Using Corpus Analysis to Inform Research into Opinion Detection in Blogs

Deanna Osman[1] **John Yearwood**[2] **Peter Vamplew**[3]

School of Information Technology and Mathematical Sciences
University of Ballarat,
P.O. Box 663, Ballarat Victoria 3353, Australia,

[1]Email: d.osman@ballarat.edu.au
[2] Email: j.yearwood@ballarat.edu.au
[3] Email: p.vamplew@ballarat.edu.au

Abstract

Opinion detection research relies on labeled documents for training data, either by assumptions based on the document's origin or by using human assessors to categorise the documents. In recent years, blogs have become a source for opinion identification research (TREC Blog06). This study analyses the part-of-speech proportion and the words used within various corpora, determining key differences and similarities useful when preparing for opinion identification research. The resulting comparisons between the characteristics of the various corpora is detailed and discussed. In particular, opinion-bearing and non-opinion Blog06 documents were found to display a high level of similarity, indicating that blog documents assessed at the document level cannot be used as training data in opinion identification research.

Keywords: Blogs, Weblogs, Blog06, TREC, Opinion detection, Opinion identification

1 Introduction

Weblogs (blogs) are a fast growing phenomenon on the World Wide Web as they allow people to publish their thoughts and opinions on any topic they choose. In September 2007 a blog tracking company, Technorati, Inc., reported that it was monitoring 104.9 million blogs worldwide (*About Technorati, Accessed September* 2007), up from 4.2 million in October 2004 (Rosenbloom 2004).

The majority of blog authors surveyed by Lenhart & Fox (2006) indicated that the reason they write blogs is to share their knowledge and skills, with a high proportion of the topics being about personal and life experiences. Blog authors are inspired by the things that happen to them and want to share these experiences. Often blog authors will express their opinions about products, event and people which impacts their lives. Automatically gathering and the analysis of these opinions could prove valuable in a number of applications.

Such a search engine could be used by manufacturers to access opinions on their products or a competitor's product. For example, negative opinions about a competitor's product may provide a competitive edge for a new design. Governments could search blogs for qualitative information to support quantitative research (opinion polls) regarding new policies or upcoming elections. Small businesses, who do not have a large 'market research' budget, could gain access to millions of people who potentially have an opinion relating to them.

Searching the blogosphere for opinions about life experience and other topics within blogs, is an arduous task using traditional search engines. A search engine that searches the blogosphere for opinions on a given topic requires the inclusion of an opinion identification module in the search engine architecture. The task of opinion identification has previously been investigated in a non-blog context. Newswire articles have been used in opinion identification research (Yu & Hatzivassiloglou 2003, Kim & Hovy 2005) to create training and testing data, by dividing the articles into opinion-bearing and non-opinion-bearing categories. Editorial and Letter to editor articles were assumed to be opinion-bearing, while Business and News articles were categorised as non-opinion-bearing (Yu & Hatzivassiloglou 2003, Kim & Hovy 2005).

These documents formed the training and testing data for Naive Bayes machine-learning (Yu & Hatzivassiloglou 2003), and were used to create a list of opinion-bearing and non-opinion-bearing words and opinion scores[1] (Yu & Hatzivassiloglou 2003, Kim & Hovy 2005). This list was expanded by adding synonyms and antonyms of opinion-bearing and non-opinion-bearing words (Kim & Hovy 2005). The original list (Yu & Hatzivassiloglou 2003) comprised of adjectives, adverbs, nouns and verbs.

The resulting list of words (adjectives, nouns, verbs & adverbs) was used to identify opinion-bearing sentences (Kim & Hovy 2005) by applying the scores to the words within the sentences. Sentences were assessed by three evaluators to enable precision and recall to be calculated. The Wall Street Journal articles in each category were not evaluated to determine the validity of the hypothesis that Editorial/Letter to editor articles are opinion-bearing and Business/News articles are non-opinion-bearing.

Opinion identification research was sponsored by the Text REtrieval Conference (TREC) within blogs for the first time in 2006. TREC collected blog posts and comments over an eleven week period to create a blog track (Blog06). One of the tasks for participants was to identify opinion-bearing blogs on a given topic. This task was made more

[1]Opinion scores indicate how strongly a word expresses an opinion.

difficult by the lack of an annotated blog corpus for training data (Yang, Si & Callan 2006, Zhang & Zhang 2006). Various other corpora were used by Blog06 participants as the training data, including the list created from the Wall Street Journal collection (Yu & Hatzivassiloglou 2003, Kim & Hovy 2005).

The results of the opinion identification task within Blog06 were varied (Ounis, de Rijke, Macdonald, Mishne & Soboroff 2006), with Mean Average Precision (MAP) ranging from 0.2983 to 0.0001. A question that arises from the Blog06 results is *'Does identifying opinions within blog posts and comments require different training data to identifying opinions within more traditional corpora?'*

Blogs are an informal form of communicating, where usually the audience of the blog is known to the author (Nardi, Schiano, Gumbrecht & Swartz 2004). The author of a blog is free to write informally and use any language required to express their thoughts and opinions.

On the other hand, newswire articles are written using a formal structure, using proper English without slang and word abbreviation. These articles have trained, experienced authors and an editor to ensure high quality writing techniques are used and words are not used out of context or with ambiguous meaning.

Therefore, it might be expected that blogs may exhibit different language usage and characteristics from other document corpora and training data developed from those corpora may not be applicable to a blog corpora. This study provides an analysis of various corpora and reports on the differences, with the view to gaining an insight into how blogs differ from traditional opinion identification corpora. A broad view of the characteristics of opinion-bearing versus non-opinion-bearing text within different corpora is also provided. The corpora analysed are listed in Table 1.

There are two main approaches to the corpora analysed:

1. The proportion of part-of-speech types within each corpus (Section 3.2).
2. The use of unique or 'weird' (Gillam & Ahmad 2005) words and slang (Section 3.3).

