Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone.

Farah Benamara
Institut de Recherche en Informatique de Toulouse, Univ. Paul Sabatier.
benamara@irit.fr

Carmine Cesarano, Antonio Picariello
Dipartimento di Informatica, Univ. di Napoli Federico II, Napoli, Italy
cacesara,picus@unina.it

Diego Reforgiato, VS Subrahmanian
Dept. of Computer Science and Institute for Advanced Computer Studies, University of Maryland
College Park, MD 20742
diegoref,vs@umiacs.umd.edu

Abstract

Most past work on determining the strength of subjective expressions within a sentence or a document use specific parts of speech such as adjectives, verbs and nouns. To date, there is almost no work on the use of adverbs in sentiment analysis, nor has there been any work on the use of adverb-adjective combinations (AACs). We propose an AAC-based sentiment analysis technique that uses a linguistic analysis of adverbs of degree. We define a set of general axioms (based on a classification of adverbs of degree into five categories) that all adverb scoring techniques must satisfy. Instead of aggregating scores of both adverbs and adjectives using simple scoring functions, we propose an axiomatic treatment of AACs based on the linguistic classification of adverbs. Three specific AAC scoring methods that satisfy the axioms are presented. We describe the results of experiments on an annotated set of 200 news articles (annotated by 10 students) and compare our algorithms with some existing sentiment analysis algorithms. We show that our results lead to higher accuracy based on Pearson correlation with human subjects.

Keywords

Sentiment analysis, adverbs of degree, Adverb-adjective combinations.

1. Introduction

There is growing interest in sentiment analysis. Companies are interested in what bloggers are saying about their products. Politicians are interested in how different news media are portraying them. Governments are interested in how foreign news media are representing their actions.

The current state of the art in sentiment analysis focuses on assigning a polarity or a strength to subjective expressions (words and phrases that express opinions, emotions, sentiments, etc.) in order to decide the objectivity/subjectivity orientation of a document [7][4] or the positive/negative/neutral polarity of an opinion sentence within a document [**?**][10][5]. Additional work has focused on the strength of an opinion expression where each clause within a sentence can have a neutral, low, medium or a high strength [6].

Though much work on determining term orientation has focused on nouns, verbs and adjectives, almost no work to date has focused on (i) the use of adverbs and (ii) the use of adverb-adjective combinations. However, the following simple example shows that adverbs do have an impact on the strength of a given sentiment.

- (S1) The concert was enjoyable.
- (S2) The concert was very enjoyable.
- (S3) The concert was thoroughly enjoyable.

All three sentences are positive - yet, most of us would agree that the sentiments expressed get progressively stronger as we go from (S1) to (S3).

The use of adverbs and adverbial phrases to improve the performance of sentiment analysis was shown in some recent studies. In [2], complex adjective phrases such as: "excessively affluent" or "more bureaucratic" are used to extract opinion propositions. Given a set of manually annotated adjectives, the score of an adverb depends on how often it co-occurs in the same sentence with the seed words in this set [5]. The overall score of a sentence is then obtained by aggregating the scores (mainly based on a score sum feature) assigned to both adverbs and adjectives. [3] uses a template based methods to map expressions of degree such as "sometimes", "very", "not too", "extremely very" to a [-2, 10] scale. This approach does not take adjective scoring into account.

In this paper, we propose a linguistic approach to the problem of sentiment analysis. Our goal is to assign a number from -1 to +1 to denote the strength of sentiment on a given topic t in a sentence or document based on the score assigned to the applicable adverb-adjective combinations found in sentences. A score of -1 reflects a maximally negative opinion about the topic, while a score of +1 reflects a maximally positive opinion about the topic. Scores in between reflect relatively more positive (resp. more negative) opinions depending on how close they are to +1 (resp. -1).

The primary contributions of this paper are the following:

1. We study the intensity of adverbs of degree at the linguistic level in order to define general axioms to score *adverbs of degree* on a 0 to 1 scale. These axioms use linguistic classifications of adverbs of degree in order to lay out axioms governing what the score of a given adverb should be, relative to the linguistic classification. These axioms are satisfied by a number of specific scoring functions, some of which are described in the paper. The axioms as well as the scoring method is described in Section 2

ICWSM '2006 Boulder, CO USA

2. We propose the novel concept of an adverb-adjective combinations (AACs for short). Intuitively, an AAC (e.g. "very bad") consists of an adjective (e.g. "bad") modified by at least one adverb (e.g. "very"). Using the linguistic classification of adverbs of degree, we provide an *axiomatic treatment* of how to score the strength of sentiment expressed by an AAC. These AAC scoring methods can be built on top of any existing method to score adjective intensity [8][10]. The AAC scoring axioms are described in section 3.

