

# Sentiment Analysis: Does Coreference Matter?

Nicolas Nicolov\*, Franco Salvetti<sup>◇</sup> & Steliana Ivanova\*<sup>1</sup>

**Abstract.** We investigate the boost in sentiment performance by taking coreference information into account. We mine user generated content for market research. Frequent topic shifts in the data lead us to investigate sentiment algorithms which look at a window of text around topic key phrases. We describe and implement a lightweight sentiment analysis system and show how the system performance can be improved by about 10% when taking into account nominal and pronominal coreference elements.

## 1 Introduction

User generated content is viewed as a valuable source for mining market intelligence data. We study the unsolicited opinions of millions of users regarding companies, their products and services. As data sources we consider weblogs, message board data, surveys, etc.

In blogs there is substantial topic drift as users are describing different experiences and mentions of companies and products are often peripheral. Hence, we determine sentiment around occurrences of topic phrases and not the entire document (blogpost). More specifically, we consider proximity sentiment approaches where the algorithms look at certain left and right window of the topic keyword and in the current sentence. This differs from earlier document-level sentiment work (e.g., movie or product reviews). In error analysis between system output of such algorithms and human annotated data we have noticed cases where richer sentiment contexts are present around nominal and pronominal expressions corefering with topic words. Here are some motivating examples:

**A1** *Microsoft retools Zune<sub>i</sub> to target Apple's flaws.*

**A2** *The upgraded player<sub>i</sub> and a new strategy helps Redmond gain ground in its battle to beat the iPod.*

Just looking at the first sentence A1 even humans wouldn't be able to infer sentiment. Some systems might consider *flaws* to influence the sentiment negatively. In the second sentence A2 unigrams like *upgraded*, *helps*, *gain* can influence humans and systems positively. It's the coreference Zune<sub>i</sub>—player<sub>i</sub> that enables us (as humans) to make that inference.

**B1** *Well I guess I was one of the ones who were actually able to find a Zune<sub>i</sub> 80 as from what I have read they are currently in limited supply.*

**B2** *This is actually my first MP3 player purchase, and I had been waiting for the release of these players since I had first heard about the 2nd generation Zunes a few months back.*

**B3** *Now I have only had it<sub>i</sub> a day, but so far I am completely impressed with the quality of this player<sub>i</sub>.*

In sentence B1 it is challenging for a system to figure out that *limited supply* licenses a positive context. B2 is even more difficult. It is only in sentence B3 that we are assured of the positivity and even simple systems would be able to get this case (*impressed*, *quality*).

**C1** *I can't stop playing with my new Zune<sub>i</sub> 80.*

**C2** *It<sub>i</sub>'s lovely to look at and hold, the UI is great, and the sound is super.*

Again, in sentences C1 and C2 it's the context in C2 around the pronoun that reveals the sentiment (*lovely*).

Above we have been agnostic as to what type of sentiment algorithm we assume (even though we have been mentioning unigram and bi-gram features). The approach in this paper applies to sentiment algorithms which do not work at the entire document (document-level sentiment). This paper demonstrates that expanding the context on which the sentiment determination is made based on coreference leads to better sentiment results.

Hurst & Nigam mention the possible benefits of using anaphora resolution in conjunction with sentiment identification [4]. More recently Wu and Oard in the context of the NTCIR-6 Chinese Chinese opinion tasks explore (among other tasks) "a simple approximation to anaphora resolution on the accuracy of opinion holder identification" [6].

Sentiment analysis, in particular volume of positive and negative mentions, is heavily influenced by spam. Banking, telecommunications, electronics and automotive domains are notorious for being spammed. We have developed a number of techniques to eliminate spam ([5]) and for the purposes of this work assume clean data.

The techniques we describe are used in combination with other approaches in a larger deployed system at Umbria for large-scale market analysis.

The structure of the rest of the paper is as follows: In Section 2 we describe a lightweight sentiment approach. In Section 4 we discuss the data we have annotated to research the effects of coreference on sentiment. Statistics on the annotation are presented in Section 5. In Section 6 we show the boost in sentiment from coreference. We discuss further ramifications of this approach in Section 7 and conclude in Section 8.

## 2 Proximity-based, focused sentiment identification

For our sentiment algorithm we assume we have an English blog posting along with its title (we employ a fast, binary English–non-English language identifier). The text is tokenized (we pay special attention to emoticons as they are sentiment bearing), sentences have been identified, tokens are normalized, sentences are part-of-speech tagged, and phrases (noun groups) are identified. We also have the offsets where topic phrases have matched the retrieved document (we

<sup>1</sup> \*Umbria Inc., 4888 Pearl East Circle, Suite 300W, Boulder, CO 80302, U.S.A. <sup>◇</sup>University of Colorado & Powerset Inc., emails: [nicolas@umbrialistens.com](mailto:nicolas@umbrialistens.com), [franco.salvetti@colorado.edu](mailto:franco.salvetti@colorado.edu), [sivanova@umbrialistens.com](mailto:sivanova@umbrialistens.com).

