# MINING MUSIC REVIEWS: PROMISING PRELIMINARY RESULTS

**Xiao Hu**
GSLIS
University of Illinois
at Urbana-Champaign
xiaohu@uiuc.edu

**J. Stephen Downie**
GSLIS
University of Illinois
at Urbana-Champaign
jdownie@uiuc.edu

**Kris West**
School of Computing
Sciences, University
of East Anglia
kw@cmp.uea.ac.uk

**Andreas Ehmann**
Electrical Engineering
University of Illinois
at Urbana-Champaign
aehmann@uiuc.edu

## ABSTRACT

In this paper we present a system for the automatic mining of information from music reviews. We demonstrate a system which has the ability to automatically classify reviews according to the genre of the music reviewed and to predict the simple one-to-five star rating assigned to the music by the reviewer. This experiment is the first step in the development of a system to automatically mine arbitrary bodies of text, such as weblogs (blogs) for musically relevant information.

**Keywords:** music reviews, data mining, text classification, genre, and rating

## 1 INTRODUCTION

Music information retrieval (MIR) and music digital library (MDL) systems require both content-based and metadata-based music information. In networked environments, ever-increasing numbers of users are coming together to help each other with their music seeking tasks. The sharing of online music reviews is one such sharing mechanism that operates in the metadata domain. Developing tools that can help users of MIR/MDL systems acquire and use the wealth of music information embedded in online reviews is the goal of this pilot project.

Online customer reviews represent a rich resource for examining the ways users of music describe their music preferences and the possible impacts of those preferences. Online reviews can be surprisingly detailed, covering not only the reviewers' personal opinions but also important background and contextual information about the music and musicians under discussion. In addition, there is a large amount of review data online as most major online music stores (e.g., `amazon.com`) provide customer reviews. There are also non-retail websites devoted to customer reviews (e.g., `epinions.com`). These sources of user-generated information provide us with an exploratory starting point for uncovering new mechanisms for leveraging the collective knowledge of the music-listening public.

In this work, we use customer reviews of music CDs published on `www.epinions.com`, a website devoted to online customer reviews of products available on the Web. This site was chosen because it contains a very large collection of music reviews organized into a comprehensive and detailed genre classification taxonomy. There are 28 major classes including Classical, Rock, Pop, Jazz, Blues, International World music, etc. Under most classes in this taxonomy, there are subclasses categorized by various criteria (e.g., style, composer, etc.). For example, Classical music is divided according to the period in which the music was produced, including the Renaissance, Medieval, Classical, Baroque, Romantic and 20th Century periods. Each review is associated with both a genre and a numerical rating expressed as number of stars (from 1 to 5), with higher ratings indicating more positive opinions.

## 2 BACKGROUND AND MOTIVATION

### 2.1 Music Information User Studies

In recent years, research on user issues in music retrieval has attracted growing attention. [1] employed qualitative ethnographic methods to study music seeking behaviors in public libraries and music stores. Using a combination of interviews, focus groups and observations, they collected detailed data regarding user behaviors and the users' underlying motivations and goals. However, due to the time-consuming nature of such qualitative ethnographic methods, the Cunningham et al. study could not scale up; only seven subjects were intensely interviewed. Another user study in MIR applied survey methods to reach a larger group of users [3]. However, because survey methods have to use questions general enough to be minimally appropriate for all respondents, it is possible to miss what is most appropriate to many respondents. Further, survey designs (i.e., the tools and administration of the tools) have to remain unchanged throughout the data collection process, and thus cannot collect information about newly emergent categories previously unknown to the researchers. We believe a close examination of user-generated reviews provides an opportunity to obtain the benefits of traditional ethnographic methods (i.e., a detailed understanding of user expression via their own words) combined with the generalization abilities of well-constructed surveys. Furthermore, the application of automatic data mining techniques to the data analysis of the reviews allows for economies of scale unparalleled by qualitative methods.

### 2.2 Automatic Music Reviews

Whitman and Ellis [8] recently attempted to automatically generate textual reviews from music audio signals. For that purpose, they used music reviews to learn the connections between the perceptual audio features of music and textual terms in reviews. Whitman and Ellis also acknowledged that human description is a far richer source than marketing tags in terms of describing music content. Notwithstanding our mutual interest in music reviews, it is clear that our work is quite different from theirs: our work uses the full review text while Whitman and Ellis' focused on individual terms (i.e., nouns and adjectives) related to audio features. Moreover, since they used music reviews to establish the ground truth of their text descriptions of audio music features, they preferred "clean" music reviews which were "consistently concise, short and informative". In our work, we intend to develop systems based upon *all* aspects of music information use and users, and thus we need "natural" reviews from end users, which include comments on the music as well as the context and reasons for those comments.

