Paper EPO-280
Text mining and sentiment analysis on video game user reviews using SAS® Enterprise Miner™

Mukesh Kumar Singh, Nikhil Batti, Dr. Miram Mcgaugh, Oklahoma State University

ABSTRACT

Digital gaming has a history of more than 50 years. The industry started in the late 1960's when the game titles such as Pong, Centipede and Odyssey were introduced to consumer markets. Digital gaming is now a wide spread phenomenon and at least 70% of the US and Europe households say that they play video games using different consoles such as PC, Xbox, PS4, Nintendo etc. It is reported that in 2011, the total revenue of the industry amounted to about 17 billion USD. Each game is reviewed and rated on the internet by users who played the game and the reviews are often contrasting based on the sentiments expressed by the user. Analysing those reviews and ratings to describe the positive and negative factors of a specific game could help consumers make a more informed decision about the game.

In this paper, we will analyse 10,000 reviews and ratings on a scale (1-10) of 200 games culled from two sites: metacritic.com and gamespot.com. We will then build a predictive models to classify the reviews into positive, negative and mixed based on the sentiments of users and develop a score which defines the overall performance of the game so that users get all the required information about a game before purchasing a copy.

INTRODUCTION

METACRITIC and GAMESPOT are the two popular websites for game reviews. Imagine being able to analyze the reviews and understand what exactly the customers liked or disliked. Using text mining we can find the terms that are most commonly used in the reviews and how it affects the game reputation. We can analyze each term in the text and see which other terms is strongly related to. Doing so we can gauge the customer satisfaction or dissatisfaction with the game which may affect the revenue generated by the game either positively or negatively. Using sentiment analysis we can build models on the existing reviews and be able to predict the new reviews as good or bad. Game developers can use this analysis to improve the quality of the movies to meet the expectations of the general audience and to generate maximum revenue.

DATA ACCESS

The data for this research paper contains game reviews taken from www.metacritic.com and www.gamespot.com . We have selected 200 games across all consoles like XBOX, PlayStation, PC and we have extracted all the user reviews using web crawler import.io. We have extracted all reviews and saved it to one excel file. It contains few more than 10,000 reviews.

DATA DICTIONARY

The data contains three variables

Variable	Level	Description
ID	ID	This field represents the unique review number
Game Name	Text	This field represents the name of the game
Review	Text	This filed contains actual game review posted by an user

Table1: Data Dictionary

METHODOLOGY

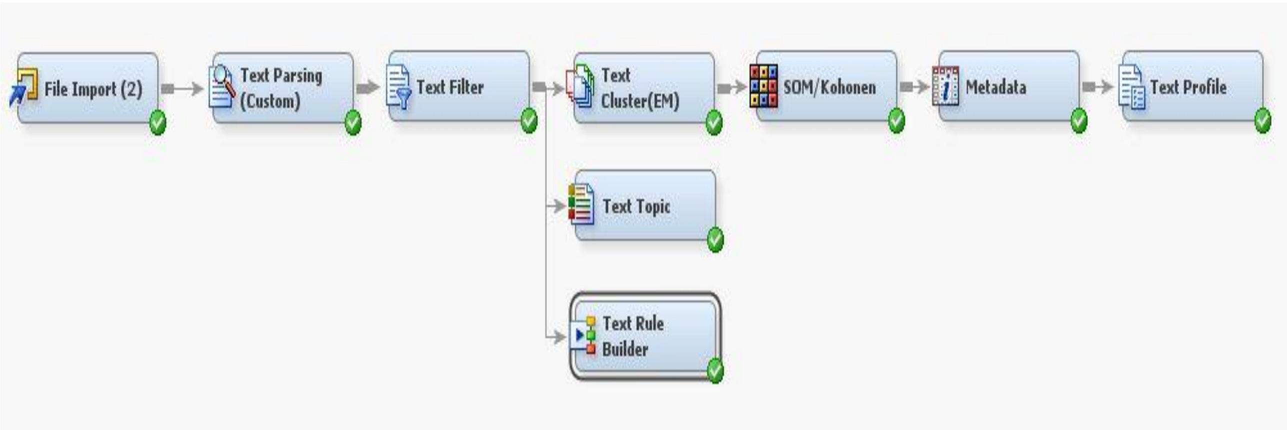

Figure1: Text Mining Process

Text Import

Since the data is available in single csv file, it is imported in SAS® Enterprise Miner™ using the file import node. The import file points to the folder that contains the reviews in CSV documents, with all the reviews in one document.

