Benjamin B. Pennell. Effectiveness of Text Representations in the Automatic Classification of Regional Game Design Trends in Video Game Reviews. A Master's Paper for the M.S. in I.S degree. December, 2008. 56 pages. Advisor: Catherine Blake

The video game industry is one of the fastest growing segments of the global entertainment market, and thus represents the design decisions of a wide spectrum of developers. This paper seeks to show that text mining can be used to predict trends in game design by identifying the region and release date automatically from video game reviews. When framed as a multi-class classification problem, a Support Vector Machine (SVM) achieves an average predictive confidence of 30.27% for noun and verb text representations, or by individual text representation: text windowing 11.22%, noun phrases 32.15%, noun phrases without game titles 31.40%, noun phrases with verbs 29.66%, individual term 27.88%. The SVM achieved better performance of 52.93% when predicting the release date trained on nouns and verbs. By text representation, the classifier found: noun phrases 62.97%, noun phrases without game titles 55.13%, noun phrases with verbs 61.10%, individual term 36.51% features.

Headings:

Text Mining

Natural Language Processing

Text Representation

Video Games

Review Mining

EFFECTIVENESS OF TEXT REPRESENTATIONS IN THE AUTOMATIC CLASSIFICATION OF REGIONAL GAME DESIGN TRENDS IN VIDEO GAME REVIEWS

by
Benjamin B. Pennell

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

December 2008

Approved by

_____

Catherine Blake

Table of Contents

1 Introduction

As one of the fastest growing segments of the global entertainment market (Szalai, 2007), video games have come to represent a wide spectrum of values in terms of expectations and design decisions of the groups involved in their development and publication. Despite a rapidly expanding market, changing demographics and growing global economic importance, regional differences in game designs have yet to be fully explored. The underlying premise of this research is that game components, genres, and themes within a descriptive review of video game products should reveal world regions. This paper seeks to explore the effectiveness of multiple text representations and classifiers to accurately predict the development region of a video game based upon the reviews.

The video game industry has matured to the extent that it has already attained a size comparable to those of both the music and movie industries (Otobe, 2007). Research by the Entertainment Software Association (ESA), the trade association for the computer and video game industry in the United States, has found that the sales of video games have more than tripled in the last 12 years (ESA, 2008). In addition to economic growth, the ESA also reports that the average age, gender ratio and size of the user base are changing. Between 2007 and 2008, the ESA reported that the average age of gamers had increased from 33 to 35 years old, the percentage of female gamers had increased from 38% to 40%, and males under the age of 17 only represent 18% of the overall market. Additionally, the number of American households owning home video game consoles

had increased from 33% to 38%, and the number of households who played either computer or video games had increased to 65%.

As a result of heightened competition due to this expanding demographic and staying up to date with new technologies, the rise in development costs has dramatically increased the level of risk versus profitability (BBC News, 2007). This may in turn result in a decline in risk taking as manufacturers specialize and tailor their products for specific target markets. The selection of hardware platform does, however, have an influence on video game design in two ways parallel to other developer specific issues like regional market expectations. First, different hardware design constraints apply to the types of features available, and second individual developers have preferences towards particular platforms (Otobe, 2007). These difference are, however, mitigated to some degree by the homogenization of features such as online connectivity and 3D rendering across hardware platforms. Therefore, given these broad changes in the industry it is an empirical question I address in this paper to determine how significant the differences in video game feature sets are between geographical regions.

Despite the rapid expansion and heightened competition of the video game industry, little has been done to formalize video game design classification beyond simple genre groupings. Although reviewers and review organizations appear to share an expanding vocabulary of terms, the lack of a formal vocabulary with a fine degree of granularity makes comparing regions more problematic. By employing text mining techniques to automatically extract features selected by reviewers; this project seeks to get at the

descriptions important to modeling the relationship between game design and geographic region.

Specifically, this project will determine the effectiveness of support vector machine classification systems to predict the development region of a video game based on the text used in a collection of reviews. Review data was extracted from two sources of electronic video game reviews and combined with a listing of developers and publishers. Further, the paper compares the performance of five text representations of the reviews, including a sliding window of terms, noun phrases, noun phrase and verbs, and individual terms with natural language processing informed stopword removal. Using the Oracle Data Mining tool, binary and multi-class classifiers were constructed based on each of the alternative text representations. Lastly, the effectiveness of region as a classifier will be compared against temporal grouping of reviews by release date over the same noun and verb based representations.

2 Related Work

2.1 General Text Mining

Although usage of text mining in relation to video games has not been extensively explored, Bragge and Storgårds describe the use text mining to perform a literature review of video game research (Bragge & Storgårds, 2007). They employ text mining to determine the topics commonly studied in video game related research, as well as the key people and organizations involved in the area. Their findings indicate that the number of gaming publications has doubled between 2004 and 2007, clearly demonstrating the growing interest in and importance of video games as an area of research. Although their focus was to review video game literature rather than to investigate the text mining techniques involved, Bragge and Storgårds do demonstrate the applicability of text mining in relation to video game knowledge discovery, as well as displaying the myriad of approaches that can be taken in studying video games.

Although there are a few examples of the use of text mining techniques in regards to video games, there is a considerable body of research related to general text mining procedures and specific applications of them in the information science field. For example “Mining MEDLINE: Abstracts, sentences, or phrases?” by Ding, et al, compared abstracts, sentences and phrases as units, similar to this paper's exploration of alternative text representations such as phrases and individual terms (Ding, et al., 2002).

One of the best generalized comparisons of text categorization techniques can be seen in Yang's article “An Evaluation of Statistical Approaches to Text Categorization” (Yang, 1999). This article provides useful descriptions of many of the available classification algorithms as well as comparisons of each algorithm's performance. Yang's survey of classifiers such as decision trees and Naïve Bayes were helpful in determining which subset of algorithm's to explore in this work, although support vector machines were not explicitly covered. In addition, Yang provides background on evaluation methods, binary classifier construction, preprocessing and feature selection, all of which were applicable in the current study.

For the selection phase of the mining process, “Eliminating Noisy Information in Web Pages for Data Mining” by Yi, Liu, and Li provided valuable information regarding how to handle and parse web resources being pulled as a corpus (Yi, Liu & Li, 2003). The previous work discusses various types of noise that can appear in websites, such as the standard navigation, advertisements and privacy notice “blocks” and how to separate them from the desired information, namely the main content block. Operating based on the theory that noise blocks will share common contents and presentation style with each other; whereas the real content should likely be much more diverse in the actual structure and contents, the researchers attempt to create a tree structure to capture the common styles found in a particular website. Although the data in this experiment had less structural variation than the previous work due to limiting the search space to only two websites, identification of unwanted or noisy blocks of text was essential to mining online resources not specifically prepared for that purpose.

2.2 Review Processing

Although video game reviews are not specifically targeted, there is a long history of data mining in regards to product reviews in general. Quite often the focus is dually between determining the sorts of features identified in the review and the opinion of those features, for example Dave, Lawrence, and Pennock investigate the process of automatically extracting and determining the opinions of the reviewer in regards to the features identified (Dave, Lawrence & Pennock, 2003). As in the case of my own study, the reviews were selected from two online review websites, presenting data selection from noisy online resources. The previous work also used both Naïve Bayes and Support Vector Machine classifiers, as well as discussed the preprocessing and feature selection required in such a project. Although the current project does not consider the opinions of reviewers in regards to the features selected, this consideration and numerous other approaches attempted in this article could be an important direction for future studies.

There has also been interest within the market analyst community in exploring the connection between sales performance and critical reception of a game. The market analyst Schachter showed in one study, reported on by New York Times writer Schiesel, that the opinions represented in critical reviews of video games correlate with sales (Schiesel, 2007). Statistically, the relationship between good reviews and high sales is much stronger in the games industry than in either film or music. Given this direct correlation it seems logical that market researchers would be interested in determining what these reviewers are stating that leads to higher sales. In addition to opinion

identification as seen in Dave, Lawrence, and Pennock's study this topic could be useful in future research.

So even though text mining has been used with product reviews, and the degree to which video game reviews influence purchasing habits, using text mining to investigate the relationship between vocabulary in reviews and the geographical region of a video game product does not seem to have been directly explored.

2.3 Classification Systems

Classification systems for video games have been studied to some extent in the past; however no consensus seems to have been reached in the community as to the best approaches. One possibility presented by Alvarez, et al. in the article "Morphological study of the video games" is to focus very heavily on gameplay oriented classifiers rather than feature oriented ones (Alvarez, et al., 2006). The central organization structure they discuss is to break down the components of video game design into small, distinct units which they refer to as "bricks." Higher level, generalized components are listed as well, including the basic metadata of the game such as author and title, the means of interaction within the game, and the logical rules that make up the game world.

Most of the focus of the paper, however, is on a ground up approach based around these "bricks." Games are mainly classified according to tiny descriptive components such as "Avoid," "Collect," "Move," or "Destroy," but there is also a hierarchy of sorts present in the concept of "metabricks." Each metabrick is a combination of singular bricks, for

example “Move” and “Avoid” gives “Drive,” which are generally more descriptive of the actual content of the video game than the base level bricks.

The objective in Alvarez's research was to develop a framework that would not be constantly outdated by advancements and trends in the industry. It is however somewhat questionable how scalable a system like this is as all of the data used for classification must be generated manually through playing the games, but their approach to gameplay classification is an interesting reference point. So although the system is built to be future proof by way of its very basic descriptors, it would likely be difficult to continue classifying large numbers of games in this way, and the classification scheme also misses aesthetic considerations which can have a critical effect in shaping a user's opinion.