## 3 EXPERIMENTAL SETUP

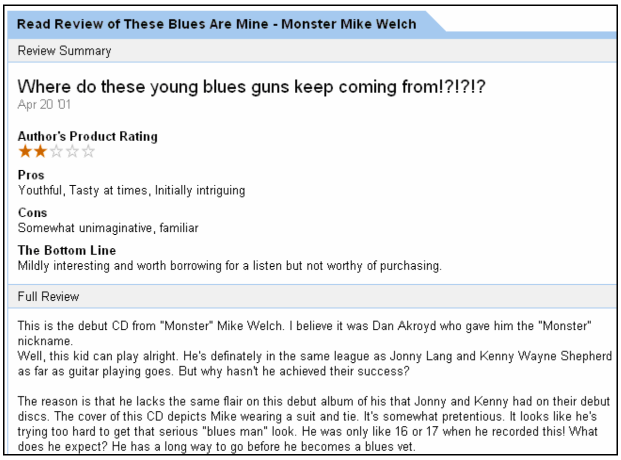
In this section we describe the experimental setup used to examine the automatic classification of reviews.

### 3.1 Data Collection

For each of the 12 genre classes used in our experiments, we crawled and downloaded CD reviews listed on the first 30 pages of the `epinions.com` product list. Each review contains a title, author's rating, a summary (expressed as "Pros", "Cons" and "The Bottom Line") and full review content. Figure 1 shows an example of a review. To simplify the process only the full review text and the rating were extracted from these documents. The title and summary are good resources to be exploited and we will do so in future work.

### 3.2 Classification Schemes

In this paper we attempt to identify: #1 the genre of the music being reviewed (Experiment #1); and #2 the rating assigned to the music by the reviewer (Experiments #2a, #2b and #2c). The same preprocessing and modeling techniques are used in both of these classification experiments. The genre of the music is not used as feature for the prediction of the rating, nor is the rating used as a feature to predict the genre to ensure that the models are entirely based on features that can be extracted from the text. We have tested the classification of reviews according to rating as a five class problem: classification into the individual ratings (1 star, 2 stars ... 5 stars) and binary classification problems: classification into negative and positive reviews (1 or 2 stars against 4 or 5 stars) and *ad extremis* (1 star against 5 stars).



Read Review of These Blues Are Mine - Monster Mike Welch

Review Summary

Where do these young blues guns keep coming from!?!?!?

Apr 20 '01

Author's Product Rating

★★☆☆☆

Pros

Youthful, Tasty at times, Initially intriguing

Cons

Somewhat unimaginative, familiar

The Bottom Line

Mildly interesting and worth borrowing for a listen but not worthy of purchasing.

Full Review

This is the debut CD from "Monster" Mike Welch. I believe it was Dan Akroyd who gave him the "Monster" nickname.
Well, this kid can play alright. He's definately in the same league as Jonny Lang and Kenny Wayne Shepherd as far as guitar playing goes. But why hasn't he achieved their success?

The reason is that he lacks the same flair on this debut album of his that Jonny and Kenny had on their debut discs. The cover of this CD depicts Mike wearing a suit and tie. It's somewhat pretentious. It looks like he's trying too hard to get that serious "blues man" look. He was only like 16 or 17 when he recorded this! What does he expect? He has a long way to go before he becomes a blues vet.

**Figure 1**. An example of a review on epinions.com

### 3.3 Dataset

The dataset used to investigate the automatic classification of reviews according to genre (Experiment #1) was composed of:

- 12 Classes (Rock, Pop, Jazz, Blues, Gospel, etc.)
- 150 examples per class
- A minimum of 3 kilobytes of text per review
- Total 1800 examples

The dataset used to investigate the automatic classification of reviews by user-assigned ratings (Experiment #2) was composed of:

- 5 Classes (1 star, 2 stars ... 5 stars)
- 200 examples per class (400 in the binary tests)
- A minimum of 3 kilobytes of text per review
- Total 1000 examples (800 in the binary tests)

#### *3.3.1 Data Preprocessing*

The first step in processing documents input to the system was to remove any residual HTML tags. The next step was to break the text down into terms and to remove all punctuations. The Porter stemming algorithm [6] was used so that different forms of the same word (e.g., plurals) would be recognized as the same term. The list of terms in each document was collected together to produce a global term list containing the frequency of each term in each document. The entire dataset was then represented as a sparse document-term matrix.

The sparse matrix produced by this process was then randomly divided into test and training matrices, with 80% of the data used to train a model and the remaining 20% held back to test the model's accuracy.

### 3.4 Modeling

The sparse training matrix is used to train a Naive Bayesian text classification model. Naive Bayes is a well-known probabilistic classification technique. Variations of the technique have been widely used in text categorization [4], [7] and [9]. As studies on multinomial mixture models have reported improved performance over multi-Bernoulli ones [4], in this paper, we have used a Naive Bayesian classifier based on a multinomial mixture model where values in document vectors are term frequencies (TF).