Text Parsing

After importing the text, the text parsing node is attached to it and a few modifications are made to clean up the unstructured text data. Using the properties panel,

- the 'find entities' option is set to standard,
- the 'detect different parts of speech' option is set to no to be able to represent one word or term as a whole and not have repetitive terms with different parts of speech
- abbr, prop and num parts of speech have been ignored apart from the default options.

The text parsing node also generates the term by frequency document matrix which is used to understand the most frequently occurring term and the number of documents it has occurred in. It is also used to analyze the terms that are rarely used. Ideally the terms that are used moderately are the ones that are the most helpful in exploration and modeling.

Terms

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ game ...		Alpha	2837	599	Y	+	4549	1
+ play ...		Alpha	779	330	Y	+	1273	5
+ good ...		Alpha	744	329	Y	+	1331	6
+ expand ...		Alpha	305	291	Y	+	3365	8
â ...		Alpha	264	258	Y		1256	10
+ story ...		Alpha	500	232	Y	+	1977	13
gameplay ...		Alpha	340	192	Y		1566	15
+ time ...		Alpha	291	157	Y	+	468	19
+ great ...		Alpha	272	153	Y	+	3475	22
graphics ...		Alpha	227	152	Y		3257	23
+ feel ...		Alpha	306	138	Y	+	4200	26
+ buy ...		Alpha	181	127	Y	+	2715	29
+ bad ...		Alpha	227	124	Y	+	4451	30
fun ...		Alpha	217	117	Y		4777	33
+ year ...		Alpha	207	112	Y	+	1499	36
+ mission ...		Alpha	252	107	Y	+	3264	40
+ thing ...		Alpha	184	104	Y	+	3572	41
+ player ...		Alpha	255	103	Y	+	998	42
+ character ...		Alpha	196	99	Y	+	5110	44
+ want ...		Alpha	154	95	Y	+	5153	48
+ people ...		Alpha	161	92	Y	+	5471	50
+ end ...		Alpha	147	90	Y	+	2911	52
+ hour ...		Alpha	159	90	Y	+	3598	52
+ world ...		Alpha	161	89	Y	+	4083	55
amazing ...		Alpha	137	87	Y		625	58
+ little ...		Alpha	134	87	Y	+	3911	58
+ look ...		Alpha	134	86	Y	+	1133	60
+ review ...		Alpha	153	85	Y	+	2150	62
series ...		Alpha	172	84	Y		510	65
+ gear ...		Alpha	193	82	Y	+	4558	67
far ...		Alpha	122	79	Y		1102	68
first ...		Alpha	138	76	Y		2473	70
+ score ...		Alpha	129	76	Y	+	437	70
+ big ...		Alpha	128	75	Y	+	376	72
+ fan ...		Alpha	147	75	Y	+	1085	72

Figure2: Text Parsing Output

The most frequently used terms are game, play, good, gameplay which makes sense since the reviews are for a game. Some of the terms which are misspelt are eliminated later using the text filter node.

Text Filter

The text filter node is added to the text parsing node and is used to eliminate the terms that occur the least number of times in all the documents by manually entering the minimum number of documents it should be present in the properties panel. We can also perform spell check by enabling the option again in the properties panel. Spell check would also suggest the terms that could be potential synonyms. The term ‘betters’ ‘bests’, better, bettered are grouped into one term ‘good’ and so on.