Another classification method outlined by Leino attempts to classify games by the emotions elicited by objects in the game (Leino, 2007). In Leino's model components of a video game are classified according to whether they are deniable or undeniable, defined colorfully as “deniable meanings are ones which can be denied without [decreasing the players ability to act in the game], like the shape of Bismarck's [mustache] in Civilization 4.” Essentially this means that Leino classifies components of a game into those that have effects on the gameplay versus those that are simply there for effect, which is certainly quite a bit different from the other classification methods.

One area where video game classification has been investigated more extensively in the past has been the genre known as Multi-User Dungeons (MUDs). In one such instance,

Keegan discusses the classification of MUDs based upon his work in constructing "family trees" of past MUDs (Keegan, 1997). His approach involves two classification schemes, the first of which involves classifying software based on the code and similarities it shares with previous multi-user dungeons, which is very much in line with the concept of a MUD family tree. The second involves looking at the characteristics of the game worlds; in particular the ways in which players can modify the game world and in how the resources in the world replenish themselves. Although it would be fascinating to build a family tree for other video game genres, Keegan's ideas on typological classification, examining the characteristics of the game worlds, bears some similarities to my own work that attempts to extract game design components.

3 Operational Definitions

In this paper, a *video game product* is defined as a piece of interactive electronic entertainment software which must be playable on a home computer, electronic mobile device, or some form of *recent video game console hardware*. This definition is in some ways similar to that proposed by Nicolas, defining games as software "which we play thanks to an audiovisual apparatus and which can be based on a story" (Nicolas, 2005). Additionally, due to the reviewing practices of the sources from which reviews were retrieved in this project, games must follow a traditional purchase model, where there is an upfront purchase price to gain access to or ownership of the game. Games which meet these criteria will be included regardless of any subsequent payment fees, such as are found in some online games.

Video game console hardware refers to specialized desktop computer used to play video games (PC Magazine, 2008), or "an electronic system that connects to a display (as a television set) and is used primarily to play video games" (Merriam-Webster, 2008). Further, PC Magazine states that video game consoles often employ proprietary operating systems and hardware not found in standard home computers, which is strictly defined by the original manufacturer. A few examples of video game consoles include: The Nintendo Wii, Nintendo Gameboy, Sony Playstation, Sony Playstation Portable, and Microsoft Xbox. For this project it must also be a *recent console*, meaning that the console is a part of the current and previous generation of console hardware. Video game hardware *generations* occur roughly every 4 years and involve phasing out the current models of console hardware and introducing new models. Alternately, a *video game*

platform is almost synonymous with console hardware except that platforms include all console hardware as well as home computers and mobile devices. Although platforms are not the focus in this paper, they are an integral component in understanding the structure of the games industry, and as is explained later, they were taken into consideration in restricting data selection.

In addition to video game console hardware platforms, two generic hardware platforms were considered; *mobile devices* and *personal computers*. In contrast to console hardware platforms that are produced by a single manufacturer, such as Nintendo or Microsoft, generic platforms are usually broad labels for hardware produced by a variety of manufacturers. *Mobile devices*, such as PDAs or cell phones, are portable devices capable of playing games, but where that is not the primary function of the device. *Personal computers* make up an even broader category, simply referring to computing platforms available to consumers. It must also be noted that these two platforms do not follow the same iterative life cycle found in console hardware due to their incremental model of improvement, which shall also be considered later during data selection..

Aside from platform association, each video game can also be defined using a series of groups of nouns, adjectives, and adverbs referred to as *game design components*. These portions include gameplay features, genres, user interfaces, plot, character development, communication features, and visual design. There is not actually a predefined list of terms, but they are instead extracted from the bodies of review articles during preprocessing phases of the text mining process.

Text mining is the technique of deriving new information from bodies of textual information and is closely related to the fields of data mining and information retrieval(Weiss et al, 2005). Text ming involves seeking patterns and performing a series of statistical operations on a corpus of text, in this case to come up with predictive models for classifying the objects related to the text. The *corpus of text* in this case is composed of *video game reviews*, which are evaluative online articles describing the authors opinion of an entire video game product in its final completed form as is to be released to consumers. The authors in this case are employees of the online video game websites through which they publish their reviews.

The predictive model created via text mining in this case is intended to predict the region in which the game was developed. A *development region* is the geographic area in which the team primarily responsible for programming the game and creating its assets is based. These regions are based off of *lock-out regions* for DVD and game sales, which prevent products from one region from being playable elsewhere (Ip & Jacobs, 2004).

4 Methodology

The structure of this project closely follows that of a standard knowledge discovery application, adhering to the main four stage model of data selection, preprocessing, transformation and classification (Fayyad, Piatetsky-Shapiro & Smyth, 1996). Ultimately, the goal is to generate a predictive model that accurately determines the region of a game according to the terms presented in a review. In addition to following the traditional knowledge discovery model, the work was broken down into five different experiments for alternate datasets based on the manner in which terms were selected from the corpus. Some portions of the process apply to all the text representations, but there are also divergences which have been documented, primarily in the transformation phase.

4.1 Data Selection

The data used in this project falls into two main categories, the review data and developer locations, where each specific type of information used is harvested from publicly accessible websites.

4.1.1 Reviews

Reviews were retrieved from two online gaming news websites, www.gamespot.com and www.ign.com, which were selected due to their high coverage of video game releases across all major hardware platforms, as well as a extensive archive of reviews on current and previous console generations since 1996. In addition to retrieving the 13,661 total reviews from both sources, other metadata related to each reviewed game was retrieved, including but not limited to the title, name of the publisher, platform, release date, and

name of the developer. In the case of the preliminary text windowing experiment, which was performed one year beforehand, 11,075 reviews were retrieved from the same sources.

The selected set of reviews used for the corpus in this project did not cover the entire span of reviews available from either source, but instead was restricted based on both the platforms and the time periods in which the games were released. For the purposes of this project, the selected platforms were restricted to: the Microsoft Xbox, Microsoft Xbox 360, Mobile Devices, Nintendo DS, Nintendo Gameboy, Nintendo Gameboy Advance, Nintendo GameCube, Nintendo Wii, Personal Computers, Sony Playstation 2, Sony Playstation 3, Sony Playstation Portable. Figure 1 illustrates the breakdown of reviews according to the platforms selected.

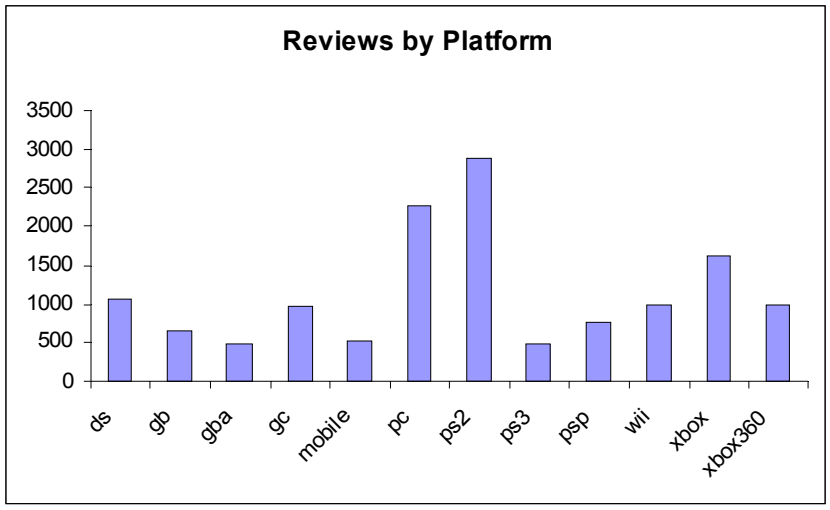

Figure 1: Reviews broken down by hardware platform within the time span of 2002 to 2008.

Only the most recent two generations of console hardware were selected even though a third was available. This decision to exclude the earliest of the three available generations was based in part on an inconsistency between the two websites in their choice of platforms to cover before 1998, which would have been partway through the third generation. Also, extending the time frame too far back likely would have muddled the classification, as the feature sets of games have been forced to change extensively over time due to innovation, competition and evolving hardware.

As this implies, restricting by console generation carries a temporal connotation as well in the case of console games, as gaming platforms are generally bound to specific development life spans, and usually begin to be phased out once a new generation of game platforms is released. For other platforms such as the PC or mobile devices, the same delineating hardware life cycles do not exist therefore requiring a superficial a cutoff of the beginning of 2002 to keep the time frame in line with that of the consoles. This date coincidences with the year marking the beginning of the console hardware generation that included the Sony Playstation 2, Nintendo GameCube, Microsoft Xbox, and Nintendo Gameboy Advance, although the Playstation 2 was released the year before. The distribution of reviews published between www.gamespot.com and www.ign.com over the time period selected for this project can be seen in figure 2.