3. We then develop three AAC scoring methods that satisfy the AAC scoring axioms. The first, called ***Variable scoring*** allows us to modify adjective scores in different ways, based on the score of the adjective. The second method, called ***Adjective priority scoring (APS)*** allows us to score an AAC by modifying the adjective score by assigning a fixed weight to the relevance of adverbs. The third, called ***Adverb First Scoring (AdvFS)*** allows us to score an AAC by modifying the score of an adverb by assigning a relevance to each adjective. Both ***APS*** and ***AdvS*** are parametrized by a number, r, between 0 and 1 that captures the relative weight of the adverb score relative to the adjective score. Part of the goal of this paper is to determine which weight most closely matches human assignments of opinions. The AAC scoring algorithms are presented in section 4.

4. Finally, we describe a set of experiments we conducted on an annotated set of about 200 documents selected randomly from a set of popular news sources. The annotations were done by 10 students. The experiments show that of the algorithms presented in this paper, the version of ***APS*** that uses $r = 0.35$) produces the best results. This means that in order to best match human subjects, the score an AAC such as "very bad"should consist of the score of the adjective ("bad") plus 35% of the score of the adverb ("very").

5. Moreover, we compare our algorithms with three existing sentiment analysis algorithms in the literature [8, 10, 4]. Our results show that using adverbs and AACs produces significantly higher Pearson correlations (of opinion analysis algorithms vs. human subjects) than these previously developed algorithms that did not use adverbs or AACs. $\boldsymbol{APS}^{0.35}$ produces a Pearson correlation of over 0.47. In contrast, our group of human annotators only had a correlation of 0.56 between them, showing that our $\boldsymbol{APS}^{0.35}$'s agreement with human annotators is quite close to agreement between pairs of human annotators. Those expeiiments (item 4 and 5) are detailed in section 6.

2. Adverb Scoring Axioms

Syntactically, adverbs may appear in different positions in a sentence. For example, they could occur as complements or modifiers of verbs (*he behaved badly*), modifiers of nouns (*only adults*), modifiers of adjectives (*a very dangerous trip*), modifiers of adverbs (*very nicely*) and clauses (*Undoubtedly, he was right*).

Semantically, adverbs are often subclassified with respect to distinct conceptual notions [11][13].

- *Adverbs of time* (e.g. yesterday, soon) tell us when an event occurred.
- *Adverbs of frequency* (e.g. never, rarely, daily) tell us how frequently an event occurs.
- *Adverbs of location* (e.g. abroad, outside) tell us where an event occurs.
- *Adverbs of manner* (e.g. slowly, carefully) tell us how something happens.
- *Adverbs of degree* (e.g. extremely, absolutely, hardly, precisely, really) tell us about the intensity with which something happens.
- *Conjunctive adverbs* (e.g. consequently, therefore) link two sentences.

In this paper, we only focus on adverbs of degree as we feel that this category of adverbs is the most relevant for sentiment analysis. We note that it is possible for adverbs that belong to other categories to have an impact on sentiment intensity (e.g. *it is never good*) - we defer a study of these other adverbs them to future work.

In this section, we outline how to provide scores between 0 and 1 to adverbs of degree that modify adjectives. A score of 1 implies that the adverb completely affirms an adjective, while a score of 0 implies that the adverb has no impact on an adjective. Adverbs of degree are classified as follows [12][14]:

1. Adverbs of affirmation: these include adverbs such as absolutely, certainly, exactly, totally, and so on.
2. Adverbs of doubt: these include adverbs such as possibly, roughly, apparently, seemingly, and so on.
3. Strong intensifying adverbs: these include adverbs such as astronomically, exceedingly, extremely, immensely, and so on.
4. Weak intensifying adverbs: these include adverbs such as barely, scarcely, weakly, slightly, and so on.
5. Negation and Minimizers: these include adverbs such as "hardly" — we treat these somewhat differently than the preceding four categories as they usually negate sentiments. We discuss these in detail in the next section.

In this section, we present a formal axiomatic model for scoring *adverbs of degree* that belong to one of the categories described above. We use two axioms when assigning scores to adverbs in these categories (except for the last category), as shown in figure 1.