EMWS1.TextFilter2_spellDS

	Parent # Docs	Term	# Docs	Parent	Role	Parent Role	Min Distance	Dictionary	Key	Parent ID
1	31.0	football	2.0	football	PROP_MISC		0.0		4311.0	2.0
2	3.0	sudden	1.0	suddenly			10.0		17.0	7.0
3	4.0	virtual	1.0	virtually			9.0		667.0	29.0
4	41.0	worthy	4.0	worth			6.0		1869.0	34.0
5	4.0	below	2.0	blow			12.0		7562.0	79.0
6	6.0	tony	1.0	ton			10.0		3981.0	102.0
7	6.0	tone	4.0	ton			10.0		3871.0	102.0
8	6.0	toni	1.0	ton			10.0		3893.0	102.0
9	29.0	definite	1.0	definitely			8.0		938.0	120.0
10	3.0	forth	1.0	fourth			10.0		7545.0	153.0
11	15.0	mmos	1.0	mmo	PROP_MISC	PROP_MISC	10.0		1255.0	156.0
12	4.0	skullface	1.0	skull face	PROP_MISC	PROP_MISC	10.0		820.0	203.0
13	5.0	nfifa	1.0	nfifa		PROP_MISC	0.0		462.0	245.0
14	7.0	handy	2.0	hand			8.0		3693.0	256.0
15	3.0	roaster	1.0	roster			8.0		4288.0	259.0
16	13.0	arena	4.0	area			12.0		724.0	268.0
17	8.0	activison	1.0	activision		PROP_MISC	5.0		3422.0	326.0
18	8.0	activision	3.0	activision		PROP_MISC	0.0		5225.0	326.0
19	3.0	footy	2.0	foot			8.0		1880.0	356.0
20	58.0	uncahrted	1.0	uncharted			5.0		5359.0	372.0
21	15.0	metal gears	1.0	metal gear	PROP_MISC	PROP_MISC	6.0		4690.0	426.0
22	15.0	metal ear	1.0	metal gear	PROP_MISC	PROP_MISC	10.0		2610.0	426.0
23	37.0	mgs5	2.0	mgs	PROP_MISC		10.0		4990.0	438.0
24	37.0	mgs3	5.0	mgs	PROP_MISC		10.0		395.0	438.0
25	9.0	innovations	1.0	innovation			3.0		4425.0	448.0
26	123.0	timeâ	2.0	time			8.0		258.0	468.0
27	123.0	timer	1.0	time			8.0		5155.0	468.0

Figure3: Text Filter Spellcheck

Terms

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Pa St
+ game ...		Alpha	Keep	0.003	2837	2841	599	599	1	+
+ be ...		Alpha	Drop	0.000	3204	3208	562	562	2	+
+ not ...		Alpha	Drop	0.000	1327	1332	413	414	3	+
+ great ...		Alpha	Keep	0.141	272	913	153	343	4	+
+ have ...		Alpha	Drop	0.000	960	963	331	331	5	+
+ play ...		Alpha	Keep	0.020	779	781	330	330	6	+
+ good ...		Alpha	Keep	0.072	744	749	329	329	7	+
+ do ...		Alpha	Drop	0.000	798	807	301	304	8	+
+ expand ...		Alpha	Drop	0.000	305	305	291	291	9	+
+ much ...		Alpha	Drop	0.000	645	645	281	281	10	+
â ...		Alpha	Drop	0.000	264	264	258	258	11	
s ...		Alpha	Drop	0.000	740	740	246	246	12	
+ story ...		Alpha	Keep	0.072	500	518	232	233	13	+
just ...		Alpha	Drop	0.000	512	512	233	233	13	
+ gameplay...		Alpha	Keep	0.084	340	361	192	204	15	+
+ get ...		Alpha	Drop	0.000	427	429	197	198	16	+
+ make ...		Alpha	Drop	0.000	349	349	182	182	17	+
n ...		Alpha	Drop	0.000	639	639	168	168	18	
+ no ...		Alpha	Drop	0.000	324	331	165	166	19	+
+ time ...		Alpha	Keep	0.050	291	295	157	158	20	+
+ go ...		Alpha	Drop	0.000	279	279	156	156	21	+
so ...		Alpha	Drop	0.000	262	262	155	155	22	
+ graphics ...		Alpha	Keep	0.076	227	229	152	153	23	+
really ...		Alpha	Drop	0.000	262	262	144	144	24	
very ...		Alpha	Drop	0.000	248	248	140	140	25	
+ feel ...		Alpha	Keep	0.008	306	306	138	138	26	+
+ give ...		Alpha	Drop	0.000	240	244	134	136	27	+
+ one ...		Alpha	Drop	0.000	229	230	135	135	28	+
+ buy ...		Alpha	Keep	0.089	181	181	127	127	29	+
+ all ...		Alpha	Drop	0.000	193	195	123	124	30	+
even ...		Alpha	Drop	0.000	195	195	121	121	31	
fun ...		Alpha	Keep	0.046	217	217	117	117	32	
only ...		Alpha	Drop	0.000	180	180	114	114	33	
+ way ...		Alpha	Drop	0.000	186	186	114	114	33	+
+ year ...		Alpha	Keep	0.053	207	208	112	113	35	+
+ other		Alpha	Drop	0.000	181	182	112	112	36	+