A similar pattern between the opinion and non-opinion corpora respectively in the above-mentioned approaches was not found, whilst the opinion and non-opinion blog corpora were found to display similar characteristics to each other. The lack of variation between BlogOp and BlogNop led to the non-relevant text being removed from these corpora and smaller sentence corpora being created, section 3.1 details the methodology applied. Further analysis of the opinion and non-opinion blog sentence corpora found a greater variation in the characteristics analysed in this study, indicating the need for analysis of the relevance of the text with blog documents prior to training and testing data being created for opinion detection research.

The remainder of this paper is organised as follows. Section 2 describes the various corpora that are listed in Table 1. Section 3 describes the methodology applied to each of the areas of analysis. The results of the analysis are discussed in section 4, with section 5 concluding the study and discussing future work.

2 Corpora

There are many document collections available for opinion identification and other research, with each one having different characteristics and features. Some are assessed to assist researchers when analysing their research, whilst others are generic document collections that need to be assessed according to the research area. This study includes the main corpora used by Blog06 participants and reports key differences between these and the analysis of blogs extracted from the assessed Blog06 data. This section describes each corpus used in this study and lists the total number of word types (distinct words) and the total number of tokens (words) in each corpus. A summary of the total number of word types and tokens in each corpus is detailed in Table 3.

Table 1: Corpora analysed in this study and the category each has been assigned to. Note: The Blog06 and Customer Review corpora has been assessed by human assessors. The Movie Review corpus is based on assumptions detailed in Section 2.8. An assumption has been made (for this study) on the category for the remaining corpora. * Indicates that the corpus is a subset of another corpus analysed in this study.

	Opinion Bearing	**Non Opinion Bearing**	**Mixed**
Formal writing/ News	WSJOp	Reuters NYT WSJNop	BNC MPQA WSJMix
Blogs	BlogOp OP3* OP5*	BlogNop NOP3* NOP5*	BlogMix
Webpages	MROp CRD	MRNop	

The corpora analysed in this study were divided into three categories: (1) Formal writing/News, (2) Blogs and (3) Webpages, with further distinctions between opinion-bearing (documents expressing an opinion), non-opinion-bearing (documents that do not express an opinion) and mixed (documents that have not been assessed). The corpora are listed in Table 1. Categories for the Wall Street Journal corpora were made using assumptions used by in the original research (Yu & Hatzivassiloglou 2003, Kim & Hovy 2005), which have not been verified by human assessors. Assumptions were made (by this researcher) for the remainder of the Formal writing/News corpora to enable them to be placed into categories. The categories for the Movie Review corpora were based on assumptions made by Pang & Lee (2004).

2.1 British National Corpus

The British National Corpus (BNC) was included in this study as a standard corpus for comparison purposes, and as a reference list of standard English words (Gillam & Ahmad 2005). BNC is made

up of written and spoken English (*What is the BNC?* 2007), and was collected over several years (1991–1994) with no new texts being added since completion. However, revisions were made in 2001 and 2007. BNC provides a general language corpus for this study. This collection is not divided into opinion-bearing and non-opinion-bearing. 4,050 documents contain a total number of 470,821 word types and the token count is 96,353,012. The mean length of each document is 23,797 tokens.

2.2 Reuters

Reuters was included in this study as an example of news articles. Reuters is a collection of 7,190 news articles dating between August 1996 and October 1996. As the articles are reporting news events, they are assumed to be non-opinion-bearing in this study. This collection contains 43,963 word types and 1,565,380 tokens. The mean length of each document is 218 tokens.

2.3 New York Times

The New York Times (NYT) corpus is a subset of the AQUAINT[2] document collection. It was included in this study as a further example of news articles. The articles range in date from June 1998 to September 2000, totaling 820 days. The corpus contains 314,452 news articles, totaling 830,075 word types and 231,856,086 tokens. This corpus contains news articles and has been categorised as non-opinion. The mean length of each document is 737 tokens.

2.4 Wall Street Journal

The Wall Street Journal (WSJ) corpus is a subset of the TIPSTER[2] document collection. The news articles have been collected in the years 1987 to 1992. Some of these articles have headings indicating which category the article originates from (*Editorial, Letter to editor, Business* or *News*). These categories were used to divide the corpus into opinion-bearing (Editorial and Letter to editor) and non-opinion-bearing (Business and News) (Yu & Hatzivassiloglou 2003, Kim & Hovy 2005). This corpus is divided into three subset corpora:

- WSJOp – 4,190 articles, 47,939 word types, 1,364,326 tokens, document mean length: 326 tokens
- WSJNop – 19,731 articles, 58,509 word types, 4,625,526 tokens, document mean length: 234 tokens
- WSJMix[3] – 29,324 articles, 288,242 word types, 60,402,701 tokens, document mean length: 2,060

The WSJ collections were included in this study to allow the comparison of opinion-bearing and non-opinion-bearing news articles to opinion-bearing and non-opinion-bearing blogs. Opinion identification research (Yu & Hatzivassiloglou 2003, Kim & Hovy 2005) used WSJ articles to develop training and testing data. A Blog06 participant (Eguchi & Shah 2006) used this word list for training data in the opinion identification task.

[2]http://www.ldc.upenn.edu/

[3]The remainder of the documents where the title did not indicate into which category the document should be placed.

2.5 MPQA

The MPQA corpus contains 535 news articles collected from various sources (*MPQA Releases* 2007, Wiebe 2002). It contains 6,867 word types, and totals 50,502 tokens. A Blog06 participant (Eguchi & Shah 2006) used this corpus for training data in the opinion identification task. The mean length of each document is 94 tokens.

2.6 Blog06

100,649 blogs were crawled over an eleven week period, made up of 70,701 'top blogs'[4], 17,969 splogs and 11,979 'other blogs'[5]. The resulting corpus contains 3,215,171 blog documents (MacDonald & Ounis 2006).