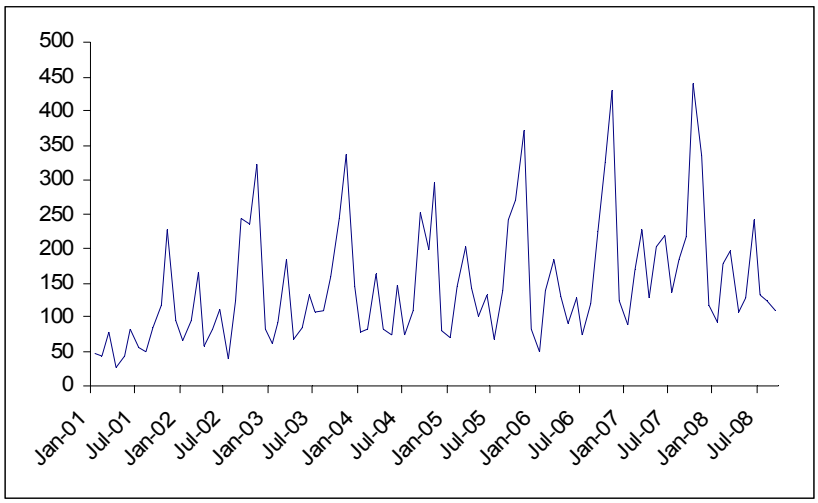

Figure 2: Quantity of reviews released over time. The large spikes in reviews follow the flurry of game releases that lead up to the winter holiday season.

In some cases there are multiple reviews of games with the same review body and title, as it is a common practice to release a game onto multiple platforms without changing its contents significantly. However, it is not trivial to eliminate duplicate titles on different platforms as there are instances in which a game for the PS2 and Nintendo DS are based off the same franchise, but are structured completely differently, and sometimes even created by different developers. For example, the games created by Ubisoft based off the Peter Jackson King Kong movie(Navarro, 2005) shared identical review texts for the Playstation 2 and Xbox versions, meanwhile the Nintendo DS version used the same title but had completely different content and was viciously panned by critics (Provo, 2005).

Data harvesting was performed using Java based crawlers and HTML parsers made specifically for this project. The harvesting process consisted of two majors steps: generating a list of URLs to each review article, and downloading review articles spanning one to eight web pages. Each Java harvester program also accounted for issues

such as navigation past randomly inserted full page click-through advertisements or pages that failed to load due to server or network issues.

4.1.2 Video Game Organizations

Developer locations were retrieved from the website www.gamedevmap.com, which maps video game developers and publishers to their city and country of origin. There are 1945 organizations listed in this dataset at present, although the website is updated multiple times per year. Organizations are divided into six categories according to their role, including developer, publisher, developer and publisher, or organization, as well as their primary distribution medium, including mobile/handheld and online developer. Figure 3 shows how the selected organizations are distributed according to these six roles. The listings from this website do not have complete coverage of the video games selected in the review dataset, in part due to www.gamedevmap.com's exclusion of free-lance developers or other very small teams, which the website defines as groups of fewer than 5 employees.

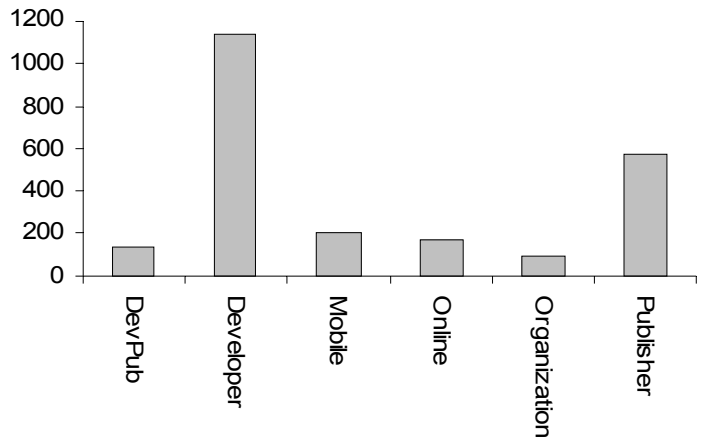

Figure 3: Number of organizations by role.

After taking into account naming discrepancies, as is describe in section 4.2.1, between the metadata indicating developers or publishers in review sources and the names provided in the game organization listings, there were 374 organizational entities from the reviews that did not have regions. The additional named entities include aliases for preexisting organizations, abbreviations, stock names, branch offices of major companies, small independent studios, second party developers, and companies that no longer exist. In these cases, I manually assigned the entity to a country of origin based on information available online, from official company websites if available, or other online resources such as wikipedia.org and news articles if no authoritative source was found. The procedure used for matching reviews with organizations will be discussed in further detail in the preprocessing section of this paper.

Extraction of data for game organization listings was centralized into six individual pages based on organizational role which could simply be copied and pasted from a web browser in tab delimited format.

4.2 Preprocessing

Two steps were required for preprocessing, first processing the raw data from reviews into a usable format for natural language processing and insertion into the database, and then building associations between organization and review metadata.

4.2.1 Preprocessing Reviews

The first phase of locating and extracting metadata and review bodies was taken care of with a combination of the crawlers and parsers used in the data selection phase of this project. Specialized parsers were created to understand the HTML templates used by each website to allow for extraction of the review text and its associated metadata.

After reducing the original HTML documents to tagless review bodies, the text was split into sentences using an algorithm similar to that outlined by Weiss, et al., which focused on identifying end of sentence punctuation and reconstruction of acronyms or numbers based on the surrounding characters (Weiss, et al., 2005). All sentences were then associated with identifiers to indicate the review, sentence number, and word number within that sentence to allow for rebuilding the structure in later phrases. A second pass was later employed to identify the sentences which would appear at the end of a 1mb file for segmenting the data before passing through natural language processing.

4.2.2 Building Region, Organization and Review Associations

Preprocessing the 2319 organizations retrieved from www.gamedevmap.com and other sources was limited to building associations with review metadata and vocabularies in

order to create a training field for building a classification model later. Organizations were associated with the major regional categories selected for this project and then matched with the developer and publisher metadata from each review.

Four regions were selected base on DVD and game lockout regions; North America, Asia, Europe/Australia, and Other, which included countries in South America and Africa. There were 65 unique countries associated with the selected game organizations, which were then assigned to a region based on geographical location.

Matching review metadata to the region names included in the main organization listing involved considerably more steps. There were 11,800 direct matches, 6935 of which were based only on the developer name, but in many cases either source's company name would contain minor variations such as common suffixes like “entertainment” or “inc”. The following steps were used to match video game developers with regions:

1) Retrieve a list of organizations where no direct match was found for the developer,
2) Join the listings of developers in each dataset by performing partial name matching between the developer and all available organization names,
3) Weight the results from step 2 with an organizational role weighting chart, which values developers highest, publisher lowest, selecting out the top rated organizations per review by role weight,
4) In the case of multiple matches with the same role, restrict to company headquarters organization,
5) If no headquarters, multiple headquarters are available, or multiple organizations with different names, restrict based on similarity and length of names.

Step 5 was rarely needed, as the previous four restrictions generally eliminated all duplicate matches. The process was repeated once more reversing the matching so as to find cases where the name from the organization list was a substring of the review

metadata field, and repeated another two times with a different weighting order for matching publishers in the same fashion.

After performing these sequences of steps, the results from each were merged back in with the direct match name sets. This data was then matched back against the review metadata in order to create a review organization join list, where matches with the developer metadata field were selected over matches with the publisher. The final join contained reviews with 6734 developer, 2690 developer/publisher, 6262 publisher, 173 online, 390 mobile, and 11 organization favored matches. Figure 4 shows the results of combining organization matching with region matching, and figure 5 shows how the ratio of the quantity of reviews breaks down by region over time.

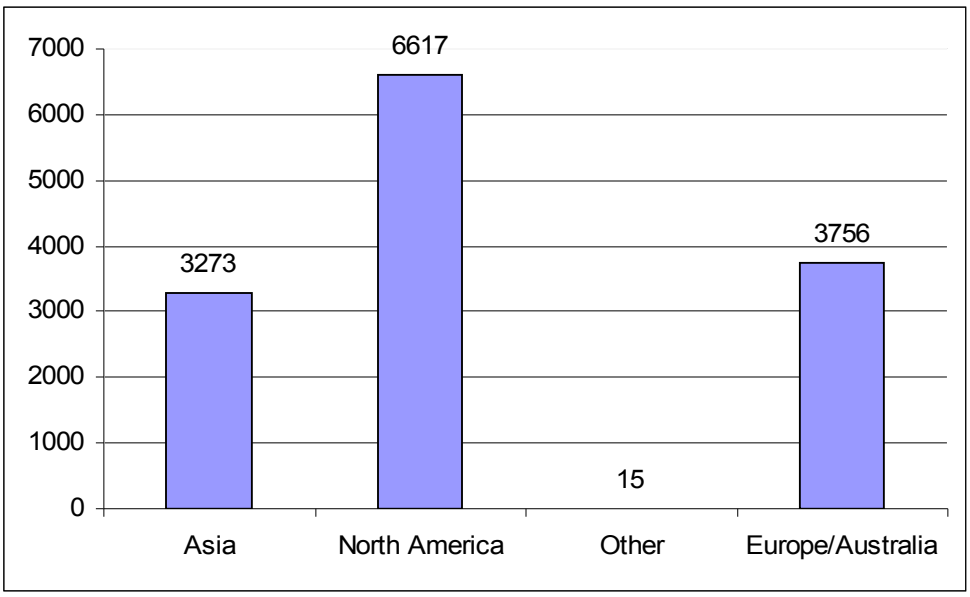

Figure 4: Quantity of reviews written about games according to the region in which the games developer or publisher is located.