1. (A1) Each weakly intensifying adverb and each adverb of doubt has a score less than or equal to each strongly intensifying adverb.
2. (A2) Each weakly intensifying adverb and each adverb of doubt has a score less than or equal to each adverb of affirmation.

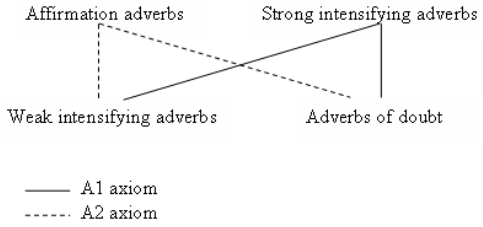

Fig. 1: *General Axioms to Score Adverbs of Degree*

Axiom (A1) is a reasonable axiom because a sentence such as *The concert will be slightly enjoyable* expresses a less strong opinion than a sentence such as *The concert will be highly enjoyable*. Axiom (A2) is a reasonable axiom because the sentence *The concert will be*

slightly enjoyable expresses a weaker sentiment than *The concert will be perfectly enjoyable*.

One may wonder whether other axioms should be added. One conundrum we faced was whether each adverb of doubt (resp. strong intensifier adverbs) gets a lower score than each weakly intensifying adverb (resp. affirmation adverbs)? The answer is unclear. For instance, *The concert will probably be enjoyable* has some doubt, but overall, it seems to assign a reasonable probability that the concert will be enjoyable. In contrast, there is no doubt in the sentence *The concert will be mildly enjoyable*, but the level of enjoyment seems low. Whether one should get higher scores than the other is debatable - hence, we decided not to require that each adverb of doubt (resp. strong intensifier adverb) get a lower or equal score than each weakly intensifying adverb (resp. affirmation adverb). We examined all possible pairs of categories to see if such axioms could be added and excluded other pairs for similar reasons.

Minimizers. There are a small number of adverbs called *minimizers* such as "hardly" that actually have a negative effect on sentiment. For example, in the sentence *The concert was hardly good*, the adverb "hardly" is a minimizer that reduces the positive score of the sentence *The concert was good.* We actually assign a negative score to minimizers. The reason is that minimizers tend to negate the score of the adjective to which they are applied. For example, the *hardly* in *hardly good* reduces the score of *good* because good is a "positive" adjective. In contrast, the use of the adverb *hardly* in the AAC *hardly bad* increases the score of *bad* because *bad* is a negative adjective.

Based on these principles, we asked a group of 10 individuals to provide scores to approximately 100 adverbs of degree - we used the average to obtain a score $sc(adv)$ for each adverb adv within each category we have defined. Some example scores we got in this way are: $sc(certainly) = 0.84, sc(possibly) = 0.22,$
$sc(exceedingly) = 0.9, sc(barely) = 0.11.$

3. Adverb Adjective Combination Scoring Axioms

In addition to the adverb scores ranging from 0 to 1 mentioned above, we assume that we have a score assigned on a -1 to +1 scale for each adjective.

There is a reason for this dichotomy of scales (0 to 1 for adverbs, -1 to +1 for adjectives). With the exception of minimizers (which are relatively few in number), all adverbs strengthen the polarity of an adjective - the difference is to the extent. The 0 to 1 score for adverbs reflects a measure of this strengthening.

In constrast, adjectives were assigned scores from -1 to +1 in [8] because they can be positive or negative. Several papers have already scored adjectives. [16] determines term orientation by bootstrapping from a set of positive terms and a set of negative terms. Their method is based on computing the pointwise mutual information (PMI) of the target term with each seed term t as a measure of their semantic association. [15] and [10] use the WordNet synonymy relation between adjectives in order to expand seed sets of opinion words using machine learning based approaches. They assign scores in the interval $[-1, +1]$ to adjectives. [8] develops scores between -1 and +1 for adjectives by using a statistical model. Our framework can work with any of these scoring methods, as long as the scores are normalized between -1 and $+1$. In our implementation, we use the scores provided by [8].

Let $sc(adj)$ denote the score of any such adjective. A score of -1 means that the adjective is maximally negative, while a score of +1 means that the adjective is maximally positive. An adjective is *positive* (resp. *negative*) if its score is greater than 0 (resp. less than 0).