Figure4: Text Filter Output

After running the text filter node, we can see that terms such as be, not, have, do are dropped from the text since they do not contribute towards any meaning in the review. Only words that are related to a game in some way are kept. Text filter is also used to group synonyms together. It can be done by importing a file with all the synonyms or manually by dragging and dropping the terms into each other.

	TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
⊞	game	46925	9180	☑	0.048		Alpha
⊟	good	12635	5410	☑	0.094		Alpha
	good	5579	2936				Alpha
	betters	1	1				Alpha
	bested	1	1				Alpha
	bests	3	2				Alpha
	better	2805	1722				Alpha
	bettering	4	3				Alpha
	goods	11	8				Alpha
	bettered	2	2				Alpha
	best	4229	2472				Alpha
⊞	play	12656	4999	☑	0.103		Alpha
⊞	story	7971	3638	☑	0.135		Alpha
⊞	great	5669	3026	☑	0.151		Alpha
⊞	time	5285	2724	☑	0.164		Alpha

Figure5: Synonyms Grouping

The above screenshot shows the synonyms for the term 'good'. Terms such as 'betters', 'bests', 'goods' are grouped together using the interactive filter viewer.

Concept Links

Concept links can be viewed in the interactive filter viewer from the properties panel of text filter node. It is a type of association analysis between the terms used. Concept links can be created for all the terms that are present in the documents, however it is meaningful to create only for a few important terms. It shows the term to be analyzed in the center and the terms that it is mostly used with as links. The width of the link depicts the strength of association. The wider the link the stronger is the association and the more important it is. Concept links also show how many times the two terms co-exist together in a sentence. A few examples are shown below.

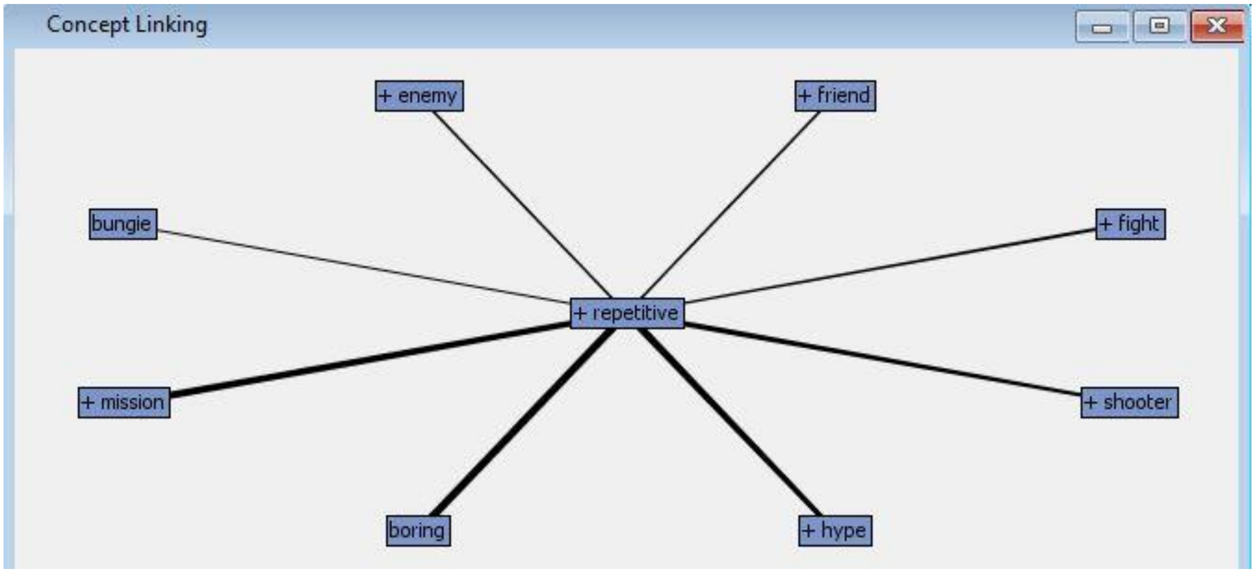

Figure6: Concept link for 'repetitive'