Judgements by human assessors Ounis et al. (2006) on 67,382 documents placed them into one of five categories (detailed in Table 2). These assessed documents were divided into three document collections for this study:

1. BlogOp – 10,446 documents, 404,131 word types, 28,713,436 tokens, blog mean length: 2,749 tokens
2. BlogNop – 8,281 documents, 338,895 word types, 19,438,021 tokens, blog mean length: 2,347 tokens
3. BlogMix[6] – 42,663 documents, 866,570 word types, 105,824,131 tokens, blog mean length: 2,480 tokens

Table 2: Number of documents allocated, by NIST assessors, to each assessment category (Ounis et al. 2006)

Relevance Scale	Label	No. of Documents
Not Judged	-1	0
Not Relevant	0	47,491
Adhoc-Relevant	1	8,361
Negative Opinion	2	3,707
Mixed Opinion	3	3,664
Positive Opinion	4	4,159
(Total)	-	67,382

2.7 BlogOpSent and BlogNopSent

The NIST assessments on the blog documents that place them into 'opinion-bearing' and 'non-opinion-bearing' categories were done at the document level, meaning that assessed documents contained text relevant to the topic (Op/Nop) and text not relevant to the topic (Op/Nop). This results in the characteristics of BlogOp and BlogNop being very similar. To enable analysis and reporting on the differences between these two corpora, they were divided into subsets by removing the text not relevant to the

[4]Top blogs were selected by Nielsen BuzzMetrics and the University of Amsterdam (Ounis et al. 2006)

[5]Blogs from a mix of genre

[6]Blogs judged as 'not relevant' were blogs retrieved in the information retrieval process and judged as not being relevant to the query topic by the NIST human assessor

given topic.

The documents within each corpus (BlogOp and BlogNop) were divided into single sentences and indexed using the Lucene Search Engine[7]. Lucene was used to create a list of individual sentences relevant to the given topic and these sentences were used as the first sentence in a block of text extracted from each blog. The blocks size was three sentences and five sentences. Full details of the methodology applied is explained in Section 3.1.

The resulting subsets are entitled:

- OP3 – 58,277 sentences, 72,018 word types, 1,867,411 tokens, sentence mean length: 32
- OP5 – 86,869 sentences, 83,231 word types, 2,305,358 tokens, sentence mean length: 26
- NOP3 – 45,821 sentences, 65,025 word types, 1,354,630 tokens, sentence mean length: 29
- NOP5 – 66,534 sentences, 72,584 word types, 1,615,991 tokens, sentence mean length: 24

2.8 Movie Review Data

This corpus[8] is made up of 5,000 subjective (MROp 13,765 word types, 100,136 tokens and sentence mean length: 20 tokens) and 5,000 objective (MRNop 14,325 word types, 110,283 tokens and sentence mean length: 22 tokens) sentences (Pang & Lee 2004). Two websites were used as the source for these sentences:

- Rotten Tomatoes[9] – these were assumed to be subjective (Pang & Lee 2004)
- Internet Movie Database[10] – these were assumed to be objective (Pang & Lee 2004)

A Blog06 participant (Yang, Yu, Valerio & Zhang 2006) used this corpus for training data in the opinion identification task.

2.9 Customer Review Data (CRD)

This corpus contains customer reviews on digital cameras, cellular phones, mp3 players and dvd players, which were collected from amazon.com, and annotated by Minqing Hu and Bing Liu[11]. There is 5,015 word types and 59,317 tokens in this corpus. The mean length of the 4,256 sentences is 14 tokens. A Blog06 participant (Yang, Yu, Valerio & Zhang 2006) used this corpus for training data in the opinion identification task.

3 Methodology

This section discusses the methodology applied in the analysis of key differences between corpora originating from different sources and the results are reported in the Results and Discussion section (Section 4). Two features of the corpora were analysed: (1) Part-of-speech (POS) Proportions and (2) Unique/Weird Words and Slang. All corpora were analysed using the above-mentioned methods and reported in results section, excluding the blog sentence corpora. The analysis on OP3, OP5, NOP3 and NOP5 is reported separately in Section 4.3.

[7] http://lucene.apache.org/java/docs/
[8] http://www.cs.cornell.edu/people/pabo/movie-review-data
[9] http://www.rottentomatoes.com/
[10] http://www.imdb.com
[11] http://www.cs.uic.edu/ liub/FBS/FBS.html

Table 3: Total word types, tokens and mean document length in the corpora analysed in this study. *Mean length is at the sentence level. ♢ Indicates that the corpus is a subset of another corpus.

Corpus	Word Types	Tokens	Mean Document Length
BNC	470,821	96,353,012	23,797
Reuters	43,963	1,565,380	218
NYT	830,075	231,856,088	737
WSJOp	47,939	1,364,326	326
WSJNop	58,509	4,625,526	234
WSJMix	288,242	60,402,701	2,060
MPQA	6,867	50,502	94
BlogOp	404,131	28,713,436	2,749
BlogNop	338,895	19,438,021	2,347
BlogMix	866,570	105,824,131	2,480
BlogOp			
OP3♢	72,018	1,867,411	32*
OP5♢	83,231	2,305,358	26*
BlogNop			
NOP3♢	65,025	1,354,630	29*
NOP5♢	72,581	1,615,991	24*
MROp	13,765	100,136	20*
MRNop	14,325	110,283	22*
CRD	5,015	59,317	14*

3.1 Removing Non-Relevant Text from the Blog Corpora

Due to the similarity between BlogOp and BlogNop, further analysis was done on these corpora. The content of blogs contained four types of text: (1) Opinion-bearing – off topic, (2) Non-opinion-bearing – off topic, (3) Opinion-bearing – on topic (BlogOp only[12]), and (4) Non-opinion-bearing – on topic.

The text within each document in the BlogOp and BlogNop respectively, was separated into single sentence blocks and indexed using The Lucene Search engine. The sentences relevant to the given topic were retrieved and placed into a list.

The structure of the text within the blogs led to some relevant text not being retrieved by the Lucene Search engine. An example of this was for the given topic 'March of the Penguins', where the title of the documentary was in one sentence and the opinion on the documentary (not mentioning the query term) was in the next sentence.