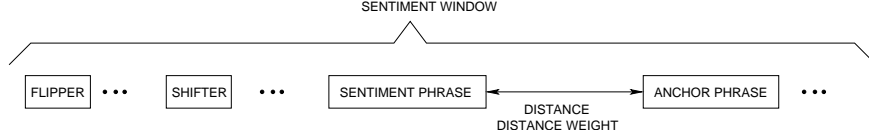


Figure 1. Elements the sentiment algorithm takes into account

will see examples for topic phrases in Section 4). We refer to the matched phrases as *anchors*.

The output of sentiment is a confidence value and a label such as: POSITIVE, NEGATIVE, MIXED, NEUTRAL or UNKNOWN

The sentiment algorithm proceeds by identifying a score for each anchor phrase and then aggregates the scores of all anchors.

### 2.1 Anchor-level sentiment score

To determine the score for an individual anchor the algorithm considers a sentiment window of 8 tokens before and 6 tokens after the anchor. We dynamically extend the boundaries of the window if the boundaries are inside noun groups. We calculate polarity scores by expanding outward from the anchor and checking if we encounter sentiment phrases. If yes we accumulate the score of the phrase multiplied by the distance weight. The distance weight function is:

$$weight(d) = \begin{cases} -\frac{1}{10}d + 2 & \text{if } d \in \{0, 1, 2, 3\} \\ 1 & \text{if } d \geq 4 \end{cases}$$

If the encountered phrase is a shifter (e.g., adverbs that enhance the sentiment—*horribly wrong*, *really good*) it is allow to influence the score of the word that follows it—we multiply the word’s value by the shifter value. The algorithm does separate passes to determine positive ( $score_{\oplus}$ ) and negative ( $score_{\ominus}$ ) scores. At this point the anchor score is:

$$anchor\_score = \frac{score_{\oplus} - score_{\ominus}}{score_{\oplus} + score_{\ominus}}$$

If flippers we encountered (e.g., negation particle *not*) we multiply the score by  $(-1)^{\#flippers}$  (e.g., *not without problems*):

$$anchor\_score = (-1)^{\#flippers} \cdot anchor\_score$$

Figure 1 depicts the elements involved in the anchor sentiment score calculation.

### 2.2 Aggregating anchor scores

The author opinion of the topic is a combination of the individual sentiment scores for each topic phrase occurrence:

$$avg = average(anchor\_score_1, \dots, anchor\_score_n)$$

$$v = variance(anchor\_score_1, \dots, anchor\_score_n)$$

The final decision rule for the sentiment label is:

- if  $avg < \tau_{\ominus}$  → NEGATIVE.
- if  $avg > \tau_{\oplus}$  → POSITIVE.
- if  $v > \tau_{mixed}$  → MIXED.
- else → NEUTRAL.

where:  $\tau_{\ominus}$ ,  $\tau_{\oplus}$  and  $\tau_{mixed}$  are negative, positive and mixed thresholds. We use the values  $\tau_{\ominus} = -0.25$ ,  $\tau_{\oplus} = 0.25$  and  $\tau_{mixed} = 0.5$ .

The confidence is calculated according to:

$$confidence = \begin{cases} v & \text{if MIXED} \\ |avg| & \text{if POSITIVE, NEGATIVE or NEUTRAL} \end{cases}$$

### 2.3 Sentiment phrases creation

Our positive and negative phrases (with qualification for strong and weak) augment initially manually created lists by automatic mining through the synonym relation in WordNet [1] using the following starting positive and negative seeds:

#### Adjectives:

- $\oplus$ : {good, nice, excellent, positive, fortunate, correct, superior, beautiful, amazing, successful}
- $\ominus$ : {bad, nasty, poor, negative, unfortunate, wrong, inferior}

