All Blogs Are Not Made Equal:

Exploring Genre Differences in Sentiment Tagging of Blogs

Alina Andreevskaia
Concordia University
andreev@cs.concordia.ca

Sabine Bergler
Concordia University
bergler@cs.concordia.ca

Monica Urseanu
Concordia University
m_ursean@cs.concordia.ca

Abstract

One of the essential characteristics of blogs is their subjectivity, which makes blogs a particularly interesting domain for research on automatic sentiment determination. In this paper, we explore the properties of two most common subgenres of blogs – personal diaries and "notebooks" – and the effects that these properties have on performance of an automatic sentiment annotation system, which we developed for binary (positive vs. negative) and ternary (positive vs. negative vs. neutral) classification of sentiment at the sentence level. We also investigate the differential effect of inclusion of negations and other valence shifters on the performance of our system on these two subgenres of blogs.

Keywords

Sentiment analysis, polarity identification

1. Introduction

Blogs (or weblogs) – online diaries displayed in reverse chronological order – are a new and fast growing genre of computer-mediated communication. Blogs contain a lot of information that is relevant for different categories of users: from blogger community to politicians, economists, sociologists, and other professionals. One of the essential characteristics of blogs is their subjectivity: most blogs contain very personal accounts of events, both private and social, along with the blogger's reactions to a variety of events and personal experiences. For that reason, blogs represent a genre of written communication particularly interesting for sentiment/subjectivity analysis. The results of sentiment tagging of blogs can be used in a large spectrum of applications: from psychological and sociological research to sales forecasting, public opinion assessment, financial markets analysis, and other.

The recent research on blog mood and sentiment is mostly focused on determination of general sentiment of an aggregate bloggers' community. For instance, Mishne and de Rijke have created the MoodViews [16] to track the general "blogosphere state-of-mind" over time [4] and to connect it with events that may affect bloggers' mood [3]. This approach captures the sentiment of the entire blogosphere at a given moment. Driven by the availability of mood indicators in posts on LiveJournal.com, this approach relies on this natural, user-made sentiment annotation for system training and evaluation. Other research, which is more in line with the general trends in sentiment research, has attempted sentiment/mood classification of individual blog postings [14, 17] in order to provide more specific information (e.g., Mishne and Glance demonstrated how bloggers' sentiment can be used for prediction of sales of a particular movie [17]). The analysis of sentiment of individual blogs has the potential to provide more detailed and more precise information that can tie sentiment not only to individual bloggers or to blog posting dates, but also to specific topics, threads and issues.

In this paper, we further advance the research in this direction by exploring sentiment annotation of blog posts at the sentence level. Sentences and clauses are often regarded as the most natural classification units for sentiment and subjectivity annotation [20], since sentiment in a single text often changes from one sentence to another, as the author develops his/her arguments or transitions from one topic to another within the same document. This fine-grained annotation has the potential to provide fairly specific information that can be used in the analysis of sentiment of a given social group or attitude toward particular event, product, or experience. Moreover, sentiment of the whole blog post can, if necessary, be derived from the sentiment of sentences that form it.

2. Data

While blogs are often treated as a special genre of computer-mediated communication [8, 15], they do not form a coherent genre in terms of style, language, and the purpose of communication. The three most commonly distinguished blog subgenres are personal journals, "notebooks", and filters [8]. While the definitions of these subgenres and the boundaries between them are still a matter of debate, some of their properties are already well defined: **personal journals** address the internal experiences and personal life of the blogger and are normally short, while **notebooks** are a mixture of comments on external and internal events and are characterized by longer, focused essays. Finally, **filters** are a collection of references (links) to external sources. Since they contain little or no text at all, the analysis of filters is beyond the scope of this study.

There are some marked differences between personal journals and "notebooks". Personal journals are typically written in a more colloquial manner, with wide use of emoticons, Internet abbreviations, and slang and with little attention to spelling and grammar. Notebooks, on the other hand, tend to be closer to newspaper text genres, such as editorials or letters to the editor, and, as newspaper texts, they tend to be more carefully proofread and use a more elaborate vocabulary and grammatical structures.

