Categorizing Written Texts by Author Gender : *Literary and Linguistic Computing* 17(4).

Argamon S., **Koppel M.**, **Fine J.**, **Shimoni A**. (2003). Gender, Genre and Writing Style in Formal Written Texts : *Text* 23(3), pp. 321–346

Argamon S., Whitelaw C., Chase P., Hota S., Dhawle S., Garg N., Levitan S. (2005) *Stylistic Text Classification using Functional Lexical Features, Journal of the Association for Information Sciences and Technology*, to appear.

Corney M., Vel O., Anderson A., Mohay G. (2002). *Gender Preferential Text Mining of E-mail Discourse :* In *Proceedings of 18th Annual Computer Security Applications Conference ACSAC*

Corney M., Vel O., Anderson A. (2001). Mining E-mail Content for Author Identification Forensics *: ACM SIGMOD Record* Volume 30, Issue 4 December

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with many relevant features. *ECML-98, Tenth European Conference on Machine Learning*.

Platt, J. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Microsoft Research Technical Report MSR-TR-98-14,

Mitchell, T. (1997) *Machine Learning*. (McGraw-Hill)

Witten I., Frank E. (1999). *Weka3: Data Mining Software in Java* http://www.cs.waikato.ac.nz/ml/weka/**Tables**

Criticism Mining: Text Mining Experiments on Book, Movie and Music Reviews

Xiao HU
J. Stephen DOWNIE
M. Cameron JONES

University of Illinois at Urbana-Champaign

1. INTRODUCTION

There are many networked resources which now provide critical consumer-generated reviews of humanities materials, such as online stores, review websites, and various forums including both public and private blogs, mailing lists and wikis. Many of these reviews are quite detailed, covering not only the reviewers' personal opinions but also important background and contextual information about the works under discussion. Humanities scholars should be given the ability to easily gather up and then analytically examine these reviews to determine, for example, how users are impacted and influenced by humanities materials. Because the ever-growing volume of consumer-generated review text precludes simple manual selection, the time has come to develop robust automated techniques that assist humanities scholars in the location, organization and then the analysis of critical review content. To this end, the authors have conducted a series of very promising large-scale experiments that bring to bear powerful text mining techniques to the problem of "criticism analysis". In particular, our experimental results concerning the application of the Naïve Bayes text mining technique to the "criticism analysis" domain indicate that "criticism mining" is not only feasible but also worthy of further exploration and refinement. In short, our results suggest that the formal development of a "criticism mining" paradigm would provide humanities scholars with a sophisticated analytic toolkit that will open rewarding new avenues of investigation and insight.

2. EXPERIMENTAL SETUP

Our principal experimental goal was to build and then evaluate a prototype criticism mining system that could automatically predict the:

1) genre of the work being reviewed (Experimental Set 1 (ES1)).

2) quality rating assigned to the reviewed item (ES2).

3) difference between book reviews and movie reviews, especially for items in the same genre(ES3).

4) difference between fiction and non-fiction book reviews (ES4).

In this work, we focused on the movie, book and music reviews published on www.epinions.com, a website devoted to consumer-generated reviews. Each review in epinions.com is associated with both a genre label and a numerical quality rating expressed as a number of stars (from 1 to 5) with higher ratings indicating more positive opinions. The genre labels and the rating information provided the ground truth for the experiments. 1800 book reviews, 1650 movie reviews and 1800 music reviews were selected and downloaded from the most popular genres represented on epinions.com. As in our earlier work (Hu et al 2005), the distribution of reviews across genres and ratings was made as evenly as possible to eliminate analytic bias. Each review contains a title, the reviewer's star rating of the item, a summary, and the full review content. To make our criticism mining approach generalizable to other sources of criticism materials, we only processed the full review text and the star rating information. Figure 1 illustrates the movie, book and music genre taxonomies used in our experiments.