To reduce the impact of this, the text within each relevant sentence was retrieved, along with the following two/four[13] sentences. Sentences were included once only in the resulting subset of text, any duplications were removed prior to the collation of the text. The analysis methods described in the remainder

[12] The assumption was made that all relevant text within the BlogNop corpus was non-opinion-bearing.
[13] Depending on the sentence block size.

of this section were applied to these four subsets of text, similarly to the other corpora listed in Section 2.

3.2 Part-of-Speech Proportions

An area of interest is whether one type of corpus has a higher proportion of a particular POS type. Three part-of-speech taggers were tested for speed and robustness by Johnson, Malhotra & Vamplew (2006): *The Stanford NLP Group Loglinear Part-Of-Speech Tagger* (2006), *The MontyLingua natural language package* (2006) and *QTag probabilistic parts-of-speech tagger* (2006). QTag was found to be the fastest and most robust of these three (Johnson, Malhotra & Vamplew 2006).

Each corpus was tagged using QTag and the proportions of the following categories[14] were summarised:

- Adjectives – general, comparative and superlative
- Nouns – common singular, common plural, proper singular and proper plural
- Pronouns – indefinite, personal, possessive (my, his), reflexive, 'wh-' (who, that) and possessive (whose)
- Adverbs – general, comparative and superlative
- Verbs – base, past tense, '-ing' (believing), past participle and '-s' (believes)
- Unclassified – words that QTag could not classify

The proportions were used as a vector and the similarity between each vector calculated using the following formula, where v is the vector, p is the position within the vector and n is the vector length. The 'norm' of the vector is calculated:

$$\parallel v1 \parallel = \sqrt{\sum_{k=1}^{n} p_k^2}$$

and the similarity $(\widehat{v1} \cdot \widehat{v2})$ is calculated:

$$\widehat{v1} \cdot \widehat{v2} = \frac{\sum_{i=1}^{n}(v1_i \cdot v2_i)}{\parallel v1 \parallel \cdot \parallel v2 \parallel}$$

The results for each corpus are detailed and discussed in the results section (4.1).

3.3 Unique/Weird Words and Slang

Another area of interest is whether blogs use a higher proportion of unique or weird words and slang. More than half of bloggers are under the age of 30 with an even split of men and women (Lenhart & Fox 2006). Bloggers form communities of common interest and link to other members of the community. These communities often create an language specific to their particular interests.

The SC reference collection of words used in the spell checking section of this research includes a wide range of words, including American, English and Canadian spelling and jargon. The list was compiled from various sources[15], on the World Wide Web. BNC is used as a reference collection for general English language when calculating weirdness values in this study, as has been done in other research (Gillam & Ahmad 2005).

3.3.1 Spell Checking

The words within each corpus were compared to the SC reference list (described above) of English words to extract uncommon words. These words were placed into a list of 'non-standard' words, they could be words that are specific to a particular community, slang or simply misspelt. The proportion of uncommon words were compared to determine whether a particular corpus is more likely to contain 'non-standard' English words.

3.3.2 Weirdness Values

'Weird' words are either not found in the reference list of words or rarely appear. Words with high frequency and weirdness values are considered high in domain specificity (Gillam & Ahmad 2005). The weirdness values were calculated for each term within each corpus, using the following formula (Gillam & Ahmad 2005), and the results are discussed in section 4.2.

$$weirdness = \frac{N_{GL} f_{SL}}{(1 + f_{GL}) N_{SL}} (Gillam\&Ahmad(2005))$$

where f_{SL} is the word frequency in the corpus, f_{GL} is the word frequency in the reference list and N_{SL} and N_{GL} are the total number of tokens in the corpus and reference list respectively.

4 Results and Discussion

The thirteen corpora analysed in this study were divided into one of three general categories and eight sub-categories:

- Formal writing/News
  - Opinion-Bearing – WSJOp
  - Non-Opinion-Bearing – WSJNop, Reuters, NYT
  - Mixed – WSJMix, BNC, MPQA
- Blogs
  - Opinion-Bearing – BlogOp
  - Non-Opinion-Bearing – BlogNop
  - Mixed – BlogMix
- Webpages
  - Opinion-Bearing – MROp, CRD
  - Non-Opinion-Bearing – MRNop

The indicators analysed in the study show a high level of similarity between the BlogOp and BlogNop corpora. However, the indicators show the BlogMix is different in many areas (detailed throughout this section). This is partly due to the existence of

[14] The Yu & Hatzivassoglou (2003) list comprised of adjectives, nouns, adverbs and verbs. Pronouns has been added to these categories for this research.

[15] http://wordlist.sourceforge.net/, http://www.mieliestronk.com/worklist.html, http://www.outpost9.com/files/WordList.html

Table 4: Mean proportion of part-of-speech categories in the corpora categories

Part-of-speech	Formal			Blog			Webpages		
Category	op	nop	mix	op	nop	mix	op	nop	mix
Adjectives	8.4	7.7	7.9	6.8	7.0	7.6	10.0	8.8	-
Nouns	31.2	34.1	31.3	31.2	32.1	40.1	27.5	32.7	-
Pronouns	4.2	2.9	4.1	5.5	4.9	4.2	5.4	6.6	-
Adverbs	3.5	2.5	3.4	4.4	4.3	3.6	5.6	3.2	-
Verbs	8.5	9.5	9.5	9.9	9.9	8.8	9.1	9.8	-

spam blogs within this corpus. These blogs contain repeated text that artificially inflates the various characteristics. This section discusses the POS proportions and Unique/Wierd words used within each individual corpus, and the characteristics found in the blog sentence corpora are discussed at the end of this section.