These differences between personal journals and notebooks suggest that these two subgenres of blogs can present different challenges for sentiment annotation and other NLP tasks. To test this hypothesis, we created two corpora, manually annotated for sentiment at the sentence level. One corpus was collected from several sites listed at cyberjournalist.net, which hosts mostly blogs written by journalists commenting on a variety of political and media-related events (thereafter, journalist blogs). The other represents a subset of the corpus distributed by the organizers of this conference and consists of English language personal journal posts from 20060501.xml file (thereafter, diaries). Both corpora, composed of 600 sentences each, were balanced to include 200 sentences with positive sentiment, 200 sentences with negative sentiment, and 200 sentences with no sentiment (objective) or with unclear sentiment (subjective neutral or mixed). For convenience, we will refer to the last group as "neutral" sentences.

In order to establish the inter-annotator agreement, 300 sentences from each corpus were independently annotated by two people. Since no sentence-level inter-annotator agreement studies on blogs have been conducted to date, such study was also important to establish the upper boundary for the system performance at the sentence level and to explore the rates of inter-annotator agreement on the two subgenres of blogs. We measured the agreement on two tasks: how consistently the two annotators can differentiate positive and negative sentences (binary classification) and how they can differentiate positive, negative, and neutral sentences (ternary classification). The agreement on positive vs. negative tags reached 95% ($kappa = 0.94$) for journalist-written blogs and 99% ($\kappa = 0.97$) for personal diaries. Separating blogs with positive or negative sentiment from neutrals, however, proved to be a much more complex task, which resulted in equally low (80%) inter-annotator agreement for both subgenres[1]. These results permit to conclude that, for human annotators, there was no significant difference in difficulty of sentiment annotation between journalist blogs and diaries, but for both subgenres, the task of tagging sentences for sentiment using ternary classification was significantly more difficult than the task of binary classification. This suggests that the expected system performance is also likely to be lower on the ternary than on binary classification.

3. Experiments

3.1 Method

The sentiment tagging of blog sentences presented here is done using a system that assigns sentiment to a sentence based on the count of sentiment clues encountered in it. The choice of this approach over statistical classifiers was motivated by several factors. First, such keyword-based methods performed best in sentiment tagging of sentences [9, 13]. Second, machine learning approaches require large amounts of labeled training data, which is hard to obtain for sentence level since it would involve extensive manual annotation.

In the experiments reported here, we tested two versions of the key-word-based system: (1) a basic system that uses only word counts and word sentiment scores that characterize system confidence in the sentiment value of a given word; and (2) a system that complements System 1 with the handling of negations and other valence shifters [18] and approximation of their scope.

[1] Corresponding Cohen's Kappa values are 0.55 and 0.59.

The system evaluation was conducted separately on both journalist blogs and diaries in order to explore the effects of these two subgenres on system performance. We wanted to find out, whether these two subgenres, equally non-problematic for human annotators, would also be processed by the system with comparable results. For a similar reason, we processed and evaluated the sets of positive, negative and neutral sentences separately, since performance of sentiment annotation systems is known to differ for positives and negatives.

3.1.1 System 1

The first set of experiments establishes the baseline for keyword-based sentiment tagging of the two blog subgenres. In these tests, the system's decision about sentence sentiment was based only on the count of sentiment-bearing words. There exist several lists of words manually classified as having positive or negative sentiment. The lists from [7](HM) and from General Inquirer [19] are used most commonly.

