CLASSIFYING SUSPICIOUS CONTENT USING FREQUENCY ANALYSIS

Obika Gellineau and George R.S. Weir[1]
Department of Computer and Information Sciences
University of Strathclyde, Glasgow, UK
obika.gellineau@strath.ac.uk, george.weir@cis.strath.ac.uk

ABSTRACT

This paper details an experiment to explore the use of chi by degrees of freedom (CBDF) and Log-Likelihood statistical similarity measures with single word and bigram frequencies as a means of discriminating subject content in order to classify samples of chat texts as dangerous, suspicious or innocent. The control for these comparisons was a set of manually ranked sample texts that were rated, in terms of eleven subject categories (five considered dangerous and six considered harmless). Results from this manual rating of chat text samples were then compared with the ranked lists generated using CBDF and Log-Likelihood measures, for both word and bigram frequency. This was achieved by combining currently available textual analysis tools with a newly implemented software application. Our results show that the CBDF method using word frequencies gave discrimination closest to the human rated samples.

1. INTRODUCTION

The rise of Internet-based social networking has sparked concerns about the use of such media for nefarious and criminal activities. In response, a number of authorities have proposed overt monitoring and data-mining of social networks in order to detect and identify prospective offenders. For example, "the NYPD plans to use online policing to find info about gang showdowns, murder cases, problematic house parties and other forms of commotion" (Kaiser, 2011). The intention is "to find criminals bragging about a crime they've committed or planning to commit a crime" (op.cit.). More recently, rioting in England raised questions about the role played by social networks in co-ordinating and promoting such events (BBC, 2011). As for other forms of communication, there is a perceived need for monitoring in order to differentiate the small proportion of suspicious from the majority of innocent content (cf. Weir, et al, 2011).

In this context, we have been exploring the use of software tools to analyse chat corpora with a view to classifying the content and to determine the likelihood that the interactions are 'dangerous' rather than 'innocent' in nature. Such techniques, based upon textual analysis, may apply equally well to social networking and other text-based forms of communication. The strategies that we applied use word and bigram (multi-word) frequencies as the basis for our classification. This approach has been evaluated in a series of tests using a newly developed Java software tool that works in conjunction with existing frequency analysis tools. In the following we describe our approach in detail and the results of this investigation.

2. APPROACH

The development and evaluation of the classification methods required several steps. Having identified prospective statistical measures that may serve to discriminate between topics as well as establish the chat room samples, we conducted a series of evaluations in order to compare the efficacy of these techniques in characterizing text samples from a chat corpus. Since our ultimate aim was to separate 'dangerous' from 'innocent' content, the chat corpus used as a basis for this development had to reflect appropriate richness in terms of subject matter and linguistic features. The corpora chosen were the NPS Chat corpus (Forsyth &

[1] Corresponding author.

Martell, 2007) and the IRC Chat corpus (Stevenson, 2000). Both chat corpora contain language commonly used in Internet chat rooms and both are amenable to textual analysis of their content. However, to classify these corpora into particular categories using statistical methods and textual analysis, a reference corpus was used to provide a standard for the linguistic features that could be associated with particular categories.

A significant issue in determining the subject content of chat interaction is the inevitable variation in subject matter within an individual session. On-line chat can have multiple topics in one session because there may be more than two people in a chat room at any point in time and topics of conversation can vary and change many times. The subject matter may range from specific topics like sports, entertainment, news and games, to very general topics like daily activities, vacations, plans for the weekend, gossip and life in general. In order to develop a means to gauge the content of any chat session, we had to select a set of categories that we could use for this experiment. Furthermore, our categories had to accommodate criminal-based activities in chat rooms, toward a classification of 'dangerous' versus 'innocent' content.

One approach to solving this problem was to use topics popularly discussed among persons of a particular age group. For instance, as adopted by Dong et al. (2006), who chose the categories of Sports, Entertainment, Games, Travel and Pornography for a research study based on the principle that most teenagers in chat room commonly discuss these topics. As well as these plausible generic (innocent) categories, our study also had to reflect topics that are considered to be offensive or dangerous activities. The content for these topics could not be readily drawn from chat rooms because users who address these topics would not want to be recorded and used for analysis. To avoid this problem, content that was readily available from the Internet black-lists was used instead. These blacklists are a collection of Uniform Resource Locators (URLs) and Internet domains used by Internet filtering programs to help web browsers and web servers filter content deemed dangerous or harmful. There are blacklists currently available for both commercial and private use, and each is categorized so as to assist the filtering programs and users to identify which content they wish to block.

