ANALYZING WRITING STYLES OF BLOGGERS WITH DIFFERENT OPINIONS

Thomas H. Park, Jiexun Li, Haozhen Zhao
College of Information Science and Technology
Drexel University
{thomas.park, jiexun.li, haozhen.zhao}@ischool.drexel.edu

Michael Chau
School of Business
The University of Hong Kong
mchau@business.hku.hk

Abstract

Understanding customers is crucial to companies' decision-making. With the advent of Web 2.0, more and more people choose to express their feelings and articulate their attitudes through online social communities such as blogs and web forums. These new sources of information offer the potential to obtain large quantities of customer feedback using automated analysis techniques. In this paper, we study how people with different opinions toward a commercial product write differently in their blogs. We define and extract four types of stylometric features – namely lexical, syntactic, structural, and sentimental features – to represent a blogger's writing style. Based on multivariate analyses on a data set of iPod-related blogs, we found various writing style patterns of bloggers. Our analysis shows that a blogger's writing style is marginally related to his or her opinion toward a product.

Keywords: Blog Analysis, Stylometry, Principal Components Analysis, MANOVA

1. Introduction

Understanding customers is crucial to companies' decision-making. Traditionally, companies have relied on expensive, time-consuming methods to collect customer feedback, including focus groups, surveys, and professional evaluations. With the advent of Web 2.0, a growing number of people are sharing their opinions about commercial products and services through online social communities such as blogs and web forums. These new forms of communication offer customer feedback that is timely, abundant, and amenable to automated text-mining techniques.

Our position is that the opinions expressed by bloggers toward a commercial product relate to the writing styles in which they express themselves. A stylometric analysis of these bloggers may yield insights into how automated techniques can be used to interpret customer feedback found online. In this study, we pose the following research questions:

- How do bloggers with different opinions toward a product differ in their writing styles?
- What are the business implications of the writing style patterns found among different groups of bloggers?

2. Literature Review

2.1 Stylometry Analysis

This study is closely related to a linguistic research area called stylometry, which refers to the statistical analysis of literary style (Holmes 1998). The central task of stylometry is to extract a set of features that can represent the writing styles of a particular author. These features are also referred to as "writeprints." Stylometry has been successfully applied to determine authorship of historical literature. For example, Mosteller and Wallace (1963) used the frequency of content-free "function words" to attribute the disputed *Federalist* papers to James Madison. Elliott and Valenza (1991) incorporated stylometric features to discredit the Earl of Oxford as the true identity of William Shakespeare. In addition to authorship identification, content analysis has been used for authorship characterization. Burrows (1989) used function words to cluster sixteen novelists by gender and chronology.

In recent years, stylometry has been directed toward the Web. Stamatatos et al. (1999) attributed essays from a newspaper's website to one of ten authors using stylometric features. de Vel (2000) identified the authors of emails based on stylometric and email-specific attributes including the use of attachments and signatures. Zheng et al. (2006) used stylometric features to identify the authors of piracy-related newsgroup and bulletin board system (BBS) messages. Abbasi and Chen (2008) developed a framework for text analysis of computer-mediated communication (CMC) in which various stylometric features are included.

2.2 Blog Analysis

Blogs serve as online diaries for people to document their life, express their views on various topics, and form and maintain social relationships. Both content analysis and link analysis techniques have been used to discover useful information from blog content and the blogosphere structure. In recent years, stylometric analysis has been applied to blog content in order to characterize its authors. Burger and Henderson (2006) explored features for predicting the age of bloggers, including stylometric features and blog-specific features like links, images, profiles, and “friend links”. Yan and Yan (2006) combined stylometric features with “non-traditional” ones, including background colors, fonts, and emoticons, to classify the gender of bloggers. Nowson et al. (2005) examined the correlation between the frequencies of different parts-of-speech and the personality traits of bloggers. Datta and Sarkar (2008) utilized stylometric features to discriminate legitimate blogs from spam blogs.

From a business intelligence perspective, Chau and Xu (2007) explored the attitudes expressed in blogs toward a commercial product using social network analysis. Their analysis showed that different opinions toward a product do not keep bloggers from interacting with one another. In this study, we aim to build on this work by looking at how bloggers with different attitudes toward a product write differently.