The above concept link is for the term repetitive. The term repetitive is strongly related to a term boring. By which we can say that if a game is repetitive it can be really boring. And also the terms shooter and mission are also strongly related to the term repetitive which infers the shooting games are mostly repetitive.

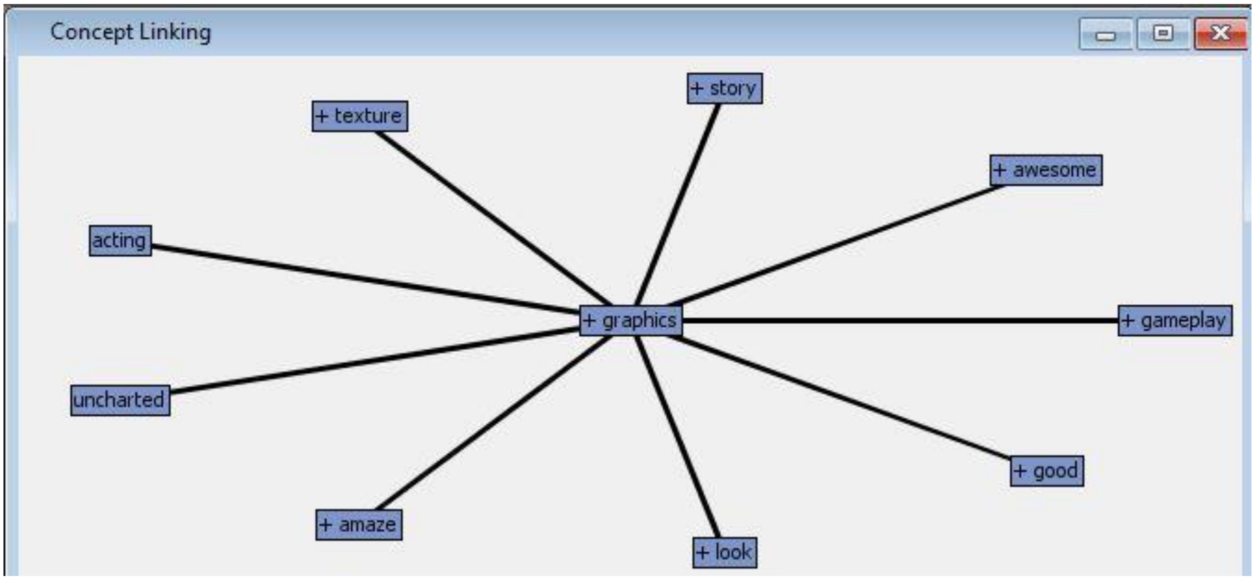

Figure7: Concept link for 'graphics'

Graphics is strongly linked with texture, story and gameplay indicating that good graphics can in the game are also dependent on good story, gameplay. The other terms such as look, texture look are always associated with the graphics of the game and thus focusing on those terms would help the game to be successful.