Books		Movies	Music
Fiction	Non-fiction		
Action & Thrillers [1]		Action/Adventure [1]	Blues
Juvenile Fiction [2]		Children [2]	Classical
	Humor [3]	Comedies [3]	Country
Horror [4]		Horror/Suspense [4]	Electronic
	Music & Performing Arts 5	Musical & Performing Arts 5	Gospel
Science Fiction & Fantasy [6]		Science-Fiction/ Fantasy 6	Hardcore / Punk
	Biography & Autobiography	Documentary	Heavy Metal
Mystery & Crime		Dramas	International
		Education/ General Interest	Jazz Instrument
Romance		Japanimation (Anime)	Pop Vocal
		War	R&B
			Rock & Pop

Figure 1: Book, movie and music genres from epinions.com used in experiments; Genres with the same superscripts are overlapping ones used in "Books vs. Movie Reviews" experiments (ES3)

The same data preprocessing and modeling techniques were applied to all experiments. HTML tags were removed, and the documents were tokenized. Stop words and punctuation marks were not stripped as previous studies suggest these provide useful stylistic information (Argamon and Levitan 2005, Stamatatos 2000). Tokens were stemmed to unify different forms of the same word (e.g., plurals). Documents were represented as vectors where each attribute value was the frequency of occurrence of a distinct term. The model selected was generated by a Naïve Bayesian text classifier which has been widely used in text mining due to its robustness and computational efficiency (Sebastiani 2002). The experiments were implemented in the Text-to-Knowledge (T2K) framework which facilitates the fast prototyping of the text mining techniques (Downie et al 2005).

3. GENRE CLASSIFICATION TESTS (ES1)

Figure 2a provides an overview of the genre classification tests. The confusion matrices (Figure 2b, 2c and 2d) illustrate which genres are more distinguishable from the others and which genres are more prone to misclassification. Bolded values represent the successful classification rate for each medium (Figure 2a) or genre (Figure 2b, 2c and 2d).

	Book	Movie	Music
Number of genres	9	11	12
Reviews in each genre	200	150	150
Term list size	41,060 terms	47,015 terms	47,864 terms
Mean of review length	1,095 words	1,514 words	1,547 words
Std Dev of review length	446 words	672 words	784 words
Mean of precision	**72.18%**	**67.70%**	**78.89%**
Std Dev of precision	1.89%	3.51%	4.11%

(a) Overview Statistics of Genre Classification Experiments

T P	Action	Bio.	Horror	Humor	Juvenile	Music	Mystery	Romance	Science
Action	**0.61**	0.01	0.06	0.01	0.02	0.03	0.20	0.05	0.02
Bio.	0.04	**0.70**	0.01	0.05	0.03	0.13	0.01	0.03	0
Horror	0.09	0	**0.66**	0	0.05	0	0.12	0.02	0.06
Humor	0.01	0.10	0	**0.74**	0.03	0.08	0.01	0.01	0.03
Juvenile	0.01	0.01	0	0.07	**0.86**	0.02	0	0.02	0
Music	0	0.09	0	0	0.01	**0.89**	0	0	0.01
Mystery	0.20	0	0.01	0	0.01	0	**0.70**	0.05	0.04
Romance	0.06	0.01	0.01	0	0.04	0	0.08	**0.78**	0.03
Science	0.03	0	0.02	0.01	0.11	0.03	0.01	0.13	**0.66**

(b) Book Review Genre Classification Confusion Matrix

T P	Action	Anime	Children	Comedy	Docu.	Drama	Edu.	Horror	Music	Science	War
Action	**0,77**	0	0	0.01	0	0.01	0.02	0	0	0.10	0.09
Anime	0	**0.89**	0.03	0.03	0	0	0	0	0	0.05	0
Children	0.02	0.01	**0.95**	0	0.01	0.01	0.01	0	0	0	0
Comedy	0.09	0.01	0.06	**0.52**	0.03	0.17	0.06	0.01	0.03	0.01	0.02
Docu.	0.02	0	0	0.04	**0.63**	0.01	0.19	0	0.09	0	0.02
Drama	0.16	0	0	0.12	0.10	**0.45**	0.05	0.03	0.03	0.01	0.04
Edu.	0	0	0.02	0.02	0.31	0.03	**0.57**	0	0	0.01	0.03
Horror	0.15	0.02	0.02	0.02	0.03	0.02	0.05	**0.69**	0	0.10	0.02
Music	0	0	0	0.01	0.18	0	0	0	**0.81**	0	0
Science	0.04	0.01	0.02	0	0.06	0.01	0.02	0.03	0	**0.76**	0.05
War	0.11	0	0.01	0.01	0.08	0.08	0.05	0.03	0.02	0.02	**0.59**