3. Data Description

Our corpus was comprised of blog entries authored by people who expressed sentiment toward Apple’s iPod portable media players through their group affiliations. Blog entries were collected from Xanga, a popular blog hosting website that supports “blogrings” - groups formed around a common interest or circle of friends. An initial set of blogrings was identified by the use of the word “iPod” in their titles or descriptions, and refined by discarding spam and other irrelevant groups. The resulting 201 groups were then manually classified as having a positive, negative, or neutral sentiment toward iPods based on their descriptions.

Members of these blogrings were assigned one of the three sentiment labels as expressed by their membership to iPod-related groups. Bloggers in both a neutral and a positive (or negative) group were assigned a positive (or negative) label. None of the bloggers were in a positive and negative blogring simultaneously. From this set of bloggers, 1,990 bloggers that had posted 1,000 or more words of text were selected. Notably, these authors seldom mentioned iPods in their blog entries despite their membership in iPod-related groups (Chau & Xu 2007). Table 1 provides an overview of the corpus.

Table 1. Summary of Corpus

	Blogrings		Bloggers	
Sentiment	**Count (Percentage)**	**Mean Number of Members**	**Count (Percentage)**	**Mean Number of Entries**
Positive	104 (51.7%)	23.11	1,346 (67.6%)	68.96
Negative	33 (16.4%)	7.85	159 (8.0%)	85.04
Neutral	64 (31.8%)	14.48	485 (24.4%)	76.14
Total	201 (100%)	17.86	1,990 (100%)	71.99

4. Sylometric Feature Extraction

Individuals possess writing styles that remain consistent in their works, even across multiple topics. By measuring characteristics of an author's writings, a "writeprint" can be created that captures his or her unique writing style (Li et al. 2006). In our analysis, we measured 39 stylometric features that fall into four categories: lexical, syntactic, structural, and sentimental.

The lexical features characterize a person’s writing style using six character-based features, such as the frequency of special characters (e.g., ‘!’, ‘$’), and eleven word-based features, including the frequency of hapax legomena (i.e., words that occur exactly once in an entry) and hapax dislegomena (i.e., words that occur exactly twice). The syntactic features measure the use of five parts of speech as well as comparatives (e.g., ‘better’, ‘faster’), superlatives (e.g., ‘best’, ‘fastest’), *wh-* words (e.g., ‘who’, ‘why’), and function words. Function words are common words that have been identified as being content-free and context-independent (e.g., ‘the’, ‘and’). The structural features describe the author’s organization of content through the use of sentences and paragraphs. Lastly, the sentimental features indicate the direction and degree of sentiment expressed by the words in the texts. These stylometric features are outlined in Table 2 and explained in detail by Li et al. (2008).

Table 2. Stylometric Features

Lexical Features	**Syntactic Features**
Total number of words	Frequency of nouns
Total number of distinct words	Frequency of proper nouns
Average word length	Frequency of verbs
Standard deviation of word length	Frequency of adjectives
7 Vocabulary richness measures	Frequency of adverbs
Total number of characters	Frequency of comparatives
Frequency of English characters	Frequency of superlatives
Frequency of uppercase characters	Frequency of *wh-* words
Frequency of lowercase characters	Frequency of function words
Frequency of numerical characters	
Frequency of special characters	
Structural Features	**Sentimental Features**
Total number of lines	Frequency of subjectivity clues
Total number of paragraphs	Frequency of strong positive words
Total number of sentences	Frequency of weak positive words
Average number of words per sentence	Frequency of strong negative words
Average number of words per paragraph	Frequency of weak negative words
Average number of sentences per paragraph	Frequency of strong neutral words
	Frequency of weak neutral words

5. Data Analysis and Discussion

After extracting the stylometric features from each blog entry, we calculated their mean values for each author. This resulted in a matrix whose rows represented the bloggers and whose columns represented the stylometric features. This matrix was analyzed using principal components analysis (PCA) and multivariate analysis of variance (MANOVA).

5.1 Principal Components Analysis

We performed a PCA with varimax rotation in order to visualize and interpret the stylometric data in a lower dimensional space. The 39 stylometric features were distilled into two principal components accounting for 39.8% of the variance in the data. The PCA results demonstrate significant correlation among the stylometric features, including between features of different categories. Figure 1 shows how the features loaded on these principal components.