It should be noted that the coverage of such manually annotated lists is relatively small and, therefore, their use results in a very low recall. In order to improve the list coverage, we have extended the two manual lists using a method described in [1, 2]. This approach makes use of the information contained in WordNet [6] glosses and relations. This WordNet-based method assumes that sentiment-bearing words are likely to be explained in a dictionary gloss through a reference to other sentiment-bearing words. The method takes a manually annotated seed list (e.g., HM) as input and expands it using synonymy, antonymy and hyponymy relations. Then WordNet glosses are searched for the occurrences of the words from the expanded seed list. The head words that contained one of the seed words in their gloss are extracted and assigned the same sentiment as that seed word. Each word retrieved using this procedure is also assigned a score calculated based on (1) the number of times the word was retrieved in multiple system runs with different non-intersecting seed lists and (2) the consistency of sentiment assigned in these multiple runs. The scores were then translated into the words' degrees of membership in the fuzzy set of sentiment using a standard S-function [22]. Using these expanded lists, we then assigned sentiment to sentences by summing up fuzzy membership scores of individual sentiment-bearing words found in the sentence into the cumulative sentiment score of this sentence. In this procedure, the scores of positive words were coded as values above zero and the scores of negative words as values below zero, thus, the summation gave us a net sentiment score of the sentence.

3.1.2 System 2

The second set of experiments explores the impact of valence shifters [18] on sentence sentiment determination. Valence shifters are language elements that can change the sentiment of sentiment-bearing words in their scope. This category includes negations (*not, never*, etc.), words with increase/decrease semantics, and some other categories. In our experiments, we were mostly interested in the words that can reverse the sentiment or set it to zero. Experiments with negations and valence shifters reported in the extant literature [11, 5] produced mixed results, suggesting the need for further research in this area. The experiments described here are intended to explore the impact of negations and other valence shifters on sentence sentiment determination for the two subgenres of blogs and assess their role in tagging of positive,

negative, and neutral sentences. A manually compiled list of 75 such words (mostly negations) was used here. For efficient processing, the scope of valence shifters was approximated in a uniform manner: it was assumed to span from the valence shifter to the closest punctuation mark.

3.2 Results

3.2.1 Binary Classification

Each of the two systems was applied to both blog subgenres – personal diaries and journalist blogs. Table 1 presents the results of the two systems on binary (positive vs. negative) classification of the diaries and journalist blogs[2]. For each experiment, we show precision and recall on positive and on negative categories separately, as well as the accuracy of the binary classification, which separates sentences into positive and negative ones. For these experiments, the baseline accuracy of correctly choosing the class by chance was 50%.

Genre	Positive		Negative		Accuracy
	Prec.	Recall	Prec.	Recall	
System 1 (no valence shifters)					
Diaries	0.69	0.74	0.80	0.49	73%
J. blogs	0.59	0.69	0.71	0.46	64%
System 2 (with valence shifters)					
Diaries	0.74	0.72	0.81	0.54	77%
J. blogs	0.62	0.69	0.75	0.53	67%

Table 1: *System performance on binary classification*

3.2.2 Ternary Classification

The results of the ternary (positive vs. negative vs. neutral) classification (Table 2) were considerably lower than those of binary classification due to greater complexity of this task: more system errors and more disagreement between human annotators occur on the tasks of separation of neutrals from sentiment-bearing (positive or negative) sentences. The accuracy of the ternary classification for personal diaries was 51% for the basic system and 53% for the system enhanced with valence shifter handling. On the journalist blogs, the numbers were 48% and 50% respectively[3]. For these experiments, the baseline accuracy of correctly choosing the class by a random pick was 33.3%.

Genre	Positive		Negative		Neutral		Acc.
	P	R	P	R	P	R	
System 1 (no valence shifters)							
Diaries	0.50	0.74	0.54	0.49	0.50	0.31	51%
J. blogs	0.42	0.69	0.51	0.46	0.60	0.28	48%
System 2 (with valence shifters)							
Diaries	0.51	0.72	0.59	0.49	0.48	0.33	53%
J. blogs	0.45	0.69	53	0.53	0.60	0.29	50%