#### Nouns:

- $\oplus$ : {joy, success}
- $\ominus$ : {disaster, failure, pain}

#### Verbs:

- $\oplus$ : {enjoy, like}
- $\ominus$ : {suffer, dislike, hate}

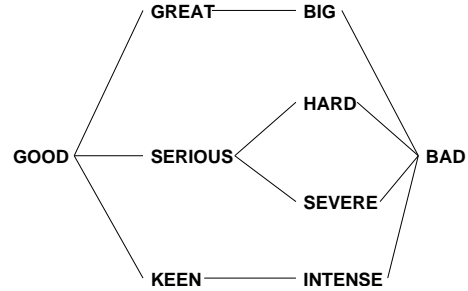


Figure 2. Various paths between ‘good’ and ‘bad’ in WordNet

For each part-of-speech tag and polarity we do a bounded, breadth-first expansion from each of the seeds, group the extracted items and associated with them the minimum distance to any of the seeds for this polarity. We then consider the positive and negative extracted items; elements for which the difference between the distance to positive and distance to a negative seed is less than a threshold of 2 are removed; the remaining elements in the intersection keep the polarity of the closest seed. Similar techniques are explored by Godbole et al. [3]. Figure 2 illustrates that ‘good’ and ‘bad’ are not that far from each other in the graph. Subsequently the automatic lists are validated by human annotators. The sentiment resources are continually tuned as part of error analysis.

TOPIC	NOMINALS	COREF EXPR.
R8 -MIDI -Leica	automobile, car, coupe, model, sportscar, supercar, conceptcar, ride, semi-exotic, vehicle, wheels	it, its, that, this
Zune	device, gadget, player	it, its, that, this

**Table 1.** Topic definitions

### 3 Coreference

The good news: Human languages have evolved to include various shortcuts to refer to previously introduced entities in the discourse. The bad news: Teaching computers the art of figuring which shortcut goes with which entity is not easy.

The latter problem comes under the heading of anaphora/coreference resolution. We consider a scenario where given a text we want to find proper names of entities (e.g., *George W. Bush*) and the other linguistic means to refer to the entities: short names (*George Bush*, *Bush*), nominals (*the U.S. president*) and pronouns (*he*, *his*, *him*). This is essentially the NIST-run automatic content extraction (ACE) task and we have been building systems to perform entity extraction, coreference resolution and relation extraction [2].

In some sense our present task is slightly easier as we can assume a human analyst providing topic terms (names of products) as well as their short forms, possible nominal expressions that can refer to the entity and corresponding pronouns. Still the challenge of referring an occurrence of *man* and *he* to *George W. Bush* vs. *Bill Clinton* who may also be mentioned in the text remains.

It is possible to create reasonably high baselines (in the order of 88% accurate assignments). The investigations in the remainder of this paper are based on perfect coreference from human annotation.

### 4 Data

We consider blog data for the topics “R8” and “Zune” for the period of October 2007–January 2008. We extract collections of blog postings that mention the topic keywords (cf. Table 1). The system keeps track of the occurrences of the topic terms. We also identify occurrences of the nominal and pronominal expressions for a topic. Each nominal and pronominal expressions is then considered in relation to the occurrences of the topic terms and is marked by human annotators according to the following scheme.

The nominal expressions are marked as referring to:

1. the previous noun group;
2. the first noun group to the left which matches by number and person but not case 1;
3. an anchor phrase but not case 1 nor 2; or
4. another noun group.

Pronominal expressions in addition to the above are marked as:

- modifier: *we didn't think it'd look this good*;
- expletive: *it seems, it appears*;
- determiner: *this car, that device*;
- wh-pronoun: *the car that captivated the audience*;
- subordinating conjunction/sentential complementizer: *I think that*;
- frozen phrases: *I miss Rowe kids very much. Well, some of them that is.*

If topic words match a different, unintended topic or if the pronoun is in text not part of the posting, we remove these examples:

- exclude: topic words wrongly chosen: *Notes R8; Email this.*

In the first example above *R8* refers to release 8; the second example is usually part of a footer of a document—identifying this is part of decomposition of a blog page which is challenging.

### 5 Annotation

We have annotated the postings extracted for the two topics with coreference information using the scheme described in Section 4. Table 2 show the distribution of different types of the coreference annotation:

Type \ Topic	R8	Zune
previous	11.32%	6.01%
num & pers match	0.93%	1.29%
coreferenced	33.49%	26.18%
modifier	0.78%	0.86%
determiner	14.26%	14.59%
expletive	6.36%	15.88%
complementizer	16.74%	24.03%
frozen phrase	0.93%	1.72%
wh-pronoun	3.57%	6.87%
other	57.36%	36.05%

**Table 2.** Distribution of coreference annotation types

The inter-annotator agreement between the two annotators participating in this project is 98.91% (overlapping annotations out of all annotations).

We have also annotated the data for sentiment. Figure 3 shows the web-based sentiment annotation tool.