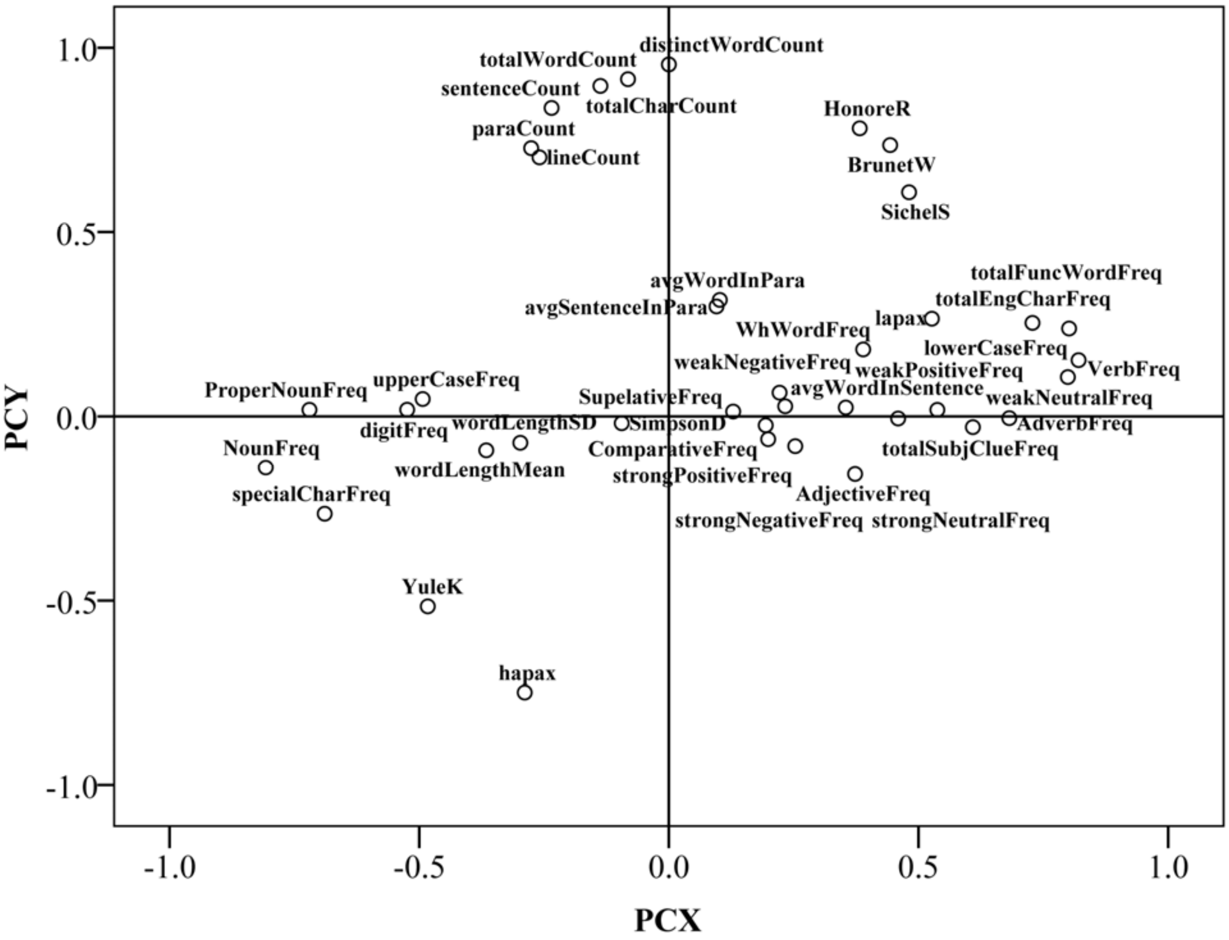

Figure 1. Plot of Variable Loadings on Principal Components

The first principal component PC_X represents a subjectivity dimension, with subjectivity words, adjectives, and adverbs as positive contributors, and the more objective nouns, proper nouns, and numerals as negative contributors. The second principal component PC_Y corresponds to the size of the blog entries, with counts of words, characters, and sentences loading most positively. The vocabulary richness measures lie on a diagonal, suggesting a third major component.

Bloggers from the three sentiment groups were plotted in the space defined by PC_X and PC_Y. Figure 2 shows that the majority of bloggers are concentrated toward the bottom right. In other words, most bloggers express personal viewpoints through relatively brief entries. However, a substantial number of bloggers radiate above and to the left of this core, representing longer entries and more objective tones, respectively.