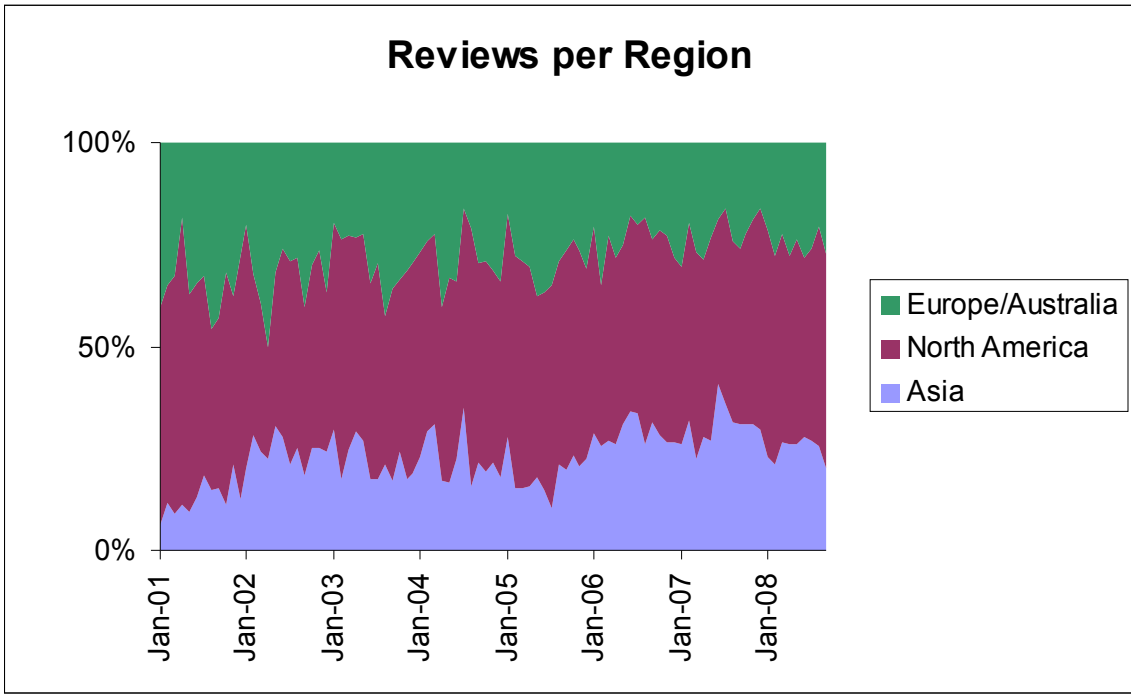

Figure 5: Ratio of reviews produced over time by game development region.

4.3 Transformation

4.3.1 Text Windowing

Although no natural language processing was used, the original experiment employed stopword removal as its primary means of cleaning up its list of terms. In this case the word removal was performed using a modified version of the MySQL Full Text Stopword removal list (Sun Microsystems, 2007), to include and exclude some terms found frequently in this specific corpus. After stopword removal was performed, all remaining terms were then processed with a Java based implementation of a Porter Stemmer (Porter, 2000).

Upon loading the data, windowing was employed to obtain multi-word phrases ranging between 2 and 5 words. In retrospect, this phase most likely would have been much easier to implement had windowing been performed before loading the data into the database, but this realization came fairly close to the end of experiment. PL/SQL functions generated and stored the multi-word phrases by using queries which took advantage of the word order information associated with each term. Some additional culling at this point helped to eliminate as many video game titles and company names from the body as possible, as these would likely bias the results and generally do not classify as design components. This was partially automated by matching titles from the metadata table to phrases. However, with some more famous games, reviewers tended to refer to the games by parts of the whole title, making it difficult to match and requiring more hands on work on my part.

4.3.2 Noun Phrases

The first step in building noun phrases, as well the 3 subsequent datasets, was to retrieve natural language processed results from the Stanford Parser. Version 1.5 of the parser was used, employing the logic and syntax labeling as described by de Marneffe, MacCartney and Manning in their paper "Generated Typed Dependency Parses From Phrase Structure Parses"(de Marneffe, MacCarntey, & Manning, 2006). The results were returned formatted so as to identify paired terms from each sentence by their Stanford Parser Label, as is shown in figure 6. This labeling includes both single pair and multi-word relationships, which can be used for reconstituting parts of or the entirety of the original sentence based on the term labeling hierarchy. It should also be noted that the

Stanford Parser results were processed for additional punctuation removal and shifted to lower case before further manipulation took place.

docid	paraid	sentid	Label	term1	termid1	term2	termid2
1	1	2	pobj	with	6	mechanics	12
1	1	2	cc	simple	7	but	9
1	1	2	conj	simple	7	responsive	10
1	1	2	cc	mechanics	12	and	13
1	1	2	amod	mechanics	12	simple	7
1	1	2	nn	mechanics	12	control	11

Figure 6: An excerpt of output from the Stanford Natural Language Processor, for the original sentence "The EAD-created title was, with simple, but responsive control mechanics and then-gorgeous graphics, as good as snowboarding games got."

Given this type of output from the parser, noun phrases were generated using the hierarchy dependencies produced in the parse results. Terms were selected based on their classification as either noun compound modifiers or adjectival modifiers, although a number elements were allowed some of the time. These number elements were included as adjectival modifiers in cases where they contained the terms "2D" or "3D" as they were actually adjectives that the parser was unable to identify. These datasets were then combined to form the basis for phrase construction

Before beginning to build up the actual noun phrases, though, a Porter Stemmer was used to generate a stemmed vocabulary of base nouns, largely to avoid fragmentation of the dataset between plural and non-plural versions of noun phrases. This list of stemmed terms was then rejoined with the core vocabulary for phrase building.

By restricting to the parser labels nouns, adjectival modifiers and number elements and utilizing the “connect by” clause provided by Oracle9i and later versions it was possible to recursively traverse the grammar hierarchy through all the modifiers for a single base noun. As can be seen in figure 7, using this method also has the added benefit of generating all the possible variations of modifiers on the base and maintaining the original word order, thus allowing for easier matching later on in cases where a reviewer might feel the need for a few extraneous adjectives.

offline multiplayer	role-playing mechanics
two-player multiplayer	role-playing game mechanics
split-screen multiplayer	role-playing game battle mechanics
offline two-player multiplayer	role-playing mechanics
two-player split-screen multiplayer	role-playing game mechanics
offline two-player split-screen multiplayer	role-playing game battle mechanics

Figure 7: Two example of phrase variations generated by using the “connect by” clause.

4.3.3 Noun Phrases without Game Titles

This section builds upon the data generated in the previous noun phrase text representation, also adding in a series of additional steps for removing game titles. The reason for removing titles mainly is due to the fact that they are not very generalizable and have a high probability of only appearing in association with the region where the game was developed, which could artificially skew classification results. Since this representation begins under the assumption that a list of noun phrases has already been generated, game titles must be matched as phrases against the noun phrases. This is further complicated by the phrase variations generated by the connect by clause, resulting

in the need to match variations of noun phrases against all the possible variations of game titles. In order to do this in the absence of a title grammar hierarchy to traverse, a simple Java application was written to construct all the possible multi-word variations on a game title and maintain the original word order.

Game title variations were then matched against the lists of noun phrases, both as exact matches and substrings. In order to reduce the search space further, the game title variations were only matched against terms present in the review from which the title originally came. Although game titles do occasionally show up in reviews of other games when the author is making direct comparisons, this is not all that common a practice and generally only applies to what could be referred to as paradigm defining games, very popular games that strongly represent a genre.

4.3.4 Noun Phrases with Game Title Removal and Verbs

Building upon the noun phrase representation without game titles, the main addition here is the inclusion of verbs. These are selected via the parser relationship labels for nominal subject, passive nominal subject, and auxiliary, as these represent the primary verbs in a subject verb sentence structure. All verbs were stemmed using a Porter Stemmer and then recombined with the noun phrase dataset.

4.3.5 Individual Terms with NLP based Stopword Removal

This representation reverted back to the original cleaned dataset from the Stanford Parser, as no phrases were used. Using a reference guide and examining the data, the primary focus here was to both identify parser labels considered meaningful and those less meaningful (Bies, et al., 1995), which fits generally with the definition of a stopword. For the first pass of stopword removal, the vocabulary was restricted to an extended list of verb labels, focusing on auxiliary, passive auxiliary, clausal subject, passive clausal subject, nominal subject, passive nominal subject, and participle modifier. These remaining terms were then stemmed using a Porter Stemmer and stored.

At this point it was necessary to identify both which other labels contained meaningful data and which term in the parser label pairing would be informative to retain. Not all meaningful words show up in the first half of a term pairing (see figure 6 for an example of the natural language processor formating), especially in the case of modifiers and adjectives, therefore information would be lost without considering both term 1 and term 2. As a result, two lists of Stanford Parser Labels were generated, where the list for term 1 is as follows: adjectival modifier, determiner, direct object, noun compound modifier, possession modifier, and predeterminers. The list for selecting the second half of the term pairing includes: adjectival modifier, auxiliary, passive auxiliary, clausal subject, passive clausal subject, noun compound modifier, participle modifier, predicate.

Terms chosen according to these two rules sets for the first and second half of the pairing were combined into a single listing of terms. With properly labeled stopwords removed

and the term listings combined, this dataset was then merged with the results of the extended verb stemming.

Unfortunately, some terms end up being misclassified as various types of nouns and verbs during natural language processing, so an additional pair of passes are run to cull uninformative terms such as prepositions. The second half of the following types of parser label pairings are deleted from the dataset: coordination, negation modifier, phrasal verb participle, relative, temporal modifier, quantity modifier, prepositional modifier, numeric modifier (where not 3D or 2D), complementizer, copula, marker, auxiliary, and passive auxiliary. Also due to the misclassification, one final pass was needed to eliminate terms which were improperly stemmed, and therefore were not caught by the previous stopword detection.