### 6 Results

To answer the question in the title of the paper—yes, sentiment benefits from coreference information. Here are two examples where originally the sentiment algorithm returned NEUTRAL sentiment. After considering the coreference anchors sentiment changed to POSITIVE:

*Being first drawn to the Zune last year because of it's [SIC] style and awesomeness, I decided with it.*

*My parents got me a Zune. It has just come out on the market. It is something i have wanted for a long time and i was very excited about it.*

Table 3 shows how the sentiment system can improve its performance by taking coreference information into account. The improvement depends on the data and is about 10%. The bottom part of the table shows that in about 14% of the cases the coreferential contexts support the original sentiment determination.

List of the last few songs that just randomly played on itunes . This **lineup** is like a drummers fantasy . Or anyone obsessed with rhythm for that matter .

positive  probably positive  neutral  probably negative  negative

spam  foreign language  other  TBD

<< >>

[Umbria Home](#) - [Help](#)

Figure 3. Sentiment annotation tool

System \ Topic	R8	Zune
nominals	8.20%	5.62%
pronouns	9.13%	6.54%
nominals $\cap$ pronouns	12.73%	8.31%
nominals support	11.05%	5.77%
pronouns support	11.05%	11.54%
nom. $\cap$ pro. support	15.24%	13.46%

Table 3. Percent improvement due to coreference over the baseline sentiment analysis. Bottom part: percent of coreference support of original sentiment determined on topic anchors only

## 7 Discussion

As part of ongoing work we are exploring the coreference effect on sentiment in other domains (in marketing referred to as verticals). We are also investigation different data sources of user generated content—message boards and survey data. In preliminary investigations we have seen a fair use of pronouns in message board data. A major challenge is to properly recognize the quoting structure in such postings. Otherwise we attribute text to an author which is not written by them. For survey data we pay special attention to imperatives (e.g., *have better help and an adaptive user interface for the new release of the software*).

Flat proximity models for sentiment are easy to implement and the performance of such systems hinge upon the quality of the sentiment lists they use. We are investigating the use of dependency parsing in order to consider only those sentiment elements which are related to the anchors. Beyond the adjective modification the nature of that relationship is quite complex and there is additional computational cost associated with the parsing.

Sentiment is expressed not only directly (e.g., *product X is great*) but also indirectly—parts of the product are good; effects of the drug are good, etc. We approach this through automatic clustering. We also benefit from the fact that in our scenario we have human analysts who adjust output suggested by the system in the explorative stage.

We are also looking at validating our techniques across different languages—initially we are looking at German.

Audi R8 has won awards as ‘Car of the Year’ for 2007 among readership of many magazines. Some have given it this title for 2008. Zune is a MP3 and video player offered by Microsoft.

## 8 Conclusions

We have considered sentiment analysis on user generated content for market intelligence. Due the frequent topic drift we have argued for focused sentiment analysis which is performed on parts of the document around topic terms. We have described a lightweight, proximity-based sentiment algorithm and have shown that that the system can be improved by about 10% (depending on the topic) by augmenting the focus area of the algorithm using contexts around nominal and pronominal coreference elements.

## Acknowledgements

We would like to thank Martha Palmer, Jim Martin and Mike Mozer from the University of Colorado for useful discussions.

## REFERENCES

- [1] *An WordNet Electronic Lexical Database*, ed., Christiane Fellbaum, The MIT Press, 1998.
- [2] Radu Florian, Hany Hassan, Abraham Ittycheriah, Hongyan Jing, Nanda Kambhatla, Xiaoqiang Luo, Nicolas Nicolov, and Salim Roukos, ‘A statistical model for multilingual entity detection and tracking’, in *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics (HLT-NAACL’2004)*, pp. 1–8, Boston, Mass., (2–7 May 2004).
- [3] Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena, ‘Large-scale sentiment analysis for news and blogs’, in *International Conference on Weblogs and Social Media (ICWSM’2007)*, ed., Nicolov et al., pp. 219–222, Boulder, CO, (26–18 March 2007).
- [4] Matthew Hurst and Kamal Nigam, ‘Retrieving topical sentiments from online document collections’, in *Document Recognition and Retrieval XI*, (2004).
- [5] Nicolas Nicolov and Franco Salvetti, ‘Efficient spam analysis for weblogs through url segmentation’, in *Recent Advances in Natural Language Processing*, volume 292 of *Current Issues in Linguistic Theory*, 125–136, John Benjamins, Amsterdam & Philadelphia, (2007).
- [6] Yejun Wu and Douglas W. Oard, ‘Chinese opinion analysis pilot task’, in *Proceedings of NTCIR-6 Workshop Meeting, May 15–18, 2007, Tokyo, Japan*, (2007).