4.1 Part-of-Speech Proportions

The QTag part-of-speech tagger was used to tag the content of corpora analysed in this study. The sum of each part-of-speech tag was compared to determine similarities and differences between the various types of corpus. The mean of the proportions for each category (detailed above) was calculated with the following results (detailed in table 4):

- Adjectives – Opinion-bearing webpages recorded the highest mean proportion (10.0%), followed by Non-opinion-bearing webpages (8.8%). The lowest mean proportion recorded was Opinion-bearing blogs (6.8%) and Non-opinion-bearing blogs (7.0%).
- Nouns – Mixed blogs recorded the highest mean proportion (40.1%), followed by Non-opinion-bearing formal writing/news. The lowest mean proportion recorded was Opinion-bearing webpages (27.5%) and Non-opinion-bearing blogs (32.1%).
- Pronouns – Non-opinion-bearing webpages recorded the highest mean proportion (6.6%), followed by Opinion-bearing blogs (5.5%). The lowest mean proportion recorded was Non-opinion-bearing formal writing/news (2.9%) and Mixed formal writing/news (4.1%).
- Adverbs – Opinion-bearing webpages recorded the highest mean proportion (5.6%), followed by Opinion-bearing blogs (4.4%). The lowest mean proportion recorded was Non-opinion-bearing formal writing/news (2.5%) and Non-opinion-bearing webpages (3.2%).
- Verbs – Opinion-bearing and Non-opinion-bearing blogs recorded the highest mean proportion (9.9%) with the lowest mean proportion being recorded by Opinion-bearing formal writing/news (8.5%) and Mixed blogs (8.8%).
- Unclassified – Of the 13 corpora analysed in this study all recorded 0.1% of words that could not be classified, except MPQA and MRNop which recorded 0.0%.

The POS proportions for each corpus was entered into a part-of-speech vector, which was used to calculate a similarity score between the various corpora. When determining similarities between different types of text documents, it is interesting to note that of the individual corpora, BlogOp and BlogNop show very little difference between the POS proportions (0.9997 where 1.0 is exactly the same), while CRD and Reuters show the highest level of difference (0.9307).

When the mean proportions for each corpus category (detailed at the start of this section) are compared, the least similar categories are Blogs Mixed and Webpages Opinion-bearing (0.9386), followed by Webpages Opinion-bearing and Formal writing/news Non-opinion-bearing (0.9507). The most similar is once again Blogs Opinion and Non-opinion (0.9997), followed by Formal writing/new Opinion and Mixed (0.9967). The mean similarity scores are detailed in Table 5.

As the POS proportions do not indicate a pattern over the various types of corpora, each corpus was analysed at an individual word level.

4.2 Unique/Weird Words and Slang

Two collections of words were used as reference collections for this analysis: (1) A collection of words including American, Canadian and English spelling and slang that was compiled from various sources on the World Wide Web[15] (SC reference list) and (2) BNC is used as the reference collection when calculating 'weirdness' (Gillam & Ahmad 2005) scores for the various corpora.

4.2.1 Spell Checking

The word types in each corpus were compared to the SC reference list, to create a list of 'non-standard' words. The proportion of word types appearing in each corpus that do not appear in the reference list is detailed in table 6. The table shows the percentage of word types not in the reference list, the percentage of tokens (word frequency) that the previous figure represents within each corpus and the percentage of those tokens that have a frequency of one within each corpus.

The blog corpora recorded the highest percentage of word types not appearing in the reference list (BlogOp 65%, BlogNop 63%, BlogMix 63%), which represents 4% (BlogMix 6%) of the tokens within each corpus, indicating that blogs use a higher proportion of unique/slang words compared to the other corpora analysed in this study. The low percentage (16%, 5%) of single frequency terms appearing in the blog corpora and not in the SC reference list indicates that unique/slang words are less likely to be 'one-off' uses compared to the other corpora. BlogMix recorded a higher proportion of tokens that were not found in the SC reference list, only 5% of

Table 5: The mean similarity between the part-of-speech vectors. Vectors that are exactly the same would score 1.0000

Corpus		Formal writing/News			Blogs			Webpages	
		Op	Nop	Mix	Op	Nop	Mix	Op	Nop
Formal writing/ News	Op		0.9942	**0.9967**	0.9956	0.9961	0.9831	0.9598	0.9667
	Nop			0.9950	0.9922	0.9936	0.9893	**0.9507**	0.9651
	Mix				0.9937	0.9934	0.9816	0.9718	0.9803
Blogs	Op					**0.9997**	0.9873	0.9563	0.9647
	Nop						0.9898	0.9529	0.9621
	Mix							**0.9386**	0.9566
Webpages	Op								0.9884
	Nop								

the of these tokens are single frequency tokens. The low proportion of single frequency tokens indicates the multiple use of unique words. This is partly due to the inclusion of spam which repeats text multiple times in the same document. To alleviate the problems found in the BlogMix corpus, spam and other repeating text will need to be removed.

Examples of the words and frequency within corpora not found in the reference list are listed in table 7. While many of the words appearing in the corpus and not appearing in the SC reference list are variations of spelling (Eg: 'heeelllooo') or words that run together (Eg: 'ofconservingcanada') other possible explanations for some of the words are listed below:

- abramoff – Jack Abramoff is a former American political lobbyist
- quicklink – QuickLink allows users to manage a set of words for which they would like links to be automatically generated[16]
- usefulrate – a term used when rating something on the Web
- alito – to overcome large amounts of adversity with ease[17]
- zinat – the name of a movie about a woman named Zinat
- korinna – ancient Greek poet or current model

4.2.2 Weirdness Values

Each corpus (other than BNC) was compared to the BNC reference collection and 'weirdness' scores were calculated for each token within the corpus. The tokens of most interest have high frequency and 'weirdness' values. Table 8 details the proportion of word types found in each corpus that are not found in the BNC reference collection. The corpora with less than 1,000,000 tokens have a low proportion of tokens ($< 10\%$) not found in the BNC reference list, with the proportion growing larger as the corpus size increases.

Word types of particular interest have a high score in frequency and weirdness. Table 9 includes an example of some of the word types with high frequency and weirdness, selected from the corpora scoring the highest in weirdness.

Table 6: Percentage of word types and tokens appearing in the corpus and not appearing in SC reference list. The number of single frequency tokens, and the percentage they represent, appearing in the corpus and not appearing in the SC reference is shown in the third and fourth columns respectively.