Any further interpretations based on these principal components must be made with caution. Conclusions drawn from stylometric analysis of books, articles, and other forms of edited text do not necessarily hold for blogs due to their unrestricted nature. For example, typical uses of capitalization include names and the start of sentences. These counts are often eclipsed, however, when bloggers use capitalization for emphasis (e.g., "*MY COMPUTER DIED*") or aesthetics (e.g., "*DiNnEr WaS sO nIcE tOo*"). These usages underscore the difficulty in drawing meaningful conclusions of bloggers based on current stylometric analysis techniques, and the need to develop features that better discern attitudes and intentions from text found online.

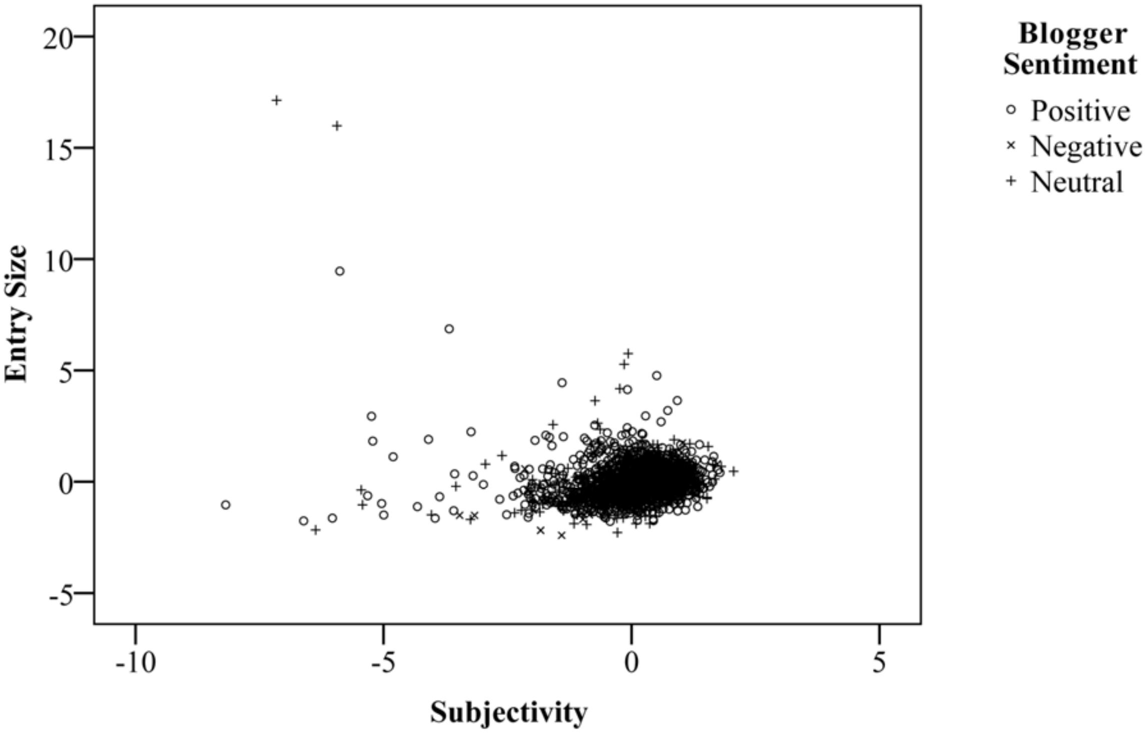

Figure 2. Plot of Bloggers by Sentiment

5.2 MANOVA

A multivariate analysis of variance (MANOVA) was used to determine whether stylistic differences existed between bloggers with differing sentiment toward iPods. The Wilks' Lambda value of 0.942 ($p < 0.05$) indicates that sentiment has a significant effect on stylometric features.

Helmert contrasts showed that the entries of neutral bloggers had, on average, 5.3 more lines, 3.4 more paragraphs, and 1.2 more words per sentence than those of non-neutral (positive and negative) bloggers, all significant at the 0.05 alpha level. Neutral bloggers predictably also had lower usages of strong negative words than non-neutral bloggers. Between the two non-neutral groups, negative bloggers used more subjectivity words with strong negative, weak negative, and surprisingly, strong positive polarities when compared with positive bloggers. Finally, bloggers with positive sentiment used 11.4 more distinct words per entry than negative bloggers, hinting at further patterns in vocabulary richness.