An unary adverb adjective combination (AAC) has the form:

$$\langle adverb \rangle \langle adjective \rangle$$

while a binary AAC has the form

$$\langle adverb_i, adverb_j \rangle \langle adjective \rangle.$$

where: $adverb_i$ can be an adverb of doubt or a strong intensifying adverb whereas $adverb_j$ can be a strong or a weak intensifying adverbs. Binary AAC are thus restricted to 4 combinations only, such as: *very very good*, *possibly less expensive*, etc. The other combinations are not often used.

Our corpus contains no cases where three or more adverbs apply to an adjective — we believe this is very rare. The reader will observe that we rarely see phrases such as *Bush's policies were really, really, very awful*, though they can occur. An interesting note is that such phrases tend to occur more in blogs and almost never in news articles.

3.1 Unary AACs

Let AFF, $DOUBT$, $WEAK$, $STRONG$ and MIN respectively be the sets of adverbs of affirmation, adverbs of doubt, adverbs of weak intensity, adverbs of strong intensity and minimizers. Suppose f is any unary AAC scoring function that takes as input, one adverb and one adjective, and returns a number between -1 and +1. We will later show how to extend this to binary AACs. According to the category an adverb belong to, f should satisfy various axioms defined below.

1. Affirmative and strongly intensifying adverbs.
   - AAC-1. If $sc(adj) > 0$ and $adv \in AFF \cup STRONG$, then $f(adv, adj) \geq sc(adj)$.
   - AAC-2. If $sc(adj) < 0$ and $adv \in AFF \cup STRONG$, then $f(adv, adj) \leq sc(adj)$.
2. Weakly intensifying adverbs.
   - AAC-3. If $sc(adj) > 0$ and $adv \in WEAK$, then $f(adv, adj) \leq sc(adj)$.
   - AAC-4. If $sc(adj) < 0$ and $adv \in WEAK$, then $f(adv, adj) \geq sc(adj)$.
3. Adverbs of doubt.
   - AAC-5. If $sc(adj) > 0$, $adv \in DOUBT$, and $adv' \in AFF \cup STRONG$, then $f(adv, adj) \leq f(adv', adj)$.
   - AAC-6. If $sc(adj) < 0$ is negative, $adv \in DOUBT$, and $adv' \in AFF \cup STRONG$, then $f(adv, adj) \geq f(adv', adj)$.
4. Minimizers.
   - AAC-7. If $sc(adj) > 0$ and $adv \in MIN$, then $f(adv, adj) \leq sc(adj)$.
   - AAC-8. If $sc(adj) < 0$ and $adv \in MIN$, then $f(adv, adj) \geq sc(adj)$.

The intuition behind AAC-1 and AAC-2 is as follows. Adjectives are either positive (e.g. *good, wonderful*) or negative (e.g. *bad, horrible*). Adverbs that are either affirmative or strong intensifiers strengthen the positivity of positive adjectives (expressed in AAC-1) and the negativity of negative adjectives (expressed in AAC-2). Thus, *very* strengthens the intensity of *good*, causing the score of *very good* to be higher than that of *good*. However, *very* also strengthens the intensity of *bad*, cuasing the score of *very bad* to be lower than that of *bad*. This is what axioms AAC-1 and AAC-2 do.

Axiom AAC-3 looks at weak intensifiers (e.g. *weakly, barely*). Axiom AAC-3 says that a positive adjective should end up with a lower intensity when used with a weak intensifier adverb. For example, *The concert was barely good* should have a lower score than *The concert was good.* Axiom AAC-4 says that a negative adjective has a higher intensity when used with a weak intensifier adverb. *The concert was slightly bad* expresses a more positive view than *The concert was bad.*

AAC-5 and AAC-6 can be explained in a manner similar to the explanation for Axioms (A1),(A2) earlier in the paper. Finally, AAC-7 and AAC-8 say that minimizers reverse the polarity of an adjective.

3.2 Binary AACs

Suppose we have an AAC consisting of the form

$$\langle adv_1 \cdot adv_2 \rangle \langle adj \rangle.$$

In this case, we assign a score as follows.

- We first compute the score $f(adv_2, adj)$. This gives us a score s_2 denoting the intensity of the unary AAC $adv_2 \cdot adj$ which we denote AAC_1.
- We then apply f to (adv_1, AAC_1) and return that value as the answer.

Here's an example of how this works.

EXAMPLE 1. *For example, suppose we have*

$sc(\textit{really}) = 0.7$;
$sc(\textit{very}) = 0.6$;
$sc(\textit{wonderful}) = 0.8$.