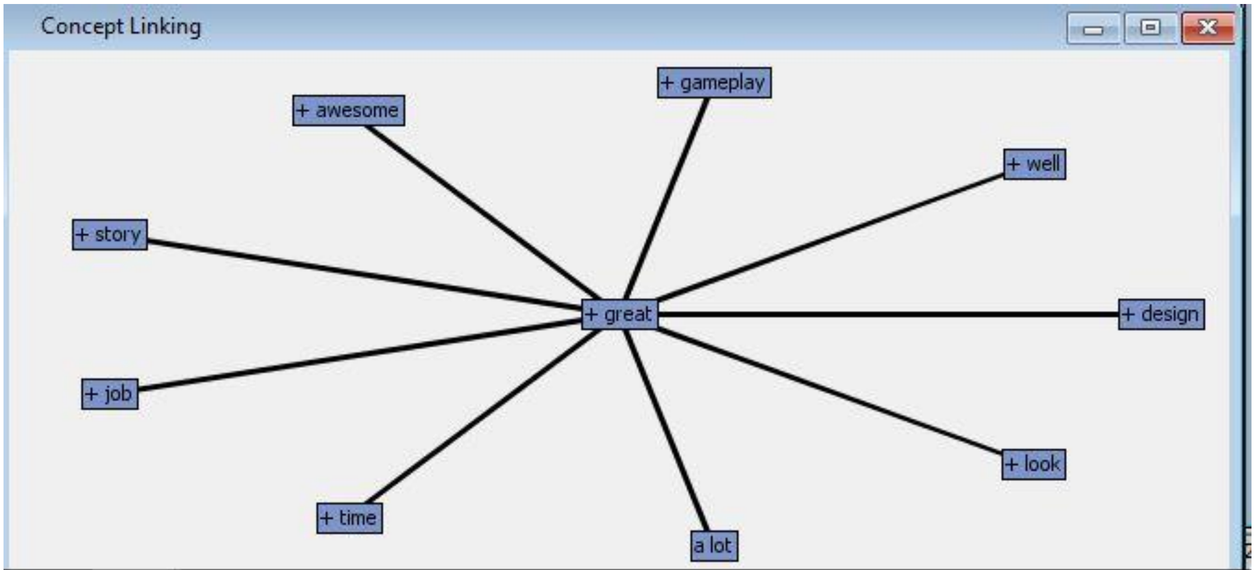

Figure8: Concept link for 'great'

The terms 'gameplay', 'design' and 'story' are strongly associated with great which could be indicating that the game was a great game if it had a good gameplay, story and design. We can say that looks of the game are equally important as the gameplay and the story.

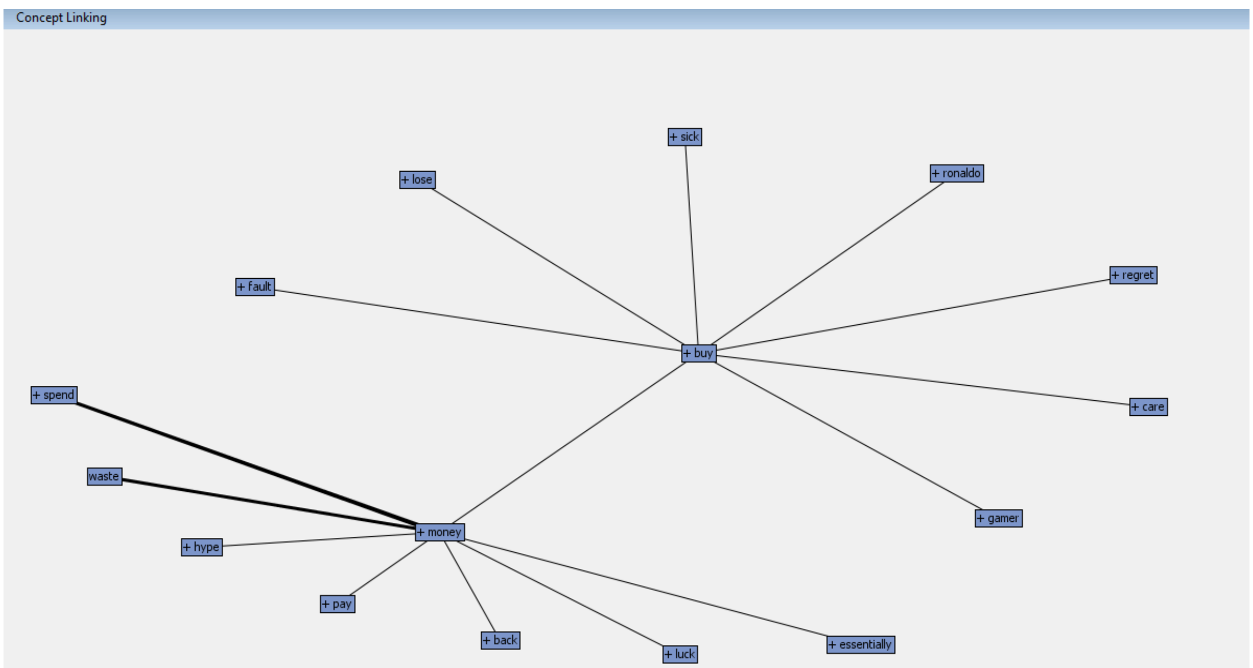

Figure9: Concept link for 'buy'