4.3.6 Weighting

Due to the size of the search space, weighting was performed in order to reduce the vocabulary before performing classification. With the five text representations constructed, TFIDF weighting was then calculated for each term or phrase in order to build the final vocabulary set for classification. All five experiments were essentially handled in the same manner during this phase, and built towards selection of the top terms or phrases by TFIDF weighting per review.

The basic process began by calculating IDF values for each term or phrase in the dataset, using the standard formula $IDF_i = \log(|D|/|\{d_j : t_i \in d_j\}|)$, where D is the total number of reviews, and the denominator is the total number of reviews in which a term occurs at

least once. Term frequency was then calculated for each term or phrase per review, which was then combined with IDF to find TFIDF using the formula $TFIDF_{ij} = Tf_{ij} * IDF_i$. Lastly, the terms with the top ten highest TFIDF values per review were selected and merged together to construct the final vocabulary for each experiment

The only variation from this standard methodology was in the case of the preliminary text windowing study where due to the method of windowing, a much larger set of terms was initially generated. As a result, more time was spent in the examining term frequencies across the entire corpus. In order to reduce the vocabulary size before calculating TFIDF weightings, TF values were used to eliminate both the extremely rare and extremely commonly occurring terms and phrases, based off of Zipf's Law and Luhn Cut-offs (Rijsbergan, 1979). This was especially helpful for reducing the number of rarely occurring multi-word phrases generated during the windowing process. After this point, the process was the same as that used for the other four representations aside from some differences in the exact method of achieving the results as the code was rewritten for the current project.

4.4 Data Mining

As was the case with weighting during the transformation phase, the five experiments followed essentially the same sequence of steps for data mining. After having built the vocabularies of highest TFIDF terms or phrases per review, the terms were then joined back in with the review to region mapping creating during the preprocessing phase.

Each experiment's dataset was built into a series of classifiers, including one binary classifier for each region as well as a single classifier covering all regions at once. The binary classifier training datasets consisted of the list of top terms per review along with a binary field indicating whether the review the term was originally extracted from was in the region being tested by this particular classifier. Whereas the "all regions" classifiers matched the same set of terms with the region code associated with the review from which the term was extracted.

Classification was performed using the Oracle Data Mining Tool version 10.2.0.4.1 to build and test models using the classifiers constructed for each experiment. Even though all available classification models were tested, including Naïve Bayes, Support Vector Machine, Adaptive Bayes Network, and even Decision Tree, only Naïve Bayes and Support Vector Machine were a reasonable match for this dataset. All four noun and verb based experiments were processed using Support Vector Machine models, and although a few datasets in each experiment were also attempt with Naïve Bayes, the results are not shown as the Support Vector Machine out performed the other classifier by a large margin in all cases.

The windowing experiment followed much the same pattern, where initially both Naïve Bayes and Support Vector Machines were employed in the classification process, but ultimately the SVM gave better results and became the primary tool. Classification involved the use of one multi-class and five binary classifiers, one for each region where the set of regions was essentially the same except with Japan separated from the Asian

region. In both the original experiment and the current, too few reviews were available for the regions Asia and Other to reliably build models, therefore they were merged and excluded in the current work, respectively.

After running the support vector machine classifiers and testing them using the Oracle Data Mining Tool, the application returns various statistics outlining the performance of a particular model over the selected dataset. This information is reported as results in section 5.

4.5 Summary of Experiments

Five experiments were performed using Support Vector Machine classifiers. Figure 8 shows the labeling of experiments as they are broken down by classifier and feature representation. Although 5 additional experiments were performed using Naïve Bayes classifiers, unfortunately the results were poor and would not have contributed meaningful data to this paper.

Text Representation	**Support Vector Machine**
Individual Terms	2
Text Windowing	1
Noun Phrases	3
Noun Phrases without Game Titles	4
Noun Phrases with Verbs	5

Figure 8: Listing of experiments by usage of feature selection and classifiers.

5 Results

5.1 Predicting Region

Figure 9 shows the predictive performance for experiments 1 through 5, measured as the percentage of times that the classification model correctly predicted the region. Comparison of each of the experiments' results as performed using binary classifiers can be seen in figure 10 and using multi-classifiers in figure 11. Comparative predictive confidence between binary classifiers is also shown in figure 12, where predictive confidence is defined as the rate of prediction error of the model compared to the baseline model with 50% error rate (Oracle, 2007). So a predictive confidence of 0% would be no better than the baseline, and 100% would mean that there were no prediction errors.

A trend in all the text representations' binary classifiers indicates that video games from Asia have the highest predictive confidence, a 33% average, whereas games from North America are lowest at 20.96%, with Europe and Australia inbetween 29.47%. A similar trend occurs in the prediction rates, with Asia having the best average at 65.32%, but with this metric North America performs better than Europe and Australia, 62.99% to 59.89%.

Results show that the multi-classifiers out perform the average of the binary classifiers in predictive confidence in all cases except for experiment 1, due to the disproportionately poor performance of the North American classifier. This may have been due to the region having almost twice as many reviews as other regions, which could also represent a broader range in the types of games developed in the region.

Experiment 2: Noun Phrases

	Asia	**Europe/ Australia**	**North America**
In Region	87.26%	87.12%	31.15%
Not In Region	48.81%	44.26%	87.88%
Multi-class	*42.43%*	*43.48%*	*78.40%*

Experiment 3: Noun Phrases, Game Title Removal

	Asia	**Europe/ Australia**	**North America**
In Region	87.88%	87.66%	29.77%
Not In Region	47.16%	42.76%	88.49%
Multi-class	*46.49%*	*37.85%*	*78.85%*

Experiment 4: Noun Phrases, Verbs

	Asia	**Europe/ Australia**	**North America**
In Region	88.06%	87.78%	28.83%
Not In Region	45.72%	40.48%	87.78%
Multi-class	*40.37%*	*40.09%*	*78.85%*

Experiment 5: Individual Terms, Stopwords

	Asia	**Europe/ Australia**	**North America**
In Region	93.14%	90.54%	52.78%
Not In Region	35.01%	37.08%	76.57%
Multi-class	*35.88%*	*39.52%*	*80.36%*

Experiment 1: Text Windowing

	Asia	**North America**	**Europe/ Australia**	**Japan**	**Other**
In Region	37.76%	45.94%	54.13%	56.71%	44.87%
Not In Region	94.19%	66.24%	61.52%	63.41%	98.63%
Multi-class	*16.93%*	*27.89%*	*19.59%*	*35.60%*	*44.87%*

Figure 9: Each table contains the percentage of times in which region was correctly predicted using models generated for each experiment using support vector machine classifiers. The rows labeled "In Region" and "Not In Region" represent the results from binary classifiers built for each region and the "multi-class" rows are from classifiers built to consider all 3 regions (or 5 in the case of the preliminary experiment) at once.

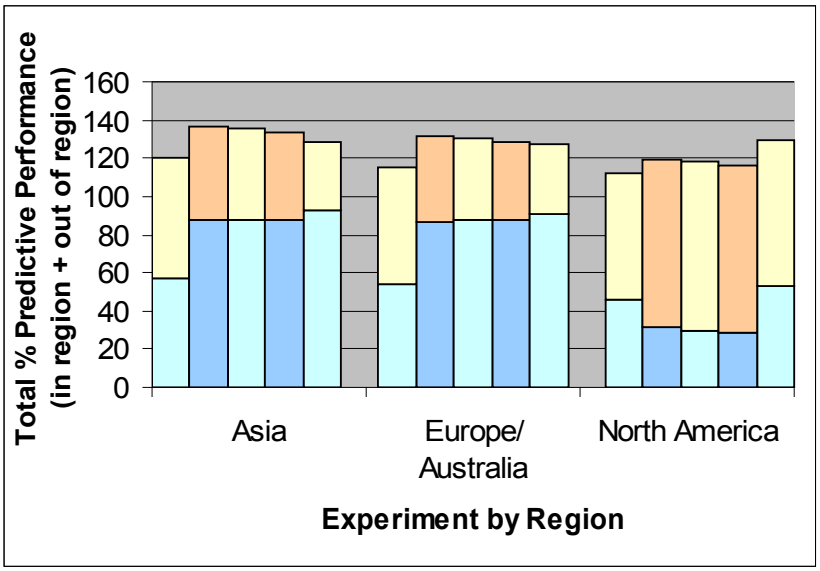

Figure 10: Binary Classifiers' percentage of phrases correctly predicted by region per experiment using support vector machines. Each cluster of columns represents one of the three regions predicted, where each column represents a binary classifiers from a single experiment, order left to right as experiment 1, 2, 3, 4, 5. The bottom portion of the column indicates the % correctly classified as "in region", the top portion is the % correctly classified as "out of region". The total height of each bar indicates the combined percentage of correctly predicted cases both inside and outside of the region.