Table 2: *System performance on ternary classification*

3.2.3 The Effects of Blog Subgenre Differences

While human annotators tagged the two corpora with fairly equal inter-annotator agreement, the Systems 1 and 2 proved to be more sensitive to the genre differences. The experiments showed that sentences from personal diaries were easier to annotate with positive and negative sentiment than sentences from the blogs written by journalists. On the task of binary classification, this observation holds for both the automatic system (77% vs. 67% accuracy respectively[4]) and for human annotators (99% vs. 95% inter-annotator agreement[5]). This can be attributed to the fact that the syntactic structures of sentences in the personal diaries are usually less complex, are much shorter (in our dataset, the average length of a sentence in the diaries is 13 words and in journalist blogs – 20 words), and include fewer sentiment-bearing words. The Table 3 below provides detailed summary of linguistic properties of the two corpora used in this study: the average length of sentences in words (Avg. length), the average number of sentences with valence shifters (V. shifters), and the average number of sentiment-bearing words per sentence (Sent. words). The difference between personal diaries and journalist blogs is particularly salient for negative sentences.

Sentiment	Avg. length	V. shifters	Sent. words
Diaries			
Positive	15.12	0.15	2.15
Negative	14.42	0.32	2.01
Neutral	14.44	0.21	1.76
Journalist Blogs			
Positive	18.37	0.20	2.68
Negative	28.10	0.33	4.14
Neutral	17.40	0.17	2.11

Table 3: *Statistics for the three sentiment categories*

4. Discussion

Overall, our system classified blog sentences with performance comparable to other similar genres explored in the literature. Our 64–67% accuracy on journalist blogs, for the system based on keyword count and scoring, is comparable with the results of similar experiments on news sentences: 67% [12] and 68% [21]. Online messages, which were classified in [10], are the closest genre to the personal diaries. The system used by Hurst and Nigam included treatment of negation and, thus, is more comparable to our System 2. The precision reported in [10] was 80% for negatives and 82% for positives, which is in line with our system's precision on these categories for personal diaries in binary classification.

The introduction of the coarse-grained handling of valence shifters into the system produced an improvement for both genres. For positive sentences, this feature affected mostly the precision which increased by 5% for personal diaries and by 3% for journalist blogs, while for negative sentences, the impact on precision was smaller (1% for personal diaries and 4% for journalist blogs), the recall, however, went up by 5% and 7% for personal journals and journalist blogs respectively (Table 1). Overall, our System 2, which included the coarse approximation of valence shifters' scope was able to correctly tag 54–56% of all sentences with valence shifters compared to 45–47% when the system is run without this module (System 1). This suggests that a more sophisticated handling of valence shifters (e.g., extended lists of valence shifters and more

[2] The difference between the two systems is statistically significant at $\alpha = 0.1$ for both subgenres.

[3] This difference was not statistically significant on our dataset.

[4] The difference between the genres is significant at $\alpha = 0.01$.

[5] This difference is significant at $\alpha = 0.01$.

refined approximation of their scope) would bring marked improvements.

The relatively low accuracy on the three-value classification, as well as the low inter-annotator agreement on this task, is not surprising and is only partially accounted for by the lower baseline probability of the correct random pick: on binary classification the chance of correct random pick is 50%, while on ternary classification it is only 33.3%. The task of ternary classification is much more complex on any genre of text, and on blogs in particular. The specifics of blogs as a genre play a role here: while blogs are almost always subjective and rarely include clearly factual, objective sentences, the specific valence of the emotion (positive or negative) is often not clearly discernible in blogs (e.g., *oh my god!*). This abundance of sentiment-laden sentences with unclear sentiment valence complicates the task of ternary sentiment classification both for humans and for computer systems.

5. Conclusion

Blogosphere, which is usually seen holistically, is composed of different types of texts and of different blog subgenres. In this study we have focused on two most common blog subgenres: personal diaries and "notebooks".

In the first set of experiments reported here, we explored the properties of these two subgenres in relation to sentiment tagging at the sentence level. The inter-annotator agreement study showed that for humans the personal diaries were easier to annotate with positive and negative sentiment than texts written by journalists. This difference was maintained in the results of the automatic sentiment tagging experiments. In this study we have identified some important properties of journalist blogs as a subgenre that may account for the observed differences: longer sentences, more extensive use of sentiment-bearing words, and more complex syntactic constructions with greater use of valence shifters.