Corpus	% types not in ref coll	% tokens not in ref coll	Num single freq tokens	% tokens with single freq.
BNC	40	1	104,737	12
Reuters	24	2	5,327	15
NYT	35	1	122,747	4
MPQA	3	1	125	27
WSJOp	10	1	3,664	40
WSJNop	21	1	6,021	10
WSJMix	33	1	47,724	7
BlogOp	65	4	171,706	16
BlogNop	63	4	134,397	16
BlogMix	63	6	277,195	5
MROp	6	1	574	49
MRNop	8	1	874	53
CRD	7	1	267	36

One problem that becomes evident when looking at word types with high frequency and high weirdness values, is the use of hyphenated words within the corpora (Eg. NYT – star-telegram and WSJMix – year-earlier). This problem is not only evident at the top end of the weirdness list (high frequency/high weirdness), it is spread throughout the list with a high concentration of single frequency word types.

In the small sample shown in table 9, there are words that could be included in a stop-word list (Eg. blog, trackback, nyt). However, simply adding high frequency/high weirdness words to the stop-word list would remove words such as 'lewinsky' and 'netflix', which is problematic as both of these words are possible query terms.

[16] http://www.majordojo.com/projects/QuickLink/
[17] http://www.urbandictionary.com/define.php?term=alito

Table 7: Examples of word types appearing in each corpus, and not appearing in the reference list. Table shows the word frequency within the corpus and the proportion of the word within the total tokens in the corpus that were not found in the reference list.

Corpus	Word Type	Corpus Freq.	% of Tokens
BlogOp	abramoff	7221	0.68
	quicklink	3746	0.35
	usefulrate	3135	0.29
	janeane	2552	0.24
	ofconservingcanada	1	0.00
	zweng	1	0.00
BlogNop	usefulrate	3643	0.42
	engadget	3010	0.35
	alito	1598	0.19
	myyahoooorbloglines	1562	0.18
	heeelllooo	1	0.00
	zinat	1	0.00
BlogMix			
	phentermine	85133	1.44
	spyware	76162	1.29
	holdem	60179	1.02
	zzzzzzzzippy	1	0.00
	korinna	2	0.00

Table 8: Word types not appearing in BNC reference corpus, detailing the total number of word types and tokens in each corpus, the number of word types found in the corpus and not found in the BNC reference corpus and the proportion of word types in the corpus that it represents. *The NYT proportion of types not in BNC is high due to the use of 'American' spelling within the corpus.

Corpus	Types	Tokens	Num types not in BNC	% types not in BNC
Reuters	43,963	1,565,380	9,663	22
NYT	830,075	231,856,086	641,395	77*
MPQA	6,867	50,502	213	3
WSJOp	47,939	1,364,326	7,666	16
WSJNop	58,509	4,625,526	15,338	26
WSJMix	288,242	60,402,701	154,727	54
BlogOp	404,131	28,713,436	292,365	72
BlogNop	338,895	19,438,021	236,700	70
BlogMix	866,570	105,824,131	695,515	80
MROp	13,765	100,136	1,150	8
MRNop	14,326	110,283	1,143	8
CRD	5,015	59,317	399	8

4.3 Blog Sentence Corpora

The blog sentence corpora were analysed using the methodology described in Section 3 for POS proportions and Unique words/slang. The OP3 and OP5 were found to be very similar in all characteristics, with the exception that OP5 was a larger corpus, the same was found for NOP3 and NOP5.

The total number of tokens in each corpus is OP3 1,867,411, OP5 2,305,358 (OP5 comprises of 23% more tokens), and NOP3 1,354,630 and NOP5 1,615,991 (NOP5 comprises of 19% more tokens). It may be expected that the difference between a corpus comprised of three sentence blocks compared to five sentence blocks would be approximately 167%[18]. The OP5 corpus comprises of 115% of the word types in to OP3, while NOP5 corpus comprises of 158% of the word types in NOP3.

Table 10 details the results of the analysis of OP3, OP5, NOP3 and NOP5 and compares them to the original blog corpora results. The figures are similar in all categories for the blog sentence except the percentage of word types found in NOP5 and not found in the SC reference list. The proportion is lower than the other corpora proportions, however this is due to there being more word types in NOP5 compared to NOP3.

The distance between the blog opinion and non-opinion corpora in the POS proportions increased in the 'pronouns', 'adverbs', 'adjectives' and 'verbs', while the distance did not change in the 'nouns' category. The opinion-bearing Yu & Hatzivassiloglou (2003) word list created from documents from within the Wall Street Journal corpus, did not include 'pronouns' as a category.

[18] 167% is calculated $\frac{5}{3}$.

The proportion of word types in the various blog corpora and not found in the SC reference list reduced substantially in the blog sentence corpora, while the proportion of tokens found in the corpus and not found in the SC reference list only changed slightly. The proportion of word types with frequency one found in each corpus and not found in the SC reference list doubled (approximately). The weirdness score also decreased dramatically. Along with the proportions changing, the distance between BlogOP and BlogNOP compared to the distance between OP3 and NOP3 increased substantially in these categories with the exception of the weirdness score which did not record a change in the distance between the opinion/non-opinion corpora. Table 10 details the percentage change to the various characteristics measured in this study.

In general the variance between the opinion and non-opinion blog corpora increased, however the difference between the corpus having three sentence blocks and five sentence blocks was not shown in these results. The remainder of this section compares the characteristics of the OP3 and NOP3 to the corpora assessed as being either Opinion or Non-opinion in this research[19]

OP3 is most similar to the WSJOp corpus in the POS proportions (0.993) with the similarity scores being 0.934 (MROp) and 0.942 (CRD) for the other opinion corpora. A major difference between the OP3 corpus and the other opinion corpora is within the 'adjective' category. The MROp corpus recorded the highest proportion of adjectives (11.9%) with CRD (8.1%) and WSJOp (8.4%) both recording a higher proportion compared to OP3 (6.5%). The 'verbs' category also showed a large variation between OP3

[19] Reuters and NYT corpora are not included in this analysis as they have not been assessed as being non-opinion, instead they were assumed to be non-opinion in this study.