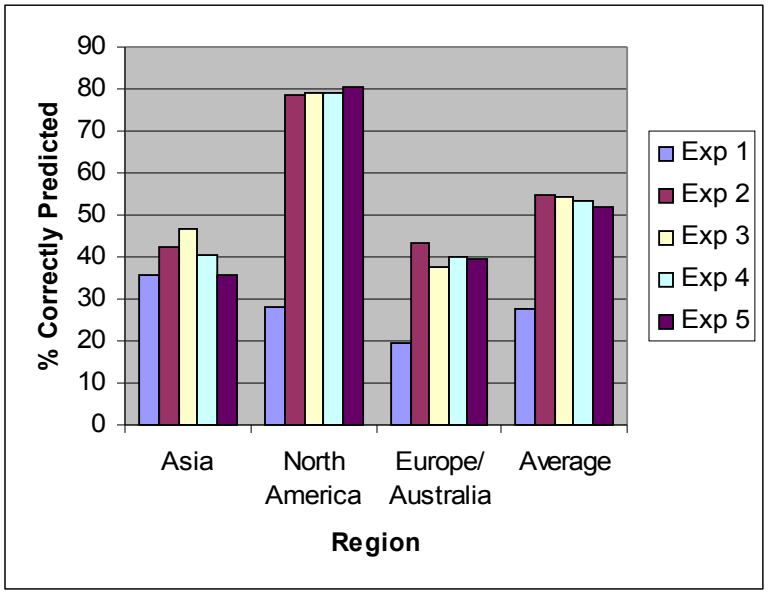

Figure 11: Multi-classifiers' percentage of phrases correctly predicted by experiment, clustered by region, where each column represents a separate experiment using support vector machines. The fourth cluster represents the average result per experiment across all regions.

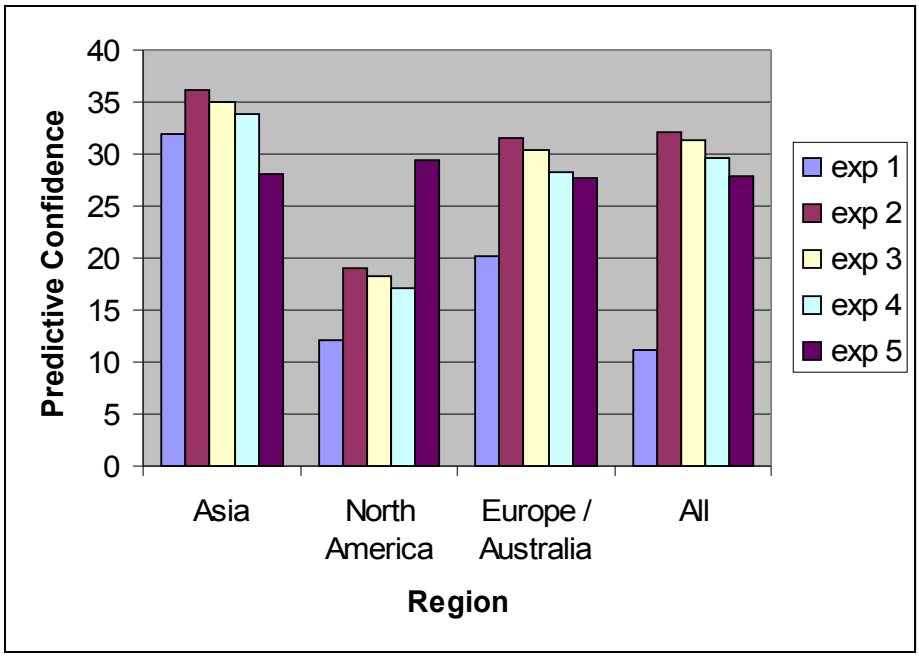

Figure 12: Predictive confidence of both binary classifiers and multi-class classifiers for experiments 1 through 5 by region using support vector machines.

The predictive performances of the four non-windowed text representations were similar, although it is not surprising that experiment 2, as is shown in figure 13a, would develop the most accurate prediction model given the inclusion of game titles. It seems reasonable to assume that titles of games should correlate fairly directly with the regions in which the game's developer or publisher is based. The problem with leaving the titles, however, is that the model may not generalize. Alternately, figure 13b compares experiments 3 and 4 to show that including stemmed verbs with noun phrases has a minor negatively impact on prediction rates.

Overall, based on the results shown in figure 13c the experiments identifying nouns and verbs outperformed the original windowing phrase generation in all three main regions. The differences were especially pronounced for the multi-class classifiers, where the

prediction rates for every region monotonically increased. For the binary classifiers, the improvement was not quite as even, but there was a small net gain in overall predictive effectiveness.

Full classification results for experiments 2 through 5 are listed in figure 14a through 14p in the form of prediction test matrices for each experiment using Support Vector Machine classifiers. Results are grouped according to experiment, with the multi-class tables preceding binary results.

Figure 13a-c: The average difference between the percentage of correctly predicted regions based on text representation using Support Vector Machine classifiers.

Experiment 2 versus Experiment 3

	Δ Percent Correctly Predicted			**Sum Δ**	**Average Sum**	**Classifier Avg**
	Asia	**Europe/ Australia**	**North America**			
In Region	-0.62%	-0.54%	1.38%	0.22%	0.07%	0.46%
Not In Region	1.65%	1.50%	-0.61%	2.54%	0.85%	
Multi-class	*-4.06%*	*5.63%*	*-0.45%*	*1.12%*	*0.37%*	*0.37%*

Figure 13a

Experiment 3 versus Experiment 4

	Δ Percent Correctly Predicted			**Sum Δ**	**Average Sum**	**Classifier Avg**
	Asia	**Europe/ Australia**	**North America**			
In Region	-0.18%	-0.12%	0.94%	0.64%	0.21%	0.84%
Not In Region	1.44%	2.28%	0.71%	4.43%	1.48%	
Multi-class	*6.12%*	*-2.24%*	*0.00%*	*3.88%*	*1.29%*	*1.29%*

Figure 13b

Experiment 1 versus the Average of Experiments 2 to 5

	Δ Percent Correctly Predicted			**Sum Δ**	**Average Sum**	**Classifier Avg**
	Asia	**Europe/ Australia**	**North America**			
In Region	32.38%	42.34%	-18.50%	56.21%	18.74%	5.92%
Not In Region	-19.24%	-25.10%	23.66%	-20.67%	-6.89%	
Multi-class	*5.69%*	*12.35%*	*59.53%*	*77.56%*	*25.85%*	*25.85%*

Figure 13c

Figure 14a-p: Prediction matrices of experiments 2 though 9 as they are represented by both multi-class and binary classifiers.

Experiment 2 Multi-class					
Row	Asia	North America	Europe/ Australia	Total	Correct %
Asia	46,668	48,543	14,778	109,989	42.42
North America	22,922	178,481	26,256	227,659	78.39
Europe/ Australia	10,768	59,904	54,356	125,028	43.47
Total	80,358	286,928	95,390	462,676	
Correct %	58.08	62.2	56.98		

Figure 14a

Experiment 2 Binary - Asia				
Row	Out	In	Total	Correct %
Out	308,096	44,968	353,064	87.26
In	56,300	53,689	109,989	48.81
Total	364,396	98,657	463,053	
Correct %	84.55	54.42		

Figure 14b

Experiment 2 Binary - North America				
Row	Out	In	Total	Correct %
Out	73,320	162,074	235,394	31.14
In	27,591	200,068	227,659	87.88
Total	100,911	362,142	463,053	
Correct %	72.65	55.24		

Figure 14c

Experiment 2 Binary - Europe and Australia				
Row	Out	In	Total	Correct %
Out	294,473	43,552	338,025	87.11
In	69,444	55,584	125,028	44.45
Total	363,917	99,136	463,053	
Correct %	80.91	56.07		

Figure 14d

Experiment 3 Multi-class					
Row	Asia	North America	Europe/ Australia	Total	Correct %
Asia	53,951	53,121	8,988	116,060	46.48
North America	26,054	187,601	24,264	237,919	78.85
Europe/ Australia	17,000	64,536	48,863	130,399	37.47
Total	97,005	305,258	82,115	484,378	
Correct %	55.62	61.45	59.51		

Figure 14e

Experiment 3 Binary - Asia				
Row	Out	In	Total	Correct %
Out	324,019	44,678	368,697	87.88
In	61,324	54,736	116,060	47.16
Total	385,343	99,414	484,757	
Correct %	84.08	55.06		

Figure 14f

Experiment 3 Binary - North America				
Row	Out	In	Total	Correct %
Out	73,493	173,345	246,838	29.77
In	27,389	210,530	237,919	88.48
Total	100,882	383,875	484,757	
Correct %	72.85	54.84		

Figure 14g

Experiment 3 Binary - Europe and Australia				
Row	Out	In	Total	Correct %
Out	310,620	43,738	354,358	87.65
In	74,642	55,757	130,399	42.75
Total	385,262	99,495	484,757	
Correct %	80.62	56.04		

Figure 14h

Experiment 4 Multi-class					
Row	Asia	North America	Europe/ Australia	Total	Correct %
Asia	48,916	57,169	15,081	121,166	40.37
North America	24,502	197,284	28,417	250,203	78.84
Europe/ Australia	11,875	72,261	56,307	140,443	40.09
Total	85,293	326,714	99,805	511,812	
Correct %	57.35	60.38	56.42		

Figure 14i

Experiment 4 Binary - Asia				
Row	Out	In	Total	Correct %
Out	344,385	46,698	391,083	88.05
In	65,764	55,402	121,166	45.72
Total	410,149	102,100	512,249	
Correct %	83.96	54.26		

Figure 14j

Experiment 4 Binary - North America				
Row	Out	In	Total	Correct %
Out	75,542	186,504	262,046	28.82
In	29,062	221,141	250,203	88.38
Total	104,604	407,645	512,249	

Correct %	72.21	54.24		

Figure 14k

Experiment 4 Binary - Europe and Australia				
Row	Out	In	Total	Correct %
Out	326,386	45,420	371,806	87.78
In	83,591	56,852	140,443	40.48
Total	409,977	102,272	512,249	
Correct %	79.61	55.58		