One of these free blacklists was the Shalla's Blacklist (2011), which contains over 1.7 million entries of URLs and domains containing Internet content that users may wish to block. These entries were recorded in 74 categories. Some of these categories were harmful and criminal-based, while others contained information that one may wish to censor for other reasons. A feature of this list was that it contained web addresses to pages containing information about these categories. Therefore, the text from these web pages could be extracted and used as the basis for our reference corpus and our classification method. From the seventy-four categories in Shalla's Blacklist, we selected eleven. Five of these categories were considered representative of dangerous Internet content, while the other six categories represented Internet content that would likely be an innocent topic of discussion in a chat room. The categories, their descriptions and type of content they represent are listed in Table 1.

Table 1. Categories for classification of chat content

Category	Description	Type
Aggressive	Aggressive and racist content, including hate speech, divisive philosophy and racist literature.	Dangerous
Drugs	Information on availability and manufacture of legal and illegal drugs..	Dangerous
Hacking	Information on security weaknesses and how to exploit them, including sites offering exploits and software to take advantage of security weaknesses.	Dangerous
Violence	Content on harming and killing people, including torture, brutality and bestiality.	Dangerous
Weapons	Information on trading weapons and accessories for weapons, (guns, knives, swords, bows, etc), including general information on weapons and their use.	Dangerous
Hobby	Information about cooking, on-line and off-line games, gardening and pets. These were considered to be hobbies that people regularly discussed.	Harmless
Military	Information about military facilities related to the armed forces.	Harmless[2]
News	Information on current events.	Harmless
Recreation	Content covering humour, such as comic strips and funny stories, sporting activities, martial arts, restaurants, travel and health.	Harmless
Religion	Information on different religious practices, sects, customs and interpretations.	Harmless
Science	Information on all topics of chemistry. (This is distinct from the drugs category.)	Harmless

[2] The content of this category might be considered dangerous, but most military websites are used for recruitment and provide information for the general public.

In order to accommodate the fact that a chat session may have multiple topics, we sought a means to characterise the 'seriousness' of the chat conversation. This involved ranking the chat content in terms of the eleven categories using an aggregated score in order to classify the content as dangerous, suspicious or harmless.

2.1 Procedure

The following procedure was followed in reaching the classification and scoring for sample texts: Firstly, a series of steps was applied on the test chat corpora:

1. From the test chat corpora, ten session text files were chosen (based on their content as considered by manually reading the files). Five text files contained subject matter considered casual and harmless, while the other five contained material considered dangerous or harmful.
2. Each of these chat text files was manually ranked by category based on the selected categories for this project. The highest category represented the main topic of the chat conversation, while the lowest topic represented the topic of the conversation that was either least evident or not present at all.
3. The word frequencies for each file were extracted using AntConc (Anthony, 2005).
4. The bigram frequencies were extracted using the Posit Tools (Weir, 2007)
5. The data from the resulting word and n-gram frequency files were used as an input to the software application that implemented the classification method.

Secondly, the following steps were employed to establish and deploy the reference corpus:

6. Website pages from the URL lists were downloaded and saved to disk as HTML documents.
7. The text from each document was extracted and saved as a text file.
8. Each file was checked to see if they contained information with respect to the category
9. The remaining text files for each category were combined into one text file per category.
10. As with the test data, the word, bigram and tri-gram frequencies were extracted.
11. The data from the resulting word and n-gram frequency files were used as an input source for the software application, as a comparison with the test files.

Using our test chat corpus and a reference corpus derived from web sites (drawn in part from Shalla's blacklist), we extracted the word and n-gram frequencies from both corpora, and used statistical comparisons between the two in order to identify the categories involved. To this end, we considered two statistical measures, chi by degrees of freedom (CBDF) and the Log-Likelihood ratio of frequencies.

Chi by degrees of freedom (CBDF), proposed by Kilgarriff (Oakes, 2003), takes the chi-square value of a corpus comparison and divides it by the degrees of freedom. The higher the CBDF value is for two corpora, the more likely that they are similar to each other.

The Log-Likelihood ratio statistic of common word frequencies, also known as the G-Score (Oakes, 2003), is a form of log linear analysis which uses the logarithmic function to determine the likelihood that the frequency data from one corpus is similar to that of the other. Like CBDF, the higher the G-Score value is for two corpora, the more likely that they are similar to each other. Using these measures, we could compare content from the test corpus with our reference corpus to determine the most likely to the least likely category to which the test sample belongs. This approach can be used with a weighted scoring scale to classify the test content as dangerous, suspicious or harmless.