We can calculate the probability *P(C_j|d_i)* that a document, $d_i$, belongs to a category, $C_j$, by applying Bayes theorem, which states that:

$$P(C_j \mid d_i) = P(C_j) * \frac{P(d_i \mid C_j)}{P(d_i)} \qquad (1)$$

where $P(C_j)$ is the prior probability of class $j$, $P(d_i|C_j)$ is the conditional probability of document $i$ given class $j$ and $P(d_i)$ is the prior probability of document $i$.

The estimation of $P(d_i|C_j)$ is problematic because almost all novel documents are different from training documents. By making the assumption that each term in a document is generated independently of the other terms in the document given the class label, Naive Bayes simplifies the estimation of $P(d_i|C_j)$ to estimating the conditional probability of a term given a class:

$$P(d_i \mid C_j) = P(w_1, w_2, ..., w_n \mid C_j) = \prod_{t=1}^{V} P(w_t \mid C_j)^{c(w_t, d_i)} \qquad (2)$$

where $w_1, w_2, ..., w_n$ are terms occurring in document $d_i$, $V$ is the vocabulary of terms occurring in the training document set, and $c(w_t, d_i)$ is the frequency count of term $w_t$ in document $d_i$.

In practice, the accumulation of probabilities from all the terms occurring in a document must be performed in the log domain (to prevent underflow) and smoothing is necessary to prevent zero probabilities for infrequently occurring terms [4]. We used Laplacian smoothing, one of the most widely used smoothing methods, to smooth the probabilities in the log domain.

### 3.5 Implementation

The experiments detailed here were implemented in the Data-to-Knowledge Toolkit (D2K), the Text-to-Knowledge framework (T2K) and the General Architecture for Text Engineering (GATE). NCSA gives a thorough introduction to text mining in D2K/T2K [5].

## 4 RESULTS

The results achieved in each of the tasks, detailed in section 3.2, are given in Table 1 and the confusion matrices for genre classification (Experiment #1) and full ratings classification (Experiment #2a) are given in Figures 2 and 3, respectively.

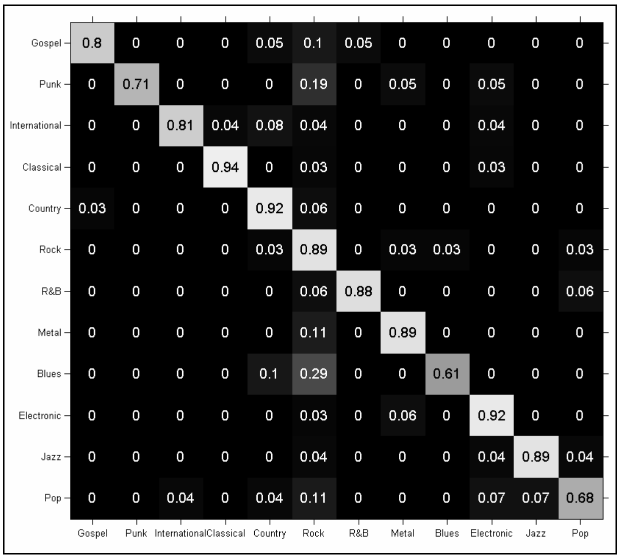


**Figure 2**. Genre classification confusion matrix

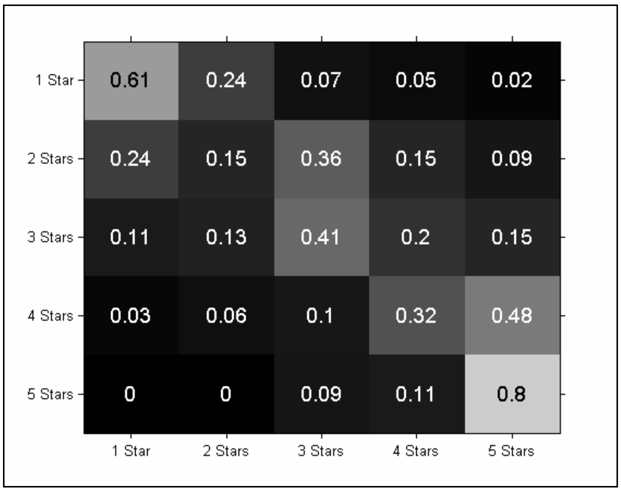


**Figure 3**. Rating classification confusion matrix

The Experiment #1 results shown in Table 1 along with the confusion matrix in Figure 2 show that the classification of music reviews, according to the genre of music reviewed, can be reliably performed. At 78.9% this performance is significantly better than the random baseline, which is 8.3%.