Figure 14l

Experiment 5 multi-class					
Row	Asia	North America	Europe/ Australia	Total	Correct %
Asia	13,910	19,260	5,594	38,764	35.88
North America	7,107	63,377	8,384	78,868	80.35
Europe/ Australia	4,624	22,147	17,492	44,263	39.51
Total	25,641	104,784	31,470	161,895	
Correct %	54.25	60.48	55.58		

Figure 14m

Experiment 5 Binary - Asia				
Row	Out	In	Total	Correct %
Out	128,202	9,436	137,638	93.14
In	26,965	14,529	41,494	35.01
Total	155,167	23,965	179,132	
Correct %	82.62	60.63		

Figure 14n

Experiment 5 Binary - North America				
Row	Out	In	Total	Correct %
Out	48,005	42,945	90,950	52.78
In	20,664	67,518	88,182	76.56
Total	68,669	110,463	179,132	
Correct %	69.91	61.12		

Figure 14o

Experiment 5 Binary - Europe and Australia				
Row	Out	In	Total	Correct %
Out	117,589	12,282	129,871	90.54
In	30,997	18,264	49,261	37.07
Total	148,586	30,546	179,132	
Correct %	79.13	59.79		

Figure 14p

5.2 Region versus Release Date as a Classifier

To evaluate the effectiveness of region as a classifier compared with other potential prediction targets, we conducted an additional experiment built using the release date of a game as the prediction target. Video games are a technology driven industry in which features traditionally follow the types of hardware that are commercially feasible, thus the types of features identified and valued by reviewers, both as the features become available and as they become the norm, would correlate over time.

A series of classifiers were built for this comparison using the text representations described in section 4.3 and substituting the release date grouped by year for prediction rather than the region. For this comparison only multi-class classifiers were used, performed using support vector machine classifiers. The full prediction matrices for each experiment are listed in figures 17a-d.

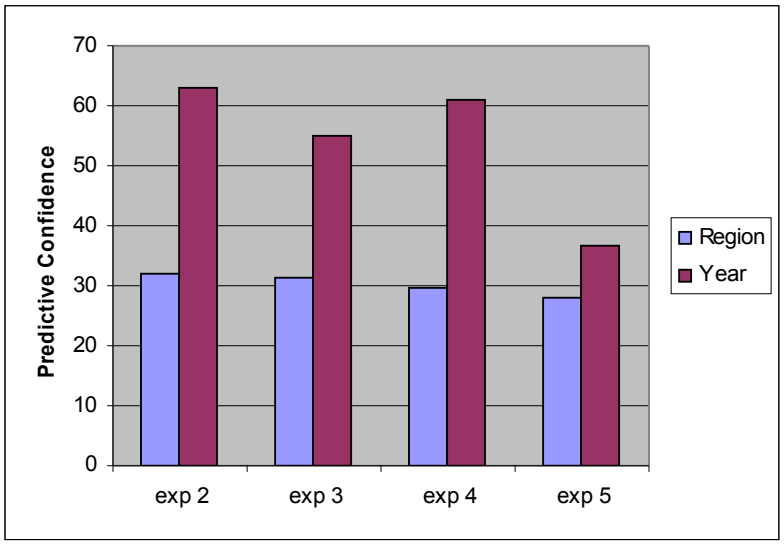

Figure 15: Predictive confidence of region versus release date across all four noun or verb based experiments.

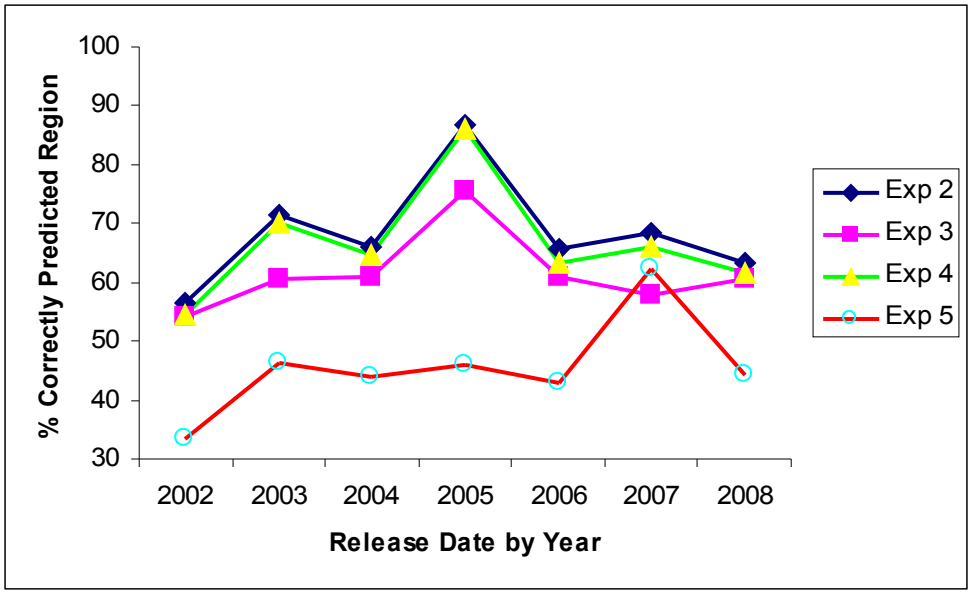

Figure 16: Percentage of correctly predicted phrases by multi-class classifiers using release date by year as the prediction target.

By comparing the predictive confidence of release date versus development region as is described in figure 15, release date clearly shows a much stronger correlation between the features and target. Although a direct comparison between the prediction rates is not possible because the individual targets are not equivalent, the average correct prediction rate for experiments 2 through 5 across all categories was 60.51% versus 53.55% for region.

Figure 16 compares the prediction rates for release date only, plotting each of the four selected experiments against each other. More variation in the prediction rate between experiments is visible with release date than was seen with region as the prediction target. Experiments 2 and 4 were extremely close across all years, even though experiment 3 does not seem to perform quite as well. This is interesting because this variation indicates that for release date, the inclusion of verbs with noun phrases has a positive

effect on correct prediction, whereas the opposite was true for region. More in line with the region results, here it seems that the use of individual terms over phrase generation also had a noticeable negative impact on both the confidence and prediction rates for the release date, although the difference is more pronounced than was seen for region.

Figure 17a-d: Prediction matrices for experiments 2 through 5 using release date grouped by year as the prediction target.

Experiment 2 Multi-class - Release Date							
Row	01/01/02	01/01/03	01/01/04	01/01/05	01/01/06	01/01/07	01/01/08
01/01/02	28,626	3,743	2,176	10,053	1,925	2,428	1,595
01/01/03	2,121	54,296	2,064	13,254	1,337	1,761	1,140
01/01/04	1,253	3,862	47,126	13,649	1,469	2,956	1,071
01/01/05	1,411	2,039	2,620	78,463	2,259	2,323	1,502
01/01/06	1,293	2,490	2,509	13,110	48,238	4,558	1,437
01/01/07	1,377	2,735	1,702	13,577	2,618	52,432	2,374
01/01/08	895	1,855	1,766	10,715	2,342	4,576	38,249
Total	36,976	71,020	59,963	152,821	60,188	71,034	47,368
Correct %	77.41	76.45	78.59	51.34	80.14	73.81	80.74

Figure 17a

Experiment 3 Multi-class - Release Date							
Row	01/01/02	01/01/03	01/01/04	01/01/05	01/01/06	01/01/07	01/01/08
01/01/02	28,971	3,054	3,005	10,791	2,529	2,771	2,446
01/01/03	4,206	50,075	4,123	14,470	3,088	3,390	3,125
01/01/04	3,442	3,651	48,286	14,342	3,169	3,228	3,108
01/01/05	3,897	3,585	4,900	75,076	4,158	4,059	3,513
01/01/06	3,503	2,896	3,543	13,255	47,639	3,940	3,470
01/01/07	3,986	3,287	4,093	14,004	4,133	46,147	4,274
01/01/08	2,711	2,359	2,707	10,843	2,683	3,512	38,396
Total	50,716	68,907	70,657	152,781	67,399	67,047	58,332
Correct %	57.12	72.67	68.34	49.14	70.68	68.83	65.82

Figure 17b

Experiment 4 Multi-class - Release Date							
Row	01/01/02	01/01/03	01/01/04	01/01/05	01/01/06	01/01/07	01/01/08
01/01/02	27,926	3,679	2,329	10,601	2,226	2,808	1,837
01/01/03	2,142	52,739	1,979	13,552	1,597	1,752	1,308
01/01/04	1,289	3,786	46,228	14,229	1,576	2,942	1,251
01/01/05	1,497	2,097	2,583	76,309	2,295	2,237	1,499
01/01/06	1,542	2,635	2,711	13,461	46,283	4,734	1,582
01/01/07	1,733	3,039	1,973	14,365	2,770	50,940	2,453
01/01/08	944	1,975	1,946	11,041	2,388	4,496	36,573
Total	37,073	69,950	59,749	153,558	59,135	69,909	46,503
Correct %	75.33	75.40	77.37	49.69	78.27	72.87	78.65

Figure 17c

Experiment 5 Multi-class - Release Date							
Row	01/01/02	01/01/03	01/01/04	01/01/05	01/01/06	01/01/07	01/01/08
01/01/02	6,279	2,461	1,676	1,342	1,495	4,272	1,197
01/01/03	527	9,607	1,936	1,497	1,560	4,387	1,246
01/01/04	410	2,186	8,866	1,657	1,513	4,398	1,209
01/01/05	401	2,548	1,897	10,733	1,625	4,811	1,366
01/01/06	470	2,546	1,821	1,636	9,831	5,056	1,483
01/01/07	556	2,735	2,037	1,702	2,177	18,123	1,742
01/01/08	254	1,546	1,233	1,063	1,155	4,366	7,607
Total	8,897	23,629	19,466	19,630	19,356	45,413	15,850
Correct %	70.57	40.66	45.55	54.68	50.79	39.91	47.99

Figure 17d

6 Future Work

The results of these experiments indicate some possible directions in which this research could be taken in the future.