To compute the score of really very wonderful, we first compute $f(\textit{very,wonderful})$. *This gives us some score - say 0.85. We set* AAC_1 *to be the AAC corresponding to the string very wonderful and set* $sc(\textit{very wonderful})$ *to be the above* f *value, i.e. 0.85. We then compute* $f(\textit{really}, AAC_1)$ *which might, for example, be 0.87. This is returned as the answer.*

3.3 Negation

Our treatment thus far does not handle negated AACs such as *The concert was not really bad*. In this case, we simply find the score for the AAC *really bad* and negate it. Thus, if the score of *really bad* was -0.6, then the score of the negated AAC, *not really bad* is +0.6. On the other hand, if the score of the sentence *really good* is 0.6, then the score of *not really good* will be -0.6.

4. Three AAC Scoring Algorithms

In this section, we propose three alternative algorithms (i.e. different f's) to assign a score to a unary AAC. Each of these three methods will be shown to satisfy our axioms. All three algorithms can be extended to apply to binary AACs and negated AACs using the methods shown above.

4.1 Variable Scoring

Suppose adj is an adjective and adv is an adverb. The variable scoring method (VS) works as follows.

- If $adv \in AFF \cup STRONG$, then:

  $$f_{\mathsf{VS}}(adv, adj) = sc(adj) + (1 - sc(adj)) \times sc(adv)$$

  if $sc(adj) > 0$. If $sc(adj) < 0$,

  $$f_{\mathsf{VS}}(adv, adj) = sc(adj) - (1 - sc(adj)) \times sc(adv).$$

- If $adv \in WEAK \cup DOUBT$, VS reverses the above and returns

  $$f_{\mathsf{VS}}(adv, adj) = sc(adj) - (1 - sc(adj)) \times sc(adv)$$

  if $sc(adj) > 0$. If $sc(adj) < 0$, it returns

  $$f_{\mathsf{VS}}(adv, adj) = sc(adj) + (1 - sc(adj)) \times sc(adv).$$

EXAMPLE 2. *Suppose we use the scores shown in Example 1 and suppose our sentence is The concert was really wonderful.* f_{VS} *would look at the ACC really wonderful and assign it the score :* $f_{\mathsf{VS}}(really,\ wonderful) = 0.8 + (1 - 0.8) \times 0.7 = 0.94$.

However, for the AAC very wonderful it would assign a score of : $f_{\mathsf{VS}}(very,\ wonderful) = 0.8 + (1 - 0.8) \times 0.6 = 0.92$ *which is a slightly lower rating because the score of the adverb really is smaller than the score of very.*

4.2 Adjective Priority Scoring

In variable scoring, the weight with which an adverb is considered depends upon the score of the adjective that it is associated with.

In contrast, in Adjective Priority Scoring (APS), we select a weight $r \in [0, 1]$. This weight denotes the importance of an adverb in comparison to an adjective that it modifies. r can vary based on different criteria. For example, if we are looking at highly reputable news media such as the BBC that have careful guidelines on what words to use in news reports [1], then r would depend on those guidelines. On the other hand, if we are looking on blogs or news media that are not subject to such strong guidelines, then experimentation is the best way to set r. Some preliminary studies, such as in [1], classify moods of blog text using a large collection of blog posts containing the authors indication of their state of mind at the time of writing: whether the author was depressed, cheerful, bored, and so on. It will be then interesting to compare the value of r depending on the nature of the opinion texts (blog or news).

The largest r is, the greater the impact. APS^r method works as follow:

- If $adv \in AFF \cup STRONG$, then

  $$f_{\mathsf{APS}^r}(adv, adj) = \mathsf{min}(1, sc(adj) + r \times sc(adv)).$$

  if $sc(adj) > 0$. If $sc(adj) > 0$,

  $$f_{\mathsf{APS}^r}(adv, adj) = \mathsf{min}(1, sc(adj) - r \times sc(adv)).$$

- If $adv \in WEAK \cup DOUBT$, then APS^r reverses the above and sets

  $$f_{\mathsf{APS}^r}(adv, adj) = \mathsf{max}(0, sc(adj) - r \times sc(adv)).$$

  if $sc(adj) > 0$. If $sc(adj) < 0$, then

  $$f_{\mathsf{APS}^r}(adv, adj) = \mathsf{max}(0, sc(adj) + r \times sc(adv)).$$

EXAMPLE 3. *Suppose we use the scores shown in Example 1 and suppose our sentence is The concert was really wonderful. Let* $r = 0.1$. *In this case,* $f_{\mathsf{APS}^{0.1}}$ *would look at the ACC really wonderful and assign it the score :*
$f_{\mathsf{APS}^{0.1}}(really,\ wonderful) = 0.8 + 0.1 \times 0.7 = 0.87$.
However, for the ACC very wonderful it would assign a score of:
$f_{\mathsf{APS}^{0.1}}(very,\ wonderful) = 0.8 + 0.1 \times 0.6 = 0.86$.
Again, as in the case of f_{VS}*, the score given to very wonderful is lower than the score given to really wonderful.*