The Experiment #2a results for the prediction of the

Table 1. Music review classification results

| | Experiment | Accuracy | Std Dev | Classes | Term list size | Average length | Std Dev |
|---|---|---|---|---|---|---|---|
| #1 | Genre | 78.89% | 4.11% | 12 | 47,864 terms | 1,547 words | 784 words |
| #2a | Rating (1 star, 2 stars … 5 stars) | 44.25% | 2.63% | 5 | 35,600 terms | 1,875 words | 913 words |
| #2b | Rating (Good vs. Bad, 1/2 stars vs. 4/5 stars) | 81.25% | N/A | 2 | 33,084 terms | 2,032 words | 912 words |
| #2c | Rating (Good vs. Bad, 1 star vs. 5 stars) | 86.25% | N/A | 2 | 32,563 terms | 1,842 words | 956 words |

Results for Experiment #1 and #2a were calculated with 3 random cross-validation tests

Results for Experiment #2b and #2c were calculated with a single iteration rating that accompanies a review initially look quite poor. However, the confusion matrix in Figure 3 shows that confusion is most likely to occur with the neighboring classes, i.e., a 2 star review is most likely to be confused with a 1 or 3 star review. Therefore, despite a relatively low overall accuracy value, the system demonstrates the ability to distinguish a positive review from a negative review. This contention is further supported by the results of the binary rating prediction tests (Experiments #2b and #2c), which show that the accurate estimation of ratings is possible (Table 1).

## 5 CONCLUSION

We have demonstrated a proof-of-concept system that can successfully mine online music reviews, by applying a Naive Bayesian classifier, to predict both the genre of the music reviewed and the rating assigned to it by the reviewer. Both experiments were highly successful in terms of classification accuracy and the logical placement of confusion in the confusion matrix. The experimental results show that the mining of music reviews is a promising line of research, from which many user-related music features could be discovered.

## 6 FUTURE WORK

User-generated reviews can provide both users and researchers with music-related metadata in great quantity and detail. This exploratory study has examined a possible approach to exploiting this resource. More powerful automatic data mining techniques and ethnographic content analysis should be applied to more fully exploit the rich data available in user reviews. We intend to build upon the promise of our preliminary results by investigating the following possible applications: the recognition of reviews within an arbitrary body of text, such as weblogs (blogs), the separation of reviews of different media such as book, movie and music reviews, and the automatic classification and indexing of those reviews.

The subjects of many opinions expressed in the reviews are nouns or noun phrases (e.g., “lyrics”, “melody”), while most opinion words are adjectives (e.g., “awesome”, “crappy”). It is natural to hypothesize that nouns and noun phrases are salient features in genre classification while adjectives are important in rating classification. Research should be conducted into this hypothesis to reveal which parts-of-speech are important for each type of classification.

Opinion feature mining (OPF) [2] is another possible method of exploiting the information available in user-generated music reviews. OPF could be used to discover what features music users frequently mention when they write reviews about music CDs and to rank those features according to the frequency with which they appear in the reviews. Those same features are likely to be important in the selection of new music, and thus the identification of those features is important for the design of MIR/MDL systems that better serve the music information needs of their users.

## ACKNOWLEDGEMENT

The Andrew W. Mellon Foundation is thanked for their financial support. This project is also supported by the National Science Foundation (NSF) under Grant Nos. NSF IIS-0340597 and NSF IIS-0327371. We also thank the content providers and the MIR/MDL community for their support. We also thank the Automated Learning Group at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign.

## REFERENCES

[1] Cunningham, S. J., Reeves, N., and Britland, M. "An ethnographic study of music information seeking: Implications for the design of a music digital library", Proceedings of the third Joint Conference on Digital Libraries, Houston, USA, 2003.

[2] Hu, M. and Liu, B. "Mining opinion features in customer reviews", Proceedings of the 19th National Conference on Artificial Intelligence, San Jose, USA, 2004.

[3] Lee, J. and Downie, J. S. "Survey of music information needs, uses, and seeking behaviours: Preliminary findings", Proceeding of the Fifth International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain, 2004.

[4] McCallum, A. and Nigam, K. "A comparison of event models for naive bayes text classification", Proceedings of the AAAI8 Workshop on Learning for Text Categorization, Palo Alto, USA, 1998.

[5] NCSA. Online tutorial: Text mining: Email classification. Webpage, April 2005. http://algdocs.ncsa.uiuc.edu/TU-20031101-1.pdf.

[6] Porter, M.F. An Algorithm for Suffix Stripping. Program, 14, 3 (1980), 130 - 137.

[7] Sebastiani, F. Machine learning in automated text categorization. ACM Computing Surveys, 34, 1 (2002), 1- 47.

[8] Whitman, B. and Ellis, D. "Automatic record reviews", Proceeding of the Fifth International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain, 2004.

[9] Yang, Y. An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1, 1-2 (1999), 69 - 90.