The statistical formulae for CBDF and Log Likelihood (i.e. G-Score) require the same input values, even though they are fundamentally different. Since the frequencies from the test and reference data sets were being used for comparison, the first values needed for both formulae were the frequencies of the common words and n-gram that exist in both test and reference data. This was known as the observed frequencies. The second values needed for these formulae were the expected frequencies of the common words and n-grams from both test and reference data. This was calculated by using the following formula (Oakes, 2003):

$$\text{Expected value} = \frac{\text{row total x column total}}{\text{grand total of items}}$$

The row total was the sum of the frequencies of a common word or n-gram and the column total is the sum of the frequencies of all common words or n-grams in the test data or the reference data. The grand total of items was the total frequencies of the common words or n-grams from both sets of data. For example, Table 2 shows the common word-forms from test (O1) and reference (O2) data sets.

Table 2. Common words with relative frequencies from test and reference data

Word	O1	O2	Row Total
the	35	78	113
you	57	32	89
it	34	79	113
Column Total	126	189	315

From this, the expected values for the words of the test (E1) and reference (E2) data would be as shown in Table 3:

Table 3. Common words with expected frequencies for test and reference data

Word	E1	E2
the	45.2	67.8
you	35.6	53.4
it	45.2	67.8

From the observed frequencies and expected frequencies, the CBDF and Log-Likelihood can both be calculated. The chi-square value was calculated from the observed and expected frequency values of each common word or n-gram, using the formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

For the examples given in Tables 2 and 3, the chi-square values for the common words are given in Table 4.

Table 4.: Chi-squared values of common words

Word	$(O1 – E1)^2 \div E1$	$(O2 – E2)^2 \div E2$	χ^2
the	2.30	1.53	3.84
you	12.86	8.58	21.44
it	2.78	1.85	4.63

The sum of the chi-squared values of all common words or n-grams was calculated. (In the given example this value would be 29.90). The degrees of freedom were calculated by counting the total number of common words or n-grams and subtracting it from one (1). (For the given example the degrees of freedom would be 3 - 1 = 2.) The CBDF was calculated by dividing the sum of the chi-squared values of all the common words or n-grams by the degrees of freedom. (In the given example, the CBDF value would be 29.90 ÷ 2 = 14.95.)

The steps taken to calculate the Log-Likelihood are similar (Rayson & Garside, 2000). The Log-Likelihood for each common word or n-gram was calculated using their observed and expected frequencies, and the following equation:

$$-2ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

For the previous example, the log likelihood values for the common word-forms would be as shown in Table 5.

Table 5. Log-Likelihood values of common words

Word-form	O1 x ln(O1 ÷ E1)	O2 x ln(O2 ÷ E2)	LL or $-2\ln\lambda$
the	-8.95	10.93	3.96
you	26.83	-16.39	20.89
it	-9.68	12.08	4.79

The sum of the Log-Likelihood of all words or n-grams was calculated to give a measurement to the comparison between the test and reference data sets. For the given example, the Log-Likelihood of the comparison is 29.64. Both CBDF and Log-Likelihood use the same values for input into their respective formulae. However, these calculations can only be completed with words or n-grams that appear in both test and reference corpora.

2.2 Scoring Scale

A scoring function was used to determine if the content of a chat conversation session was dangerous, suspicious or harmless in nature. The following steps were applied to determine the values and thereby the classification for any chat text.

1. The dangerous type categories were each given a score of ten as their weighted value (i.e, Aggressive, Hacking, Drugs, Violence and Weapons, each had a weighed value of ten).
2. The harmless type categories were each given a score of five as their weighted value (i.e., Hobby, Military, News, Science, Religion and Recreation, each had a weighed value of five).
3. When these eleven categories were ranked, the weighted scores of the top five categories were added together.
4. A scale was used to determine which class the chat session belongs to, dangerous, suspicious or harmless.

The maximum weighted score value of the top five categories was fifty, if all the categories were dangerous. If the weighted score value of the top five categories was twenty-five, then all the categories were harmless. With this in view, the procedure to determine in which class a chat conversation belonged was as follows:

```
if WS >= 40 then dangerous
else
    if WS > 30 and < 45 then suspicious
    else  harmless
```

For a chat session to be classified as ‘dangerous’, there needs to be at least three dangerous category types in the top five categories. To be classified as ‘suspicious’, a sample text needs to have at least two dangerous category types in the top five. Any sample with fewer than two dangerous category types in the top five was classified as ‘harmless’.