(c) Movie Review Genre Classification Confusion

T P	Blues	Classical	Country	Electr.	Gospel	Punk	Metal	Int'l	Jazz	Pop Vo.	R&B	Rock
Blues	**0.61**	0	0.10	0	0	0	0	0	0	0	0	0.29
Classical	0	**0.94**	0	0.03	0	0	0	0	0	0	0	0.03
Country	0	0	**0.92**	0	0.03	0	0	0	0	0	0	0.06
Electr.	0	0	0	**0.92**	0	0	0.06	0	0	0	0	0.03
Gospel	0	0	0.05	0	**0.80**	0	0	0	0	0	0.05	0.10
Punk	0	0	0	0.05	0	**0.71**	0.05	0	0	0	0	0.19
Metal	0	0	0	0	0	0	**0.89**	0	0	0	0	0.11
Int'l	0	0.04	0.00	0.04	0	0	0	**0.81**	0	0	0	0.04
Jazz	0	0	0	0.04	0	0	0	0	**0.89**	0.04	0	0.04
Pop Vo.	0	0	0.04	0.07	0	0	0	0.04	0.07	**0.68**	0	0.11
R&B	0	0	0	0	0	0	0	0	0	0.06	**0.88**	0.06
Rock	0.03	0	0.03	0	0	0	0.03	0	0	0.03	0	**0.89**

(d) Music Review Genre Classification Confusion Matrix

Figure 2: Genre classification data statistics, results and confusion matrices. The first rows in confusion matrices represent prediction (P); the first columns represent ground truth (T). 5- fold random cross-validation on book and movie reviews, 3- fold random cross-validation on music reviews

As Figure 2a shows, the overall precisions are impressively high (67.70% to 78.89%) compared to the baseline of random selection (11.11% to 8.33%). The identification of some genres is very reliable e.g., "Music & Performing Arts" book reviews (89%) and "Children" movie reviews (95%). Some understandable confusions are also apparent e.g., "Documentary" and "Education" movie reviews (31% confusion). High confusion values appear to indicate that such genres semantically overlap. Furthermore, such confusion values may also indicate pairs of genres that create similar impressions and impacts on users. For example, there might be a formal distinction between the "Documentary" and "Education" genres but the two genres appear to affect significant numbers of users in similar, interchangeable ways.

4. RATING CLASSIFICATION TESTS (ES2)

We first tested the classification of reviews according to quality rating as a five class problem (i.e., classification classes representing the individual

ratings (1, 2, 3, 4 and 5 stars)). Next we conducted two binary classification experiments: 1) negative and positive review "group" identification (i.e., 1 or 2 stars versus 4 or 5 stars); and 2) *ad extremis* identification (i.e., 1 star versus 5 stars). Figure 3 demonstrates the dataset statistics, corresponding results and confusion matrices.

Book Reviews			
Experiments	1 star … 5 stars	1, 2 stars vs. 4, 5 stars	1 star vs. 5 stars
Number of classes	5	2	2
Reviews in each class	200	400	300
Term list size	34,123 terms	28,339 terms	23,131 terms
Mean of review length	1,240 words	1,228 words	1,079 words
Std Dev of review length	549 words	557 words	612 words
Mean of precision	**36.70%**	**80.13%**	**80.67%**
Std Dev of precision	1.15%	4.01%	2.16%
Movie Reviews			
Experiments	1 star … 5 stars	1, 2 stars vs. 4, 5 stars	1 star vs. 5 stars
Number of classes	5	2	2
Reviews in each class	220	440	400
Term list size	40,235 terms	36,620 terms	31,277 terms
Mean of review length	1,640 words	1,645 words	1,409 words
Std Dev of review length	788 words	770 words	724 words
Mean of precision	**44.82%**	**82.27%**	**85.75%**
Std Dev of precision	2.27%	2.02%	1.20%
Music Reviews			
Experiments	1 star … 5 stars	1, 2 stars vs. 4, 5 stars	1 star vs. 5 stars
Number of classes	5	2	2
Reviews in each class	200	400	400
Term list size	35,600 terms	33,084 terms	32,563 terms
Mean of review length	1,875 words	2,032 words	1,842 words
Std Dev of review length	913 words	912 words	956 words
Mean of precision	**44.25%**	**81.25%**	**86.25%**
Std Dev of precision	2.63%	N/A	N/A