[1] www.bbc.co.uk/guidelines/editorialguidelines/

4.3 Adverb First Scoring

In this section, we take the complementary view that instead of weighting the adverb, we should modify the adverb score by weighting the adjective score using an r (as before) that measures the weight of an adjective's importance in an AAC, relative to the importance of the adverb - this is why this method is called Adverb First Scoring.

Our $AdvFS^r$ algorithm works as follow:

- If $adv \in AFF \cup STRONG$, then

  $$f_{\mathsf{AdvFS}^r}(adv, adj) = \mathsf{min}(1, sc(adv) + r \times sc(adj)).$$

  if $sc(adj) > 0$. If $sc(adj) < 0$,

  $$f_{\mathsf{AdvFS}^r}(adv, adj) = \mathsf{max}(0, sc(adv) - r \times sc(adj)).$$

- If $adv \in WEAK \cup DOUBT$, then we reverse the above and set

  $$f_{\mathsf{AdvFS}^r}(adv, adj) = \mathsf{max}(0, sc(adv) - r \times sc(adj))$$

  if $sc(adj) > 0$. If $sc(adj) < 0$, then

  $$f_{\mathsf{AdvFS}^r}(adv, adj) = \mathsf{min}(1, sc(adv) + r \times sc(adj)).$$

EXAMPLE 4. *Let us return to the sentence The concert was really wonderful with $r = 0.1$. In this case, $f_{\mathsf{AdvFS}^{0.1}}$ would look assign the ACC really wonderful the score :*
$f_{\mathsf{AdvFS}^{0.1}}(really,\ wonderful) = 0.7 + 0.1 \times 0.8 = 0.78.$
However, for the ACC very wonderful it would assign a score of :
$f_{\mathsf{AdvFS}^{0.1}}(very,\ wonderful) = 0.6 + 0.1 \times 0.8 = 0.68.$
Again, as in the case of f_{VS} and $f_{\mathsf{AdvFS}^{0.1}}$, the score given to very wonderful is lower than the score given to really wonderful.

5. Scoring the Strength of Sentiment on a Topic

Our algorithm for scoring the strength of sentiment on a topic t in a document d is now the following.

1. Let $Rel(t)$ be the set of all sentences in d that directly or indirectly reference the topic t.
2. For each sentence s in $Rel(t)$, let $Appl^+(s)$ (resp. $Appl^-(s)$) be the multiset of all AACs occurring in s that are positively (resp. negatively) applicable to topic t.
3. Return $strength(t, s) =$

$$\frac{\Sigma_{s \in Rel(t)} \Sigma_{a \in Appl^+(s)} score(a) - \Sigma_{s \in Rel(t)} \Sigma_{a' \in Appl^-(s)} score(a')}{card(Rel(t))}.$$

The first step uses well known algorithms [5] to identify sentences that directly or indirectly reference a topic, while the second step finds the AACs applicable to a given topic by parsing it in a straightforward manner. The third step is key: it says that we classify the applicable AACs into positive and negative ones. We sum the scores of all applicable positive AACs and subtract from it, the sum of scores of all applicable negative AACs. We then divide this by the number of sentences in the document to obtain an average strength of sentiment measure. Let us see how the above method works on a tiny example.

EXAMPLE 5. *Suppose we have a concert review that contains just two sentences in $Rel(t)$. . . . The concert was really wonderful. . . . It [the concert] was absolutely marvelous. . . . According to Example 1, the first sentence yields a score of 0.87. Similarly, suppose the second sentence yields a score of 0.95. In this case, our algorithm would yield a score of 0.91 as the average.*

On the other hand, suppose the review looked like this: . . . The concert was not bad. It was really wonderful in parts.. . .. In this case, suppose the score, $sc(bad)$ of the adjective bad is -0.5. In this case, the negated AAC not bad gets a score of $+0.5$ in step (3) of the scoring algorithm. This, combined with the score of 0.87 for really wonderful would cause the algorithm to return a score of 0.685. In a sense, the not bad reduced the strength score as it is much weaker in strength than really wonderful.