1. It has been observed that positive reviews map much more closely with the sales of video games than has been observed in other media such as films or music (Schiesel, 2007). Utilizing the results and datasets used in this study, it could be investigated to see if a relationship existed between video game review contents and game sales, as well as between predicted game region and sales in each region.
2. The extent of the effect caused by retaining duplicate cross platform reviews also warrants investigation, such as in cases where the same review text was used to describe both the Playstation 3 and Xbox 360 versions of a game. Even though in terms of sales and development time these two products would be considered distinct, duplication of the terms in the review could have a skewing effect towards regions that tend to release more cross platform titles.
3. Consider the application of this dataset in how it could be used as a means of creating an extensive growing vocabulary of game design components. Similar to the technique used in Comeau and Wilbur's study of how to create an expanding spell checker by interpreting correct spelling probabilistically from medical articles (Comeau & Wilbur, 2003), a live vocabulary could potentially be extracted from review texts.
4. As a prediction target, released date demonstrated promising classification results, indicating that the date could potentially serve as the basis for a study similar to

the current one. Other targets could also be considered, such as the platform on which games are released.

7 Conclusion

This project measured predictive performance of five text representations, (individual terms, text windowing, noun phrases, and noun phrases with verbs) using Support Vector Machine classifiers to predict a video game's development region based on reviews. Reviews were selected from two websites, www.gamespot.com and www.ign.com, to act as descriptive texts for the original game products. Using this data, five text representations were built for use in creating both binary and multi-class classifiers for building region prediction models.

The noun and verb features outperformed simple windowing, and showed the largest positive effect on prediction rates of any of the factors explored. When framed as a multi-classifier problem there was an average of 25.85% improvement in prediction rate across the three major regions, as well as a less pronounced improvement in the binary classifiers of 5.92% on average, combining the results from both “in region” and “out of region” predictions.

The four methods of feature representations that involved nouns and verbs achieved similar prediction rates, and therefore the type of representation seems to be a minor role as long as the representation utilizes sentence structure. When framed as a binary classifier, the average predictive confidence by text representation was: text windowing, 21.42%, noun phrases, 28.89%, noun phrases without game titles, 27.91%, noun phrases with verbs, 26.42%, and individual terms with stopword removal, 28.38%. Framed as a multi-class classifier, the predictive confidence by representation was: text windowing,

11.22%, noun phrases, 32.15%, noun phrases without game titles, 31.40%, noun phrases with verbs, 29.66%, and individual terms with stopword removal, 27.88. Inclusion of verbs in the feature set resulted in a minor decrease in the average prediction rate of 0.84% for binary classifiers, and 1.29% for multi-class, whereas retaining game titles had a small positive effect of 0.46% for binary and 0.37% for multi-class over the noun phrases representation without titles.

The predictive confidence of region increased by 26.57%, and 28.49% if text windowing is excluded, over the baseline classifier. However, results show that text mining has a higher predictive accuracy when predicting release date as a multi-classifier. Release date had an average predictive confidence of 53.93% across noun and verb based text representations, whereas region over the same representations has 30.27% confidence. This does not, however, discount the use of text mining in the prediction of a video game's origin, it simply gives an indication of the relative strength of it as a predictor compared to other possible factors.

This paper represents the first study of using review contents to automatically classify video games according to the region in which they were produced. Findings from the five selected text representations yielded positive results, indicating the possibility of numerous future directions for research in this area.

8 Acknowledgments

Thanks to Dr. Catherine Blake for taking the time to advise me in undertaking this project, especially in regards to phrase generation and use of natural language processors. Thanks to her and the other faculty at SILS for putting up with my various video game related projects over the years.

Thanks to my parents for allowing their children to play more video games than they probably should have, and my family in general for their support.

This research was done using resources provided by the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

9 References

Alvarez, J., Djaouti, D., Ghassempouri, R., Jessel, J., and Methel, G. (2006). Morphological study of the video games. *In Procedings of the 3rd Australasian Conference on interactive Entertainment*. ACM International Conference Proceeding Series, 207, 36-43. Murdoch University, Murdoch University, Australia.

BBC News (2007). Game profitability 'under threat'. *BBC News*. Retrieved on April 18, 2008, from http://news.bbc.co.uk/2/hi/technology/6397527.stm

Bies, Ann, et al. (1995). Bracketing Guidelines for Treebank II Style. *Penn Treebank Project*. Retrieved October 18, 2008, from http://www.ldc.upenn.edu/Catalog/desc/addenda/LDC1999T42/PRSGUID1.PDF

Bragge, J., Storgårds, J. (2007). Profiling Academic Research on Digital Games Using Text Mining Tools. *Situated Play*, 714-729. The University of Tokyo, Tokyo, Japan.

Comeau, D., Wilbur, J. (2004). Non-word identification or spell checking without a dictionary. Journal of the American Society for Information Science and Technology, 55(2), 169-177.

console. (2008). In *Merriam-Webster Online Dictionary*. Retrieved November 14, 2008, from http://www.merriam-webster.com/dictionary/console

Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *In Proceedings of the 12th international Conference on World Wide Web*. ACM, New York, NY.

Ding, J., et al. (2002). Mining MEDLINE: Abstracts, sentences, or phrases? *Pacific Symposium on Biocomputing*. Kaua'i, HI.

ESA (2008). 2008 Sales, Demographic and Usage Data: Essential Facts About the Computer and Video Game Industry. *The Entertainment Software Association.* Retrieved on November 2, 2008, from http://www.theesa.com/facts/pdfs/ESA_EF_2008.pdf

Fayyad,U., Piatetsky-Shapiro,G., Smyth,P.(1996) The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM,* 39(11).

Ip, B. & Jacobs, G. (2004). Territorial lockout – an international issue in the videogames industry. *European Business Review*, 16(5), 511-521.

Keegan, M. (1997). A Classification of MUDs. *Journal of MUD Research*, 2(2). Retrieved November 13, 2007, from http://www.brandeis.edu/pubs/jove/HTML/v2/keegan.html

Leino, O. (2007). Emotions about the Deniable/Undeniable: Sketch for a Classification of Game Content as Experienced. *Situated Play*, 113-120. The University of Tokyo, Tokyo, Japan.

de Marneffe, M., MacCartney, B., Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

Navarro, A (2005). Peter Jackson's King Kong: The Official Game of the Movie for Playstation 2 Review. *gamespot.com*. Retrieved October 23, 2007, from http://www.gamespot.com/ps2/action/kingkong/review.html

Nicolas, Esposito (2005). A Short and Simple Definition of What a Videogame Is. *Changing Views: Worlds in Play*. University of Vancouver. Vancouver, CA.

NPD Group, Inc. (2008). NPD Market Research. NPD Group, Inc. http://www.npd.com/

Oracle (2007). Oracle Data Mining Frequently Asked Questions. Retrieved November 14, 2008, from http://www.oracle.com/technology/products/bi/odm/odm_10g_faq.pdf

Otobe, I. (2007). Innovations in the video game industry. *Autumn 2007-2008 Seminar/Public Lecture Series: Innovation Systems and Processes in Asia*. Stanford University.

PC Magazie. Definition of: video game console. *PC Magazine*. Retrieved on November 13, 2008, from http://www.pcmag.com/encyclopedia_term/0,2542,t=video+game+ console&i= 53848,00.asp

Porter, M. (2000). *Porter Stemmer*. http://tartarus.org/martin/PorterStemmer/java.txt

Provo, F (2005). Peter Jackson's King Kong: The Official Game of the Movie for DS Review. *gamespot.com*. Retrieved October 23, 2007, from http://www.gamespot.com/ds/action/kingkong/review.html

van Rijsbergan, C. (1979). Automatic Text Analysis. *Information Retrieval*. Buttersworth. London, UK.

Rajman, M., Besançon, R. (1997). Text Mining: Natural Language Techniques and Text Mining applications. In: *Proceedings of the Seventh IFIP 2.6 Working Conference on Database Semantics (DS 7), Chapam & Hall IFIP Proceedings series*.

Schiesel, S. (2007). Their Thumbs May Be Too Busy to Raise, but Gamers Agree With Critics. *New York Times*. http://www.nytimes.com/2007/07/11/arts/television/11game.html

Sun Microsystems, Inc. (2007). MySQL Stopword List. *MySQL 5.0 Reference Manual*. MySQL AB. Retrieved November 5, 2007, from http://dev.mysql.com/doc/refman/5.0/en/fulltext-stopwords.html

Szalai, G. (2007). Study shows video game industry growth still strong. *Reuters*. http://uk.reuters.com/article/entertainmentNews/idUKN2132172920070621

Weiss, S. et al. (2005). Overview of Text Mining. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Spring Science and Business Media, Inc. New York, NY.

Wingfield, N. (2007). High Scores Matter to Game Makers, Too. *Wall Street Journal*. Santa Monica, CA.

Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, 1(1-2). Kluwer Academic Publishers, MA.

Yi, L., Liu, B., and Li, X. (2003). Eliminating noisy information in Web pages for data mining. *Proceedings of the Ninth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining.* New York, NY. ACM.