The term 'buy' is associated with the money and gamer. Few reviews mentioned that the game is not worth buying as it was costly. And many were confused thinking it's just the hype and originally it's not worth buying. Ronaldo in the link shows that he is the reason for which people think of buying FIFA game.

Text Clustering

Once the text has been filtered using the Text Filter node we group similar terms in the dataset together. SAS® Enterprise Miner™ allows us to group terms closely related to each other into separate clusters of related terms. The properties settings for the Text Cluster Node are set to generate an exact ten cluster solution using Expectation-Maximization Cluster Algorithm and 8 descriptive terms that describe the cluster. The ten clusters generated are well separated from each other as seen in Figure 10.

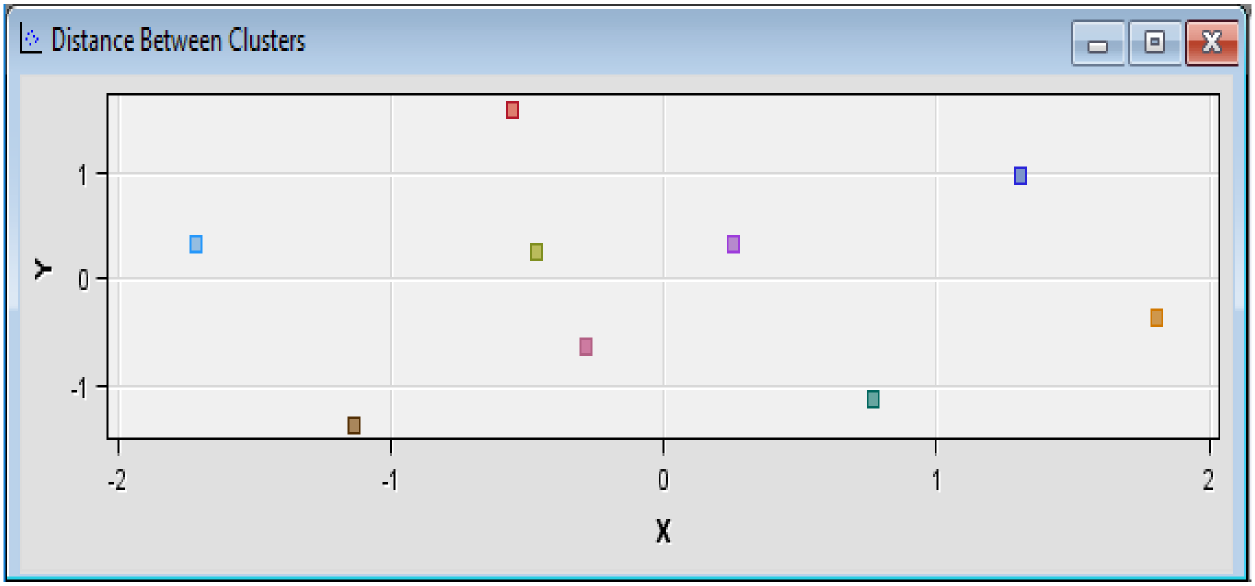

Figure10: Distance between the Clusters

The pie chart shows the distribution of the cluster frequencies.