(a) Overview Statistics of Rating Classification Experiments

T P	1 star	2 stars	3 stars	4 stars	5 stars
1 star	**0.45**	0.21	0.15	0.09	0.10
2 stars	0.24	**0.36**	0.19	0.12	0.09
3 stars	0.11	0.17	**0.28**	0.22	0.21
4 stars	0.05	0.06	0.17	**0.41**	0.31
5 stars	0.04	0.07	0.17	0.26	**0.46**

(b) Book Review Rating Classification Confusion Matrix (5 ratings)

T P	1 star	2 stars	3 stars	4 stars	5 stars
1 star	**0.49**	0.19	0.17	0.08	0.07
2 stars	0.15	**0.45**	0.23	0.11	0.06
3 stars	0.04	0.24	**0.28**	0.27	0.17
4 stars	0.05	0.13	0.13	**0.41**	0.27
5 stars	0.07	0.03	0.16	0.20	**0.54**

(c) Movie Review Rating Classification Confusion Matrix (5 ratings)

T P	1 star	2 stars	3 stars	4 stars	5 stars
1 star	**0.61**	0.24	0.07	0.05	0.02
2 stars	0.24	**0.15**	0.36	0.15	0.09
3 stars	0.11	0.13	**0.41**	0.20	0.15
4 stars	0.03	0.06	0.10	**0.32**	0.48
5 stars	0	0	0.09	0.11	**0.80**

(d) Music Review Rating Classification Confusion Matrix (5 ratings)

Figure 3: Rating classification data statistics, results and confusion matrices. The first rows in confusion matrices represent prediction (P); the first columns represent ground truth (T). 5- fold random cross-validation on book and movie reviews, one single iteration on music reviews

The classification precision scores for the binary rating tasks are quite strong (80.13% to 86.25%), while the five class scores are substantially weaker (36.70% to 44.82%). However, upon examination of the five class confusion matrices it is apparent that the system is "reasonably" confusing adjacent categories (e.g., 1 star with 2 stars, 4 stars with 5 stars, etc.).

5. MOVIE VS. BOOK REVIEW TESTS (ES3)

We first formed a binary classification experiment with movie and book reviews of all genres. We then compared reviews in each of the six genres common to books and movies. To prevent the oversimplification of the classification task we eliminated words that can directly suggest the categories: "book", "movie", "fiction", "film", "novel", "actor", "actress", "read", "watch", "scene", etc. Eliminated terms were selected from those which occurred most frequently in either category but not both.

Genre	All Genres	Action	Horror	Humor/Comedy
Number of classes	2	2	2	2
Reviews in each class	800	400	400	400
Term list size	49,263 terms	24,552 terms	25,509 terms	26,713 terms
Mean of review length	1,608 words	933 words	1,779 words	1,091 words
Std Dev of review length	697 words	478 words	546 words	625 words
Mean of precision	**94.28%**	**95.63%**	**98.12%**	**99.13%**
Std Dev of precision	1.18%	0.99%	1.40%	1.05%

Genre	Juvenile Fiction /Children	Music & perorming Aarts	Science Fiction & Fantasy
Number of classes	2	2	2
Reviews in each class	400	400	400
Term list size	21,326 terms	23,217 terms	25,088 terms
Mean of review length	849 words	791 words	1,011 words
Std Dev of review length	333 words	531 words	544 words
Mean of precision	**97.87%**	**97.02%**	**97.25%**
Std Dev of precision	0.71%	1.49%	1.91%

Figure 4: Overview statistics of book and movie review classification experiments. All results are from 5 - fold random cross validation

The results in Figure 4 show the classifier is amazingly accurate (consistently above 94.28% precision) in distinguishing movie reviews from book reviews both in mixed genres and within single genre classes. We conducted a post-experiment examination of the reviews to ensure that the results were not simply based upon suggestive terms like those we had eliminated pre-experiment. Therefore, it can be inferred that users criticize books and movies in quite different ways. This is an important finding that prompts for future work the identification of key features contributing to such differences.