3. RESULTS

Our experiment compared the use of CBDF and Log-Likelihood in each of two conditions (single word frequencies and bigram frequencies) to determine how effectively these techniques would discriminate across the subject categories and contribute to classification of samples as dangerous, suspicious or innocent. The control for these comparisons was a manual rating of the sample texts. The five files with dangerous content were classified as dangerous or suspicious depending on their manually weighted score, while the other five files were classified as harmless on the same basis. The ten chat conversation sessions selected from the IRC and NPS Chat corpora were analysed for their textual content and their frequency files were analysed by our software application (TXTClassify) in relation to the collected reference data. The ranked category lists of these files were then compared with the ranked category lists generated from their CBDF and Log-Likelihood results, for both word and bigram frequency files.

3.1 Results Format

For each input sample, the TXTClassify application indicates CBDF and Log Likelihood values against each of the eleven content categories and classifies the sample as dangerous, suspicious or harmless, based upon the scoring scale described above. For instance, Table 6 shows the results using word frequencies for an input file called 'weed_town.txt'. In this example the CBDF and Log-Likelihood values for each of the subject categories determines the ranking and subsequent scoring of 35 and 30 respectively. This indicates a classification of suspicious on the basis of the CBDF ranking and harmless in terms of the Log-Likelihood ranking. These can be compared to the manual ranking, which gave a weighted score of 35 and the corresponding classification of suspicious.

Table 6.: Classification using word frequencies for 'weed_town.txt'

Rank	CBDF Results		Log-Likelihood Results		Manual Results
	Category	Value	Category	Value	
1	Hobby	307.007	Hobby	1589.293	Drugs
2	Recreation	217.959	Recreation	1435.366	Recreation
3	News	202.198	News	1406.732	Hobby
4	Drugs	52.069	Drugs	1082.698	Religion
5	Weapons	51.855	Religion	921.355	Aggressive
6	Religion	48.683	Weapons	805.462	Science
7	Military	29.897	Violence	733.402	News
8	Violence	25.990	Military	671.625	Hacking
9	Science	23.925	Science	584.201	Military
10	Hacking	19.483	Hacking	539.775	Violence
11	Aggressive	5.111	Aggressive	204.809	Weapons
Weighted Score	35		30		35
Classification	Suspicious		Harmless		Suspicious

The results for the same input file when using bigram frequencies are shown in Table 7.

Table 7. Classification using bigram frequencies from 'weed_town.txt'

Rank	CBDF Results		Log-Likelihood Results		Manual Results
	Category	Value	Category	Value	
1	Hobby	31.599	Hobby	389.019	Drugs
2	News	23.612	News	314.826	Recreation
3	Recreation	18.270	Recreation	250.330	Hobby
4	Drugs	3.251	Drugs	68.454	Religion
5	Science	2.628	Religion	48.946	Aggressive
6	Religion	2.628	Science	38.716	Science
7	Violence	2.174	Violence	37.032	News
8	Weapons	2.124	Hacking	31.188	Hacking
9	Hacking	2.114	Weapons	28.013	Military
10	Military	2.035	Military	19.649	Violence
11	Aggressive	0.649	Aggressive	4.147	Weapons
Weighted Score	30		30		35
Classification	Harmless		Harmless		Suspicious

For this example chat file (weed_town.txt), the results only show agreement between the manual classification and that of the CBDF with word frequency.

3.3 Classifications

A similar process of analysis was applied to the other nine sample chat files in order to compare the classification results for each of the two statistical similarity measures, in each of the two conditions (word frequency and bigram frequency), against the manual classification for each sample. The summarized results for all ten files are shown in Table 8, below.

Table 8. Classification results for all test files in all conditions

File	Classification				
	CBDF		Log-Likelihood		Manual
	Word	Bigram	Word	Bigram	
weed_town.txt	Suspicious	Harmless	Harmless	Harmless	Suspicious
casual.txt	Suspicious	Harmless	Harmless	Harmless	Suspicious
planetchat2.txt	Suspicious	Harmless	Suspicious	Harmless	Suspicious
11-08-teens.txt	Suspicious	Harmless	Suspicious	Harmless	Dangerous
11-08-adults.txt	Suspicious	Harmless	Suspicious	Harmless	Dangerous
englishbar.txt	Suspicious	Harmless	Suspicious	Harmless	Harmless
10-26-teens.txt	Harmless	Harmless	Harmless	Harmless	Harmless
11-09-40s.txt	Suspicious	Suspicious	Harmless	Suspicious	Harmless
10-19-adults.txt	Harmless	Harmless	Suspicious	Harmless	Harmless
11-09-adults.txt	Harmless	Suspicious	Harmless	Suspicious	Harmless

4. DISCUSSION

As indicated above, our results show that the software application, TXTClassify was able to take word and bigram frequencies from the collected test data and use it to complete the classification task. However, the classification results vary across the CBDF and Log-Likelihood statistical methods. Furthermore, results also vary when using word frequency data and bigram frequency data. There are also trends that were observed with respect to the ranking of the categories, with certain categories usually found in the top five for most of the classifications, even though the classification methods are fundamentally different.