Table 9: Word types with high frequency and weirdness

Corpus	Word Type	Corpus Freq.	BNC Freq.	Weirdness
NYT	nyt	111,517	1	23,171
	nytimes	27,403	0	11,387
	coxnet	26,387	0	10,965
	star-telegram	24,938	0	10,363
	lewinsky	24,718	0	10,272
Blog Op	blog	30,088	0	100,965
	trackback	16,083	0	53,969
	permalink	10,417	0	34,956
	netflix	6,437	0	21,600
	google	5,498	0	18,449
Blog Nop	blog	22,015	0	109,126
	permalink	12,615	0	62,531
	google	4,868	0	24,130
	usefulrate	3,643	0	18,058
	url	3,498	0	17,339
Blog Mix	blog	147269	0	134088
	phentermine	82953	0	75528
	spyware	76162	0	69345
	holdem	60179	0	54793
	permalink	50474	0	45956
WSJ Mix	totaling	3637	0	5801
	calif	14251	4	4546
	year-earlier	8272	2	4398
	totaled	5381	1	4291
	bankruptcy-law	2218	0	3538

and the remaining corpora. OP3 recorded 10.3% which is higher than MROp (8.4%), CRD (9.8%) and WSJOp (8.5%).

The proportion of word types in the SC reference list and not in OP3 (36%) is much higher than the other opinion corpora (MROp 6%, CRD 7% & WSJOp 10%), this represents 3% of the tokens within OP3 corpus and 1% of the remaining corpora. The proportion of these word types with a frequency of one ranges between 34–49%, with OP3 recording the lowest proportion.

OP3 recorded the highest level of weirdness (39%), with the remaining corpora recording 16% (WSJOp) and 8% for MROp and CRD. This indicates a high level of domain specific word types being used within the OP3 corpus. Table 11 details the results of the analysis using POS proportions, Spell checking and Weirdness.

When analysing the non-opinion corpora, it was found that once again the NOP3 corpus was more similar to the WSJNOP corpus (0.991) compared to the MRNop corpus (0.955). Similar to the opinion corpora, the proportion of adjectives in the NOP3 corpus (7.1%) was lower than the other corpora (MRNop 8.8% & WSJNop 7.7%), however the proportion of verbs was similar across the non-opinion corpora. The proportion of adverbs was higher in the NOP3 corpus compared to the remaining corpora (MRNOP 3.2% & WSJNop 2%).

The proportion of 49% of word types found in NOP3 and not in the SC reference collection was much higher compared to MRNop (8%) and WSJNop (21%). The proportion of tokens found in the non-opinion corpora and not found in the SC reference list was slightly higher in NOP3 (4%) compared to MRNop (2%) and WSJNop (1%). MRNop recorded the highest proportion of word types with a frequency one (53%) compared to NOP3 (33%) and WSJNop (10%).

As was found in the opinion corpora, NOP3 was much higher in weirdness (38%) compared to WSJNop (26%) and MRNop (8%). This reinforces the belief that blogs are more likely to contain domain specific word types compared to other corpora. Table 11 details the results of the analysis using POS proportions, Spell checking and Weirdness.

5 Conclusions and Future Work

The indicators analysed in this study reveal the opinion and non-opinion blog corpora to be different in their characteristics to the other corpora analysed in this study, especially in the use of non-standard words where a high proportion of words used in the blog documents were not found in standard English word lists. Contrasting this is the corpora collected from sources other than blogs that recorded a much lower proportion of non-standard words. It is expected that opinion identification research training data collected from outside the Blogosphere will not contain the same high level of non-standard words, making it unlikely to produce accurate results when attempting to identify opinions within the blogosphere.

Comparing the results of analysis for the opinion and non-opinion blog corpora indicates very little variation between the two corpora. The similarity of the POS proportions in the two corpora was close to the perfect score (for corpora that is exactly the same). While the distance between the remaining indicators was not substantial. The similarity between the opinion and non-opinion blog corpora is not mirrored by the opinion and non-opinion Movie Review corpora where there was greater distance between the various indicators analysed. One major difference between the blog and Movie Review corpora is that the Movie Review corpora contains opinion or non-opinion sentences. The non-opinion text has been removed from the opinion documents in the MROp corpus.

To determine if a case exists for separating the blog corpora into sentence blocks, the opinion and non-opinion blog corpora were divided into subsets comprising sentences relevant to their given topic. These corpora were compared to the original blog corpora to determine whether the distance between the opinion and non-opinion corpora increased. The distance between the opinion and non-opinion blog sentence corpora generally increased compared to the distance between the original opinion/non-opinion blog corpora. This was particularly evident in the Spell categories which recorded large distance increases in the percentage of word types and the percentage of tokens found in the blogs and not in the SC reference list. This, coupled with the distance increases in the POS categories will lead to more detailed research at the 'word' level in future research.

Comparing the OP3 corpus to other opinion corpora in this study (MROp, CRD & WSJOp)

Table 10: Characteristics of various blog corpora. The 'Variance Change' column indicates the % change in the distance between BlogOP/BlogNOP and OP3/NOP3 in the indicators within this study. *indicates mean document length.

	OP3	OP5	BlogOP	BlogNOP	NOP3	NOP5	Variance Change %
Word types	72,018	83,231	404,131	338,895	65,025	72,581	
Tokens	1,867,411	2,305,358	28,713,436	19,438,021	1,354,630	1,615,991	
Mean Sentence Length	32	26	2,749*	2,347*	29	24	
% Adjectives	6.5	6.5	6.8	7.0	7.1	7.1	6
% Nouns	32.7	32.3	31.2	32.1	33.6	33.6	0
% Pronouns	5.3	5.3	5.5	4.9	4.3	4.3	8
% Adverbs	4.1	4.2	4.4	4.3	3.8	3.8	5
% Verbs	10.3	10.2	9.9	9.9	10.0	10.0	3
Spell							
% Types	36	39	65	63	49	37	33
% Tokens	3	3	4	4	4	4	33
% Single Freq. Word Types	34	33	16	16	33	33	3
Weirdness	39	43	72	70	38	40	0

showed that OP3 was most similar to WSJOp in POS proportions, whilst the Spell and Weirdness categories recorded a large variation between OP3 and the remaining opinion corpora. Of the three opinion corpora in this section of the study, WSJOp content has been assumed to be opinion-bearing text in other research (Yu & Hatzivassiloglou 2003, Kim & Hovy 2005) at the document level, whilst MROp and CRD were assessed at the sentence level. These similarities and differences are repeated in the comparison of NOP3 to the non-opinion corpora in this study (MRNop & WSJNop).