6. FICTION VS. NON-FICTION BOOK REVIEW TEST (ES4)

As in ES3, we eliminated such suggestive words as "fiction", "non", "novel", "character", "plot", and "story" after examining high-frequency terms of each category. The classification results are shown in Figure 5.

Fiction vs. Non-fiction	
Number of classes	2
Reviews in each class	600
Term list size	35,210 terms
Mean of review length	1,220 words
Std Dev of review length	493 words
Mean of precision	**94.67%**
Std Dev of precision	1.16%

(a) Overview Statistics of Fiction and Non-fiction Book Review Classification Experiment

T P	Fiction	Non-fiction
Fiction	**0.98**	0.02
Non-Fiction	0.09	**0.91**

(b) Fiction and Non-fiction Book Review Classification Confusion Matrix

Figure 5: Fiction and non-fiction book review classification data statistics, results and confusion matrix. The first row in confusion matrix represents prediction (P); the first column represents ground truth (T). Results are from 5- fold random cross validation

The precision of 94.67% not only verifies our system is good at this classification task but also indicates reviews on the two categories are significantly different. It is also noteworthy that more non-fiction book reviews (9%) were mistakenly predicted as fiction book reviews than the other way around (2%). Closer analysis on features causing such behaviors will be our future work.

7. CONCLUSIONS AND FUTURE WORK

Consumer-generated reviews of humanities materials represent a valuable research resource for humanities scholars. Our series of experiments on the automated classification of reviews verify that important information about the materials being reviewed can be found using text mining techniques. All our experiments were highly successful in terms of both classification accuracy and the logical placement of confusion in the confusion matrices. Thus, the development of "criticism mining" techniques based upon the relatively simple Naïve Bayes model has been shown to be simultaneously viable and robust. This finding promises to make the ever-growing consumer-generated review resources useful to humanities scholars.

In our future work, we plan to undertake a broadening of our understanding by exploring the application of text

mining techniques beyond the Naïve Bayes model (e.g., decision trees, neural nets, support vector machines, etc.). We will also work towards the development of a system to automatically mine arbitrary bodies of critical review text such as blogs, mailing lists, and wikis. We also hope to construct content and ethnographic analyses to help answer the "why" questions that pertain to the results.

References:

Argamon, **S**., and **Levitan**, **S**. (2005). Measuring the Usefulness of Function Words for Authorship Attribution. *Proceedings of the 17th Joined International Conference of ACH/ALLC*.

Downie, **J. S**., **Unsworth**, **J**., **Yu**, **B**., **Tcheng**, **D**., **Rockwell**, **G**., and **Ramsay**, **S. J**. (2005). A Revolutionary Approach to Humanities Computing?: Tools Development and the D2K Data-Mining Framework. *Proceedings of the 17th Joined International Conference of ACH/ALLC*.

Hu, **X**., **Downie**, **J. S**., **West**, K., and **Ehmann**, **A**. (2005). Mining Music Reviews: Promising Preliminary Results. *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR)*.

Sebastiani, **F**. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34, 1.

Stamatatos, E., Fakotakis, N., and **Kokkinakis, G.** (2000). Text Genre Detection Using Common Word Frequencies. *Proceedings of 18th International Conference on Computational Linguistics*.

Markup Languages for Complex Documents – an Interim Project Report

Claus HUITFELDT
Department of Philosophy, University of Bergen, Norway

Michael SPERBERG-MCQUEEN
World Wide Web Consortium, MIT Computer Science and Artificial Intelligence Laboratory (CSAIL)

David DUBIN
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Lars G. JOHNSEN
Department of Linguistics and Comparative Literature, University of Bergen, Norway

Background

Before the advent of standards for generic markup, the lack of publicly documented and generally accepted standards made exchange and reuse of electronic documents and document processing software difficult and expensive.

SGML (Standard Generalized Markup Language) became an international standard in 1986. But it was only in 1993, with the introduction of the World Wide Web and its SGML-inspired markup language HTML (Hypertext Markup Language), that generic markup started to gain widespread acceptance in networked publishing and communication.

In 1998, the World Wide Web Consortium (W3C) released XML (Extensible Markup Language). XML is a simplified subset of SGML, aimed at retaining HTML's simplicity for managing Web documents, while exploiting more of SGML's power and flexibility. A large family of applications and related specifications has since emerged around XML. The scope of XML processing and the