In the second set of experiments on the two subgenres, we studied the contribution of valence shifter handling module with a basic approximation of the valence shifter scope. We discovered that, for both subgenres, the inclusion of this module resulted in improved system precision with no negative effect on recall. Our future efforts will focus on further refinement of valence shifter handling, since, given the current system accuracy in annotation of sentences with valence shifters, the potential for performance gains in this area is substantial.

References

[1] A. Andreevskaia and S. Bergler. Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *Proc. EACL-06*, Trento, Italy, 2006.

[2] A. Andreevskaia and S. Bergler. Semantic tag extraction from wordnet glosses. In *Proc. of LREC-06*, Genova, Italy, 2006.

[3] K. Balog and M. de Rijke. Decomposing bloggers boods: Towards a time series analysis of moods in the blogosphere. In *Proc. of WWW2006*, Edinburgh, UK, 2006.

[4] K. Balog, G. Mishne, and M. de Rijke. Why are they excited: Identifying and explaining spikes in blog mood levels. In *Proc. of EACL-06*, Trento, Italy, 2006.

[5] K. Dave, S. Lawrence, and D. M. Pennock. Mining the Peanut gallery: opinion extraction and semantic classification of product reviews. In *Proc. of WWW03*, Budapest, Hungary, 2003.

[6] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[7] V. Hatzivassiloglou and K. B. McKeown. Predicting the Semantic Orientation of Adjectives. In *Proc. of the 35th ACL*, Somerset, NJ, 1997.

[8] S. C. Herring, L. A. Scheidt, S. Bonus, and E. White. Bridging the Gap: A Genre Analysis of Weblogs. In *Proc. of the 37th Hawaii International Conference on System Sciences*, Los Alamitos, CA, 2004.

[9] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proc. of KDD-04*, 2004.

[10] M. Hurst and K. Nigam. Retrieving topical sentiments from online document collection. In *Exploring Attitude and Affect in Text: theories and application (AAAI-EAAT 2004)*, Stanford University, 2004.

[11] A. Kennedy and D. Inkpen. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22(2):110–125, 2006.

[12] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proc. of COLING-04*, Geneva, SZ, 2004.

[13] S.-M. Kim and E. Hovy. Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proc. of IJCNLP-05*, Jeju Island, Korea, 2005.

[14] G. Leshed and J. J. Kaye. Understanding How Bloggers Feel: Recognizing Affect in Blog Posts. In *Proc. of CHI-2006*, Montreal, Canada, 2006.

[15] C. R. Miller and D. Shepherd. Blogging as social action: A genre analysis of the weblog. In L. Gurak, S. Antonijevic, L. Johnson, C. Ratliff, and J. Reyman, editors, *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. Minneapolis, MN, 2004.

[16] G. Mishne and M. de Rijke. Moodbiews: Tools for blog mood analysis. In *Proc. of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, Stanford, CA, 2006.

[17] G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In *Proc. of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, Stanford, CA, 2006.

[18] L. Polanyi and A. Zaenen. Contextual Valence Shifters. In J. G. Shanahan, Y. Qu, and J. Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application*. Springer Verlag, 2006.

[19] P. J. Stone, D. C. Dunphy, M. S. Smith, D. M. Ogilvie, and associates. *General Inquirer. A computer Approach to content analysis*. M.I.T. press, 1997.

[20] J. Wiebe, R. Bruce, M. Bell, M. Martin, and T. Wilson. A corpus study of evaluative and speculative language. In *Proc. of the 2nd ACL SIGDial Workshop on Discourse and Dialogue*, Aalberg, Denmark, 2001.

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

[22] L. A. Zadeh. Calculus of Fuzzy Restrictions. In L. Zadeh, K.-S. Fu, K. Tanaka, and M. Shimura, editors, *Fuzzy Sets and their Applications to cognitive and decision processes*, pages 1–40. Academic Press Inc., New-York, 1975.