Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone

Farah Benamara
Univ. Paul Sabatier (IRIT), France
benamara@irit.fr

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

Diego Reforgiato, VS Subrahmanian
University of Maryland
College Park, MD 20742
diegoref,vs@umiacs.umd.edu

Abstract

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

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 orientation of a document [6][3] or the positive/negative/neutral polarity of an opinion sentence within a document [8][9][4]. 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 [5]. Adverbs were used for opinion mining in [1] where adjective phrases such as "excessively affluent" were used to extract opinion carrying sentences. [4] uses sum based scoring with manually scored adjectives and adverbs, while [2] uses a template based methods to map expressions of degree such as "sometimes", "very", "not too", "extremely very" to a [-2, 10] scale. However, almost no work to date has focused on (i) the use of adverbs and (ii) the use of adverb-adjective combinations.

We propose a linguistic approach to sentiment analysis where we assign a number from -1 (maximally negative opinion) to +1 (maximally positive opinion) 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. 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. Section 2 shows how we use linguistic classifications of adverbs of degree (AoD), we define general axioms to score AoDs on a 0 to 1 scale. T These axioms are satisfied by a number of specific scoring functions, some of which are described in the paper.

2. Section 3 proposes the novel concept of an adverb-adjective combination (AAC). Intuitively, an AAC (e.g. "very bad") consists of an adjective (e.g. "bad") modified by at least one adverb (e.g. "very"). 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 [7][9].

3. Section 4 presents the ***Variable scoring***, ***Adjective priority scoring (APS)***, and ***Adverb First Scoring (AdvFS)*** algorithms – all these methods satisfy the AAC scoring axioms. T

4. Section 6 describes experiments we conducted with an annotated corpus of 200 news articles (10 annotators) and 400 blog posts (5 annotators). 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"). Moreover, we compare our algorithms with three existing sentiment analysis algorithms [7, 9, 3]. 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.

2. Adverb scoring axioms

In this paper, we only focus on *adverbs of degree* [10] such as extremely, absolutely, hardly, precisely, really - such adverbs tell us about the intensity with which something happens. 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. We now describe 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 [11][12]:

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).

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.

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 (maximally negative) to +1 (maximally positive) scale for each adjective.[1] Instead of scoring adjectives from scratch, we used the framedwork in [7] that provides a score for adjectives on the -1 to $+1$ scale. Several other papers also score adjectives in other ways and could be plugged in here instead[13, 9].

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

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

[1] 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.

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 $f(adv, adj) \geq sc(adj)$.

Binary AACs We assign a score to a binary AAC $\langle adv_1 \cdot adv_2 \rangle \langle adj \rangle$ as follows. First, we 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.

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.

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 1. *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.

Adjective Priority Scoring. In Adjective Priority Scoring (APS), we select a weight $r \in [0, 1]$ that denotes the importance of an adverb compared to an adjective that it modifies. r can vary based on different criteria. The larger r is, the greater the impact of the adverb. 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 2. *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.

Adverb First Scoring. This algorithm is exactly like the previous algorithm except that the r parameter is applied to the adjective rather than to the adverb. 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 3. *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 can be implemented using well known algorithms [4]. Let us see how the above method works on a tiny example.

EXAMPLE 4. *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* **??**, *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 implemented all algorithms proposed in this paper on top of the OASYS system[7], as well as the algorithms described in [9, 3]. The implementation was 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 news articles scored by 10 students and 400 blog posts scored by 5 students.[2] We then conducted two sets of experiments on both blogs and news articles.

[2] The *training set used in OASYS was different from the experimental suite of 200 documents.*

Experiment 1 (Comparing correlations of algorithms in this paper). The first experiment tried to find the value of r that makes APS^r and $advFS^r$ provide the best performance using Pearson correlation as the measure of "best". Our news experiments gave the best r value as 0.35, while the blog experiments yielded a best value of 0.30. The figure below shows the Pearson correlation on the blog data as we vary r.

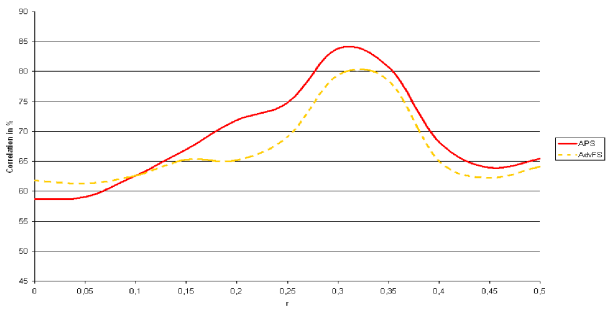

Fig. 1: *Pearson correlation coefficient for* APS^r *and* AdvFS^r

Experiment 2 (Correlation with human subjects). We We compared the algorithms in this paper with those described in [7, 9, 3]. The table below shows the Pearson correlations of the algorithms in this paper (with $r = 0.35$ for news data) compared to the algorithms of [7, 3, 9]. Similar results apply to blog posts.

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

Results. It is easy to see that APS^r with r in the 0.3 to 0.35 range has the highest Pearson correlation coefficient when compared to human subjects. It seems to imply two things: (i) First, that adjectives are more important than adverbs in terms of how a human being views sentiment and (ii)

that when identifying the strength of opinion expressed about a topic, the "weight" given to adverb scores should be about 30 to 35% of the weight given to adjective scores.

Inter-human correlations. Note that we also compared the correlations between the human subjects (on the news data). This correlation turned out to be 0.56. As a consequence, on a relative scale, $\text{APS}^{0.35}$ 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 [1][4], adverb scores depend on their collocation frequency with an adjective within a sentence, whereas in [2], scores are assigned manually by only one English speaker. These works do not distinguish between adverbs that belong to different classes. 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. Our experiments show that APS^r method is the best with a r around 0.3 or 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.

Acknowledgment. Work funded in part by AFOSR contract FA95500610405.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

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

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

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

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

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

[13] 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, volume IV, 2004, pages 11151118, Lisbon, Portugal.

[14] 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.