6. Implementation and Experimentation

We have implemented all the algorithms mentioned in this paper (VS, APS^r, $ADVFS^r$) on top of the OASYS system[8], We also implemented the algorithms described in [10, 4]. And of course, as our algorithms are built on top of OASYS[8], we can compare our algorithms with [8] as well.

The algorithms were implemented in approximately 4200 lines of Java on a Pentium III 730MHz machine with 2GB RAM PC running Red Hat Enterprise Linux release 3. We ran experiments using a suite of 200 documents. The *training set used in OASYS was different from the experimental suite of 200 documents.*

We manually identified 3 topics in each document in the experimental dataset, and asked about 10 students (not affiliated with this paper) to rank the strength of sentiment on each of the three topics associated with each document.

We then conducted two sets of experiments.

6.1 Experiment 1 (Comparing correlations of algorithms in this paper).

The first experiment compared just the algorithms described in this paper in order to determine which one exhibits the best performance. More specifically, we were interested in finding out the value of r that makes APS^r and $advFS^r$ provide the best performance. The performance of an algorithm is based on the use of Pearson correlation coefficients between the opinion scores returned by the algorithm and the opinion scores provided by the same of human subjects.

The goal of our first experiment was to determine how well the algorithms APS^r, AdvFS^r did as we varied r. The graphs shown in Figure 2 below show how the Pearson correlation coefficient of our algorithms varied as we varied r for each of the two algorithms.

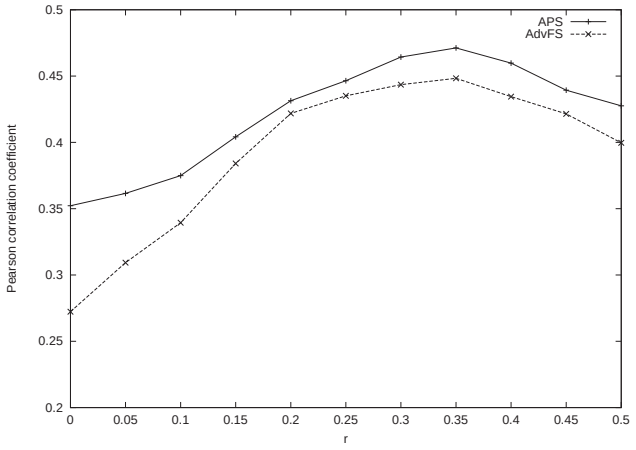

Fig. 2: *Pearson correlation coefficient for APS^r and AdvFS^r*

6.2 Experiment 2 (Correlation with human subjects).

The second experiment compared the algorithms from this paper with the algorithms described in [8, 10, 4]. Again, what we were interested in was the Pearson correlation of the algorithms described in this paper (showing the correlation between our algorithms and human subjects) with the corresponding Pearson correlations for the algorithms in [8, 10, 4].

Our algorithms apply to finding strength of sentiment in an entire document, not just in a single sentence. The table below shows the Pearson correlations of the algorithms in this paper (with $r = 0.35$) compared to the algorithms of [8, 4, 10].

Algorithm	Pearson correlation
Turney	0.132105644
Hovy	0.194580548
VS	0.342173328
AdvFS$^{0.35}$	0.448322524
APS$^{0.35}$	0.471219646

6.3 Results

It is easy to see that $\mathsf{APS}^{0.35}$ has the highest Pearson correlation coefficient when compared to human subjects. It seems to imply two things:

1. First, that adjectives are more important than adverbs in terms of how a human being views sentiment - this is because Adjective Priority Scoring (APS) beats Adverb First Scoring.

2. Second, the results seem to imply that when identifying the strength of opinion expressed about a topic, the "weight" given to adverb scores should be about 35% of the weight given to adjective scores. The fact that previous methods to measure sentiment strength did not take adverbs and AACs into account seems to account for the improved correlations of $\mathsf{APS}^{0.35}$. Moreover, past work did not make this observation about the relative degrees of importance of adverbs vs. adjectives in sentiment intensity scoring.

Inter-human correlations. Note that we also compared the correlations between the human subjects. This correlation turned out to be 0.56. As a consequence, on a relative scale, APS$^{0.35}$ seems to perform almost as well as humans.