The application of the two statistical formulae used in the classification method, was to examine which would be more effective in classification of on-line chat data. When comparing the CBDF and Log-Likelihood ranked results to the manual ranked categories for each file, it was clear that none of the ranked category lists were the same. This was expected since the manual ranked lists were based on reading the content of the test data, while the other lists were based on the calculated CBDF and Log-Likelihood values. However, for the test data considered as dangerous content, both the calculated and manually categorised results had dangerous categories in the top five for each list.

Notably, for the CBDF and Log-Likelihood results, none of the test data was classified as Dangerous. Therefore, none of the ranked category lists for CBDF and Log-Likelihood contained more than two dangerous categories in the top five. Overall, it appears that the classification of the test data using the CBDF method with word frequency was able to distinguish between the dangerous and innocent test data, with 100% of the dangerous test data files classified as Suspicious, and 60% of the innocent test data files classified as Harmless. None of the other methods came close to this result, as indicated in Tables .9 and 10, below.

Table 9: Percentages of dangerous data test files classified

Classification	CBDF		Log-Likelihood	
	Word	Bigram	Word	Bigram
Dangerous	0%	0%	0%	0%
Suspicious	100%	0%	60%	0%
Harmless	0%	100%	40%	100%

Table 10: Percentages of innocent data test files classified

Classification	CBDF		Log-Likelihood	
	Word	Bigram	Word	Bigram
Dangerous	0%	0%	0%	0%
Suspicious	40%	40%	40%	40%
Harmless	60%	60%	60%	60%

Inevitably, there are other factors that may impact upon the operation of these classification examples. The quality and character of the reference and data samples must directly influence the results. For example, the Hobby, Recreation and News categories appeared in the top five categories for most of the results while the Aggressive category was the lowest category for all results. Perhaps there was not enough data for the Aggressive category to make a proper comparison with the test data with the results that this category may always be toward the bottom of the ranked list. Furthermore, the Hobby, Recreation and News categories may have contained very common data instances and thus may not effectively discriminate topics within the chat conversations of the test data. This aspect requires further investigation.

In summary, from the observed results the classification method that used the CBDF statistical calculation and word frequencies from test and reference data, was able to come closest to the manually classified results and thereby was the most effective of the considered classification methods. However, the similarity of ranking results for the reference data in the Hobby, Recreation, News and Aggressive categories throughout the ranking lists, suggests that further content refinement is required.

REFERENCES

Anthony, L. (2005). AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning.

BBC, (2011). ‘Arrest over social network site damage incitement’, http://www.bbc.co.uk/news/uk-england-tyne-14521031. Last visited: 1st September 2011.

Dong, H., Hui, S. C., He, Y. (2006). ‘Structural analysis of chat messages for topic detection’, *Online Information Review*, 30(5), 496–516.

Forsyth, E.N. and Martell, C.H. (2007). ‘Lexical and Discourse Analysis of Online Chat Dialog’, *Proceedings of Semantic Computing 2007*, International Conference on Semantic Computing, IEEE, 19-26.

Kaiser, T. (2011). ‘NYPD Looks to Mine Social Networks for Info on Criminal Activity’, http://www.dailytech.com/NYPD+Looks+to+Mine+Social+Networks+for+Info+on+Criminal+Activity/article22417.htm. Last visited: 1st September 2011.

Oakes, M. (2003). ‘Text categorization: Automatic discrimination between US and UK English using the chi-square text and high ratio pairs’, *Research in Language*, 1, 143–156.

Rayson, P. and Garside, R. (2000). ‘Comparing corpora using frequency profiling’, in *WCC '00 Proceedings of the Workshop on Comparing Corpora* - Volume 9, ACM, 1-6.

Shalla’s Blacklist (2011). http://www.shallalist.de/. Last visited: 1st September 2011.

Stevenson, J. (2000). ‘The language of Internet Relay Chat’, http://www.demo.inty.net/Units/Internet%20Relay%20Chat.htm. Last visited: 1st September 2011.

Weir, G.R.S. (2007). ‘The Posit Text Profiling Toolset.’ In *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics*. Pan-Pacific Association of Applied Linguistics.

Weir, G.R.S., Toolan, F. and Smeed, D. (2011). ‘The threats of social networking: old wine in new bottles?’ Submitted for publication.