In short, MANOVA revealed that bloggers who expressed neutral opinions toward iPods tended to have long entries with few subjective words, negative bloggers had short entries with many subjective words, and positive bloggers were in between. Figure 2 offers some visual support for these results. Nevertheless, there is a high degree of overlap between the different sentiment groups, echoing earlier findings that "even if a blogger is negative (or positive) about iPod[s], he/she still interacts with other bloggers who may be positive (or negative)" (Chau & Xu 2007).

It should be noted that the stylistic differences found among the sentiment groups do not manifest exclusively in writings about iPods; rather, they correspond to the underlying personalities of the bloggers. Some use blogs as a platform for expressing their opinions about a wide range of topics, whereas others use blogs for posting more objective or reflective information. This raises questions of how bloggers with different writing styles influence the opinions of others differently, and how shared opinions can grow into trends online. Are bloggers who evaluate products passionately or dispassionately more influential within their communities? Does one opinionated blogger with a negative sentiment drown out the many positive bloggers? Further study is needed to understand these phenomena.

6. Conclusion

In our study, we discovered that the writing styles of bloggers could be characterized in terms of subjectivity and entry size, and captured using specific stylometric features. We extracted stylometric features from the blog entries of people who expressed opinions toward a commercial product through their membership in blogrings. We then used multivariate statistical techniques to analyze and visualize these bloggers based on their writing styles, finding a marginal relationship between the writing styles of bloggers and their attitudes toward a product.

These results are an early step in realizing the potential of data found in blogs and web forums for business intelligence. Nevertheless, our results suffered from the gulf between blogring affiliation and blog content. We hope to discover stronger patterns by incorporating more sophisticated stylometric features and by applying our methodology to content more closely tied to the product of interest. Lastly, we believe that by integrating our stylometric approach with topical and network analyses, a more complete understanding of bloggers' online expressions toward a product can be achieved.

References

Abbasi, A. & Chen, H. 2008. “CyberGate: A design framework and system for text analysis of computer-mediated communication,” *Management Information Systems Quarterly* (32:4), pp. 811-837.

Burger, J.D. & Henderson, J.C. 2006. “An exploration of observable features related to blogger age,” *AAAI 2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, pp. 15-20.

Burrows, J.D. 1989. “‘An ocean where each kind…': Statistical analysis and some major determinants of literary style,” *Computers and the Humanities* (23), pp. 309-312.

Chau, M. & Xu, J. 2007. “Studying customer groups from blogs,” *6th Workshop on e-Business*.

Datta, S. & Sarkar, S. 2008. “A comparative study of statistical features of language in blogs-vs-splogs,” *2nd Workshop on Analytics for Noisy Unstructured Text Data* (303), pp. 63-66.

de Vel, O. 2000. “Mining e-mail authorship,” *Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining*.

Elliott, W. & Valenza, R. 1991. “Was the Earl of Oxford the true Shakespeare?” *Notes and Queries* (38), pp. 501-506.

Holmes, D.I. 1998. “The evolution of stylometry in humanities scholarship,” *Literary and Linguistic Computing* (13:3), pp. 111-117.

Li, J., Zheng, R., & Chen, H. 2006. “From fingerprint to writeprint: Identifying the key features to help identify and trace online authorship,” *Communications of the ACM* (49:4), pp. 76-82.

Li, J., MacDonald, C.M., & Zheng, R. 2008. “Stylometric feature selection for assessing review helpfulness,” *18th Annual Workshop on Information Technologies and Systems*.

Mosteller, F. & Wallace, D. 1963. “Inference in an authorship problem,” *Journal of the American Statistical Association* (58:302), pp. 275-309.

Nowson, S., Oberlander, J., & Gill, A. 2005. “Weblogs, genres, and individual differences,” *27th Annual Conferences of the Cognitive Science Society*, pp. 1666-1671.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. 1999. “Automatic authorship attribution,” *9th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 158-164.

Wilson, T., Wiebe, J., & Hoffmann, P. 2005. “Recognizing contextual polarity in phrase-level sentimental analysis,” *Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing.*

Yan, X. & Yan, L. 2006. “Gender classification of weblog authors,” *AAAI 2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, pp. 228-230.

Zheng, R., Li, J., Chen, H., & Huang, Z. 2006. “A framework for authorship identification of online messages: Writing-style features and classification techniques,” *Journal of the American Society for Information Science and Technology* (57:3), pp. 378-393.