7. Discussions and Conclusion

In this paper, we study the use of AACs in sentiment analysis based on a linguistic analysis of adverbs of degree. We differ from past work in three ways.

1. In [2][5], adverb scores depend on their collocation frequency with an adjective within a sentence, whereas in [3], scores are assigned manually by only one English speaker. These works do not distinguish between adverbs that belong to different conceptual notions, such as : "sometimes", "therefore", "daily" or "very". We propose a methodology for scoring adverbs by defining a set of general axioms based on a classification of adverbs of degree into five categories. Following those axioms, our scoring was performed by 10 people.

2. Instead of aggregating the scores of both adverbs and adjectives using simple scoring functions, we propose an axiomatic treatment of AACs based on the linguistic categories of adverbs we have defined. This is totally independent from any existing adjective scoring. Moreover, it is conceivable that there are other ways of scoring AACs (other than those proposed here) that would satisfy the axioms and do better - this is a topic for future exploration.

3. Based on the AAC scoring axioms, we developed three specific adverb-adjective scoring methods, namely, ***Variable scoring***, ***Adjective priority scoring (APS)*** and ***Adverb First Scoring (AdvFS)***. Our experiments show that the second method is the best with a weight of 0.35. We compared our methods with 3 existing algorithms that do not use any adverb scoring and our results show that using adverbs and AACs produces significantly higher precision and recall than these previously developed algorithms.

Our first experiments are very encouraging and open the door to several future directions. These include:

1. We plan to extend our set of adverb scoring axioms in order to handle other categories of adverbs, such as adverbs of time or adverbs of frequency.

2. We also plan to study other syntactic constructions, such as: adverb verb combinations (like in: *He strongly affirmed that ....*) as well as their use for scoring the overall opinion expression.

3. We plan to study the impact of style guidelines (such as news guidelines) on the evaluation process of the strength of opinion expressions.

References

[1] G. Mishne, Experiments with Mood Classification in Blog Posts, Proc. Style2005 - the 1st Workshop on Stylistic Analysis Of Text For Information Access, at SIGIR 2005, 2005.

[2] S. Bethard and H. Yu and A. Thornton and V. Hativassiloglou and D. Jurafsky, Automatic Extraction of Opinion Propositions and their Holders, Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text, 2004.

[3] T. Chklovski, Deriving Quantitative Overviews of Free Text Assessments on the Web, In Proceedings of 2006 International Conference on Intelligent User Interfaces (IUI06), January 29-Feb 1, 2006, Sydney, Australia, 2006.

[4] P. Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, In Proceedings of 2006 International Conference on Intelligent User Interfaces (IUI06), 2002.

[5] H. Yu and V. Hatzivassiloglou, Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, In Proceedings of EMNLP-03, 2003.

[6] T. Wilson and J. Wiebe and R. Hwa, Just how mad are you? Finding strong and weak opinion clauses, AAAI-04, 2004.

[7] B. Pang and L. Lee and S. Vaithyanathan, Thumbs up? Sentiment Classification Using Machine Learning Techniques, 2002.

[8] C. Cesarano and B. Dorr and A. Picariello and D. Reforgiato and A. Sagoff and V.S. Subrahmanian, OASYS: An Opinion Analysis System, AAAI 06 spring symposium on Computational Approaches to Analyzing Weblogs, 2004.

[9] V. Hatzivassiloglou and K. McKeown, Predicting the Semantic Orientation of Adjectives, ACL-97, 1997.

[10] S.O Kim and E. Hovy, Determining the Sentiment of Opinions, Coling04, 2004.

[11] A. Lobeck, Discovering Grammar. An Introduction to English Sentence Structure, New York/Oxford: Oxford University Press, 2000.

[12] R. Quirk and S. Greenbaum and G. Leech and J. Svartvik, A Comprehensive Grammar of the English Language, London: Longman, 1985.

[13] Shuan-Fan Huang , A Study of Adverbs, The Hague: Mouton, 1975 .

[14] D. Bolinger, Degree Words, The Hague: Mouton, 1972.

[15] J. Kamps and M. Marx and R.J. Mokken and M. De Rijke, Using WordNet to measure semantic orientation of adjectives, In Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, volume IV, 2004, pages 11151118, Lisbon, Portugal.

[16] P.D Turney and M.L. Littman, Measuring praise and criticism: Inference of semantic orientation from association, ACM Transactions on Information Systems, 2003, Vol. 21(4), pages 315346.