There was however, a lack of variation[20] between the blog sentence corpora with three sentence blocks and five sentence blocks. Whether there is a difference when using the corpora as training data will be determined in future research into opinion identification with blog documents.

Blogs contain a high level of 'non-standard' word types when comparing them to a reference list of either standard English words (BNC) or an expanded list of words containing various spellings of words (English, American, Canadian, etc.), proper names and abbreviations, with a low percentage of these being a singular use of the word. This dramatic variation indicates that blogs use a higher proportion of specific words, demonstrating a substantial variation in the words used within blogs compared to other corpora, and that training data for blog opinion identification should not be extracted from the other corpora.

When asking the question *'Does identifying opinions within blog posts and comments require different training data to identifying opinions within more traditional corpora?'*, it is clear that there is no simple approach to dealing with Blogs. The difference between opinion-bearing and non-opinion-bearing blog documents is not great enough to warrant using blogs assessed at the document level. It cannot be assumed that an entire blog document will contain opinion-bearing words as has been assumed in other research (Yu & Hatzivassiloglou 2003, Kim & Hovy 2005). The high level of 'non-standard' words found within blogs indicates that specific blog training data is needed when attempting to identify opinions within blogs.

[20]Excluding the percentage of word types found in NOP3 and NOP5 and not in the SC reference list.

References

About Technorati, Accessed September (2007). Available: http://www.technorati.com/about/.

Eguchi, K. & Shah, C. (2006), Opinion retrieval experiments using generative models: Experiments for the trec 2006 blog track, *in* E. M. Voorhees & L. P. Buckland, eds, 'The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)', Gaithersburg, Maryland.

Gillam, L. & Ahmad, K. (2005), Pattern mining across domain-specific text collections., *in* P. Perner & A. Imiya, eds, 'Machine Learning and Data Mining in Pattern Recognition', pp. 570–579.

Johnson, D., Malhotra, V. & Vamplew, P. (2006), 'More effective web search using bigrams and trigrams', *Webology* **3**(4).

Kim, S.-M. & Hovy, E. (2005), Automatic detection of opinion bearing words and sentences, *in* 'Natural Language Processing - IJCNLP 2005', Springer, New York.

Lenhart, A. & Fox, S. (2006), Bloggers: A portrait of the internet's new storytellers, Technical report, Pew Internet & American Life Project.

MacDonald, C. & Ounis, I. (2006), 'The trec blogs06 collection : Creating and analysing a blog test collection', *DCS Technical Report Series* p. 8.

MPQA Releases (2007). Available: http://www.cs.pitt.edu/mpqa/.

Table 11: Characteristics of Opinion and Non-opinion-bearing corpora. Note: Similarity scores closest to 1.000 are the most similar.

Category	OP3	MROp	CRD	WSJOp	NOP3	MRNop	WSJNop
Adjectives	6.5	11.9	8.1	8.4	7.1	8.8	7.7
Nouns	32.7	28.7	26.3	31.2	33.6	32.7	35.1
Pronouns	5.3	4.2	6.7	4.2	4.3	6.6	2.1
Adverbs	4.1	5.7	5.4	3.5	3.8	3.2	2
Verbs	10.3	8.4	9.8	8.5	10	9.8	10
Similarity		0.934	0.942	0.993		0.955	0.991
Spell							
Types	36	6	7	10	49	8	21
Tokens	3	1	1	1	4	2	1
Single Freq.	34	49	36	40	33	53	10
Weirdness	39	8	8	16	38	8	26

Nardi, B. A., Schiano, D. J., Gumbrecht, M. & Swartz, L. (2004), 'Why we blog', *Commun. ACM* **47**(12), 41–46.

Ounis, I., de Rijke, M., Macdonald, C., Mishne, G. & Soboroff, I. (2006), Overview of the trec-2006 blog track, *in* E. M. Voorhees & L. P. Buckland, eds, 'The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)', Gaithersburg, Maryland.

Pang, B. & Lee, L. (2004), A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, *in* 'Proceeding of the ACL'.

QTag probabilistic parts-of-speech tagger (2006). http://www.english.bham.ac.uk/staff/ omason/software/qtag.html.

Rosenbloom, A. (2004), 'Introduction: The blogosphere', *Commun. ACM* **47**(12), 30–33.

The MontyLingua natural language package (2006). http://web.media.mit.edu/ hugo/ montylingua/index.html.

The Stanford NLP Group Loglinear Part-Of-Speech Tagger (2006). http://nlp.stanford.edu/ software/tagger.shtml.

What is the BNC? (2007). Available: http://www.natcorp.ox.ac.uk/corpus/index.xml.

Wiebe, J. (2002), Instructions for annotating opinions in newspaper articles, Technical Report TR-02-101, Department of Computer Science, University of Pittsburgh, Pittsburgh, PA.

Yang, H., Si, L. & Callan, J. (2006), Knowledge transfer and opinion detection in the trec2006 blog track, *in* E. M. Voorhees & L. P. Buckland, eds, 'The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)', Gaithersburg, Maryland.

Yang, K., Yu, N., Valerio, A. & Zhang, H. (2006), Widit in trec-2006 blog track, *in* E. M. Voorhees & L. P. Buckland, eds, 'The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)', Gaithersburg, Maryland.

Yu, H. & Hatzivassiloglou, V. (2003), Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences, *in* 'Proceedings of the 2003 conference on Empirical methods in natural language processing', Association for Computational Linguistics, Morristown, NJ, USA, pp. 129–136.

Zhang, E. & Zhang, Y. (2006), Ucsc on trec 2006 blog opinion mining, *in* E. M. Voorhees & L. P. Buckland, eds, 'The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)', Gaithersburg, Maryland.