

Let us analyse the individual dimensions of the problem. Concerning the three approaches to classification described in Section 4.3, Approach II outperforms the other two, but by an extremely narrow margin. As for the choice of learners, on average the best performer is NB, but again by a very small margin wrt the others. On average, the best reduction factor for feature selection turns out to be 50%, but the performance drop we witness in approaching 99% (a dramatic reduction factor) is extremely graceful. As for the choice of Tr_o^K, we note that Tr_o^3 and Tr_o^4 elicit comparable levels of performance, with the former performing best at SO-accuracy and the latter performing best at PNO-accuracy.

An interesting observation on the learners we have used is that NB, PrTFIDF and SVMs, unlike Rocchio, generate classifiers that depend on $P(c_i)$, the prior probabilities of the classes, which are normally estimated as the proportion of training documents that belong to c_i. In many classification applications this is reasonable, as we may assume that the training data are sampled from the same distribution from which the test data are sampled, and that these proportions are thus indicative of the proportions that we are going to encounter in the test data. However, in our application this is not the case, since we do not have a "natural" sample of training terms. What we have is one human-labelled training term for each category in {Positive,Negative,Objective}, and as many machine-labelled terms as we deem reasonable to include, in possibly different numbers for the different categories; and we have no indication whatsoever as to what the "natural" proportions among the three might be. This means that the proportions of Positive, Negative, and Objective terms we decide to include in the training set will strongly bias the classification results if the learner is one of NB, PrTFIDF and SVMs. We may notice this by looking at Table 3, which shows the average proportion of test terms classified as Objective by each learner, depending on whether we have chosen Tr_o to coincide with Tr_o^3 or Tr_o^4; note that the former (resp. latter) choice means having roughly as many (resp. roughly five times as many) Objective training terms as there are Positive and Negative ones. Table 3 shows that, the more Objective training terms there are, the more test terms NB, PrTFIDF and (in particular) SVMs will classify as Objective; this is not true for Rocchio, which is basically unaffected by the variation in size of Tr_o.

6 Conclusions

We have presented a method for determining *both* term subjectivity *and* term orientation for opinion mining applications. This is a valuable advance with respect to the state of the art, since past work in this area had mostly confined to determining term orientation alone, a task that (as we have ar-

[10]We observed this trend in all of our experiments.

Table 3: Average proportion of test terms classified as Objective, for each learner and for each choice of the Tr_o^K set.

Learner	Tr_o^3	Tr_o^4	Variation
NB	.564 ($\sigma = .069$)	.693 (.069)	+23.0%
SVMs	.601 (.108)	.814 (.083)	+35.4%
Rocchio	.572 (.043)	.544 (.061)	**-4.8%**
PrTFIDF	.636 (.059)	.763 (.085)	+20.0%

gued) has limited practical significance in itself, given the generalized absence of lexical resources that tag terms as being either Subjective or Objective. Our algorithms have tagged by orientation *and* subjectivity the entire General Inquirer lexicon, a complete general-purpose lexicon that is the *de facto* standard benchmark for researchers in this field. Our results thus constitute, for this task, the first baseline for other researchers to improve upon.

Unfortunately, our results have shown that an algorithm that had shown excellent, state-of-the-art performance in deciding term orientation (Esuli and Sebastiani, 2005), once modified for the purposes of deciding term subjectivity, performs more poorly. This has been shown by testing several variants of the basic algorithm, some of them involving radically different supervised learning policies. The results suggest that deciding term subjectivity is a substantially harder task that deciding term orientation alone.

References

M. Baroni and S. Vegnaduzzo. 2004. Identifying subjective adjectives through Web-based mutual information. In *Proceedings of KONVENS-04, 7th Konferenz zur Verarbeitung Natürlicher Sprache (German Conference on Natural Language Processing)*, pages 17–24, Vienna, AU.

Ann Devitt and Carl Vogel. 2004. The topology of WordNet: Some metrics. In *Proceedings of GWC-04, 2nd Global WordNet Conference*, pages 106–111, Brno, CZ.

Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss analysis. In *Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management*, pages 617–624, Bremen, DE.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*, pages 174–181, Madrid, ES.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING-00, 18th International Conference on Computational Linguistics*, pages 174–181, Saarbrücken, DE.

Thorsten Joachims. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 143–151, Nashville, US.

Thorsten Joachims. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Chemnitz, DE.

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

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of COLING-04, 20th International Conference on Computational Linguistics*, pages 1367–1373, Geneva, CH.

Andrew K. McCallum and Kamal Nigam. 1998. A comparison of event models for naive Bayes text classification. In *Proceedings of the AAAI Workshop on Learning for Text Categorization*, pages 41–48, Madison, US.

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

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of CONLL-03, 7th Conference on Natural Language Learning*, pages 25–32, Edmonton, CA.

P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, US.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting emotional polarity of words using spin model. In *Proceedings of ACL-05, 43rd Annual Meeting of the Association for Computational Linguistics*, pages 133–140, Ann Arbor, US.

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

Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, US.

Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of AAAI-04, 21st Conference of the American Association for Artificial Intelligence*, pages 761–769, San Jose, US.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing*, pages 129–136, Sapporo, JP.