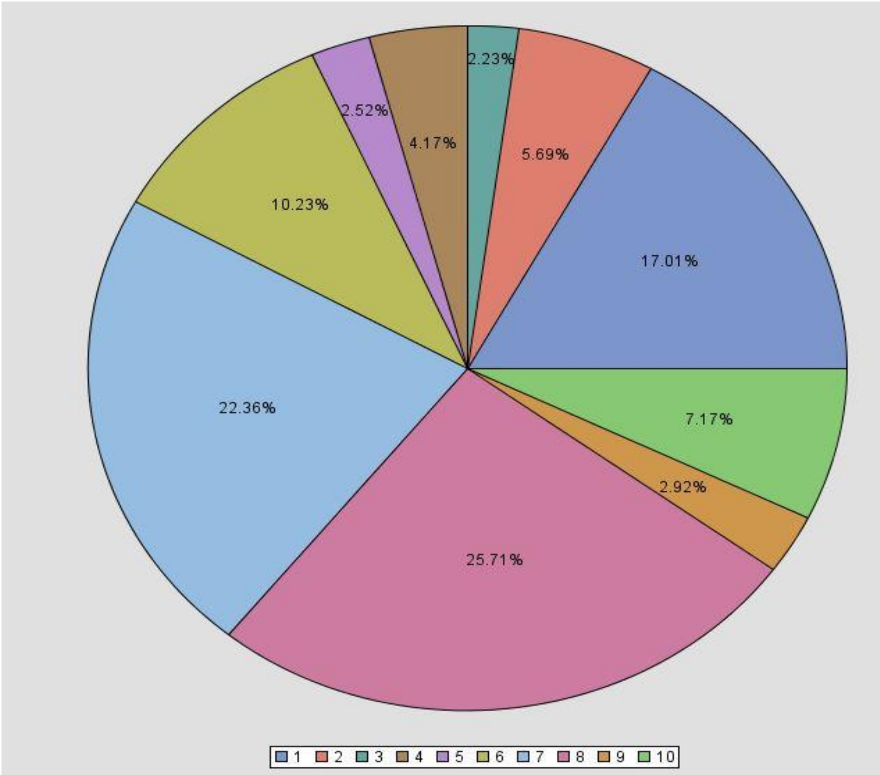

Figure11: Distribution of Clusters

Text Clusters Generated

Cluster ID	Descriptive Terms	Percentage	Explanation
1	+mission, destiny, +level, repetitive, +find, +gear, +shooter, +suck	17 %	This cluster has grouped the reviewers sentiments about the movie and whether or not it was worth their money
2	Uncharted, +dog, naughty, +graphics, +end, series, +amaze, +last	6 %	This cluster is a group of terms for the classification of certain parts and types of the movie.
3	+far, +cry, cry, +open, +world, +story, +graphics, +drive	2 %	This cluster clearly groups terms related to television series.
4	+dark, +soul, +challenge, +level, +find, +little, +feel, +world	4 %	This cluster groups all the musical and dance related terms together.
5	+metal, +gear, +solid, +mission, +open, +pain, series +world	3 %	This cluster is a grouping of the attributes that are generally associated to wars.
6	+system, +hour, +world, +race, +character +find, +first, series	10 %	This cluster is a grouping of the terms that are related to movies that come under the comedy genre.
7	+game, +good, +play, +story, gta, +graphics, +great, +awesome	22 %	This cluster groups the terms that determine movies that maybe adopted from books and novels.
8	+game, +play, +fun, +review, +player, +multiplayer, +people, +mode	26 %	This cluster groups together all the people involved with the movie and the story and roles of everyone.
9	+car, +race, +drive, +track, +mode, +good, online, +graphics	3 %	This cluster is a grouping of the terms that are related to horror and thriller categories of movies.
10	+mode, +player, online +bad, +buy, +year, fifa, +team	7 %	This cluster is a grouping of the terms that are related to action and war related movies.

Table2: Distribution and Explanation of 'Text Clusters'

Text Topic

After connecting the Text Filter node in SAS® Enterprise Miner™ we join the Text Topic node which will enable us to combine the term into topics so that we can analyze further. The properties settings for the Text Topic node have been set to generate 7 topics.

Topic ID	Topic Terms	Explanation
1	destiny, +mmo, +people , +review, +mission	This topic shows the presence of the game destiny and also discusses about it.
2	Uncharted, +dog, naughty, +uncharted series	This topic discuss about series games such as uncharted
3	gta, theft, gta, +car, auto	This topic clearly groups terms related to racing games such as Grand Theft Auto
4	fifa, +player, +ball, +year, +bad,	This topic clearly groups the soccer games like FIFA
5	+enemy, +weapon, +level, +kill, +combat	This topic is a grouping of the attributes that are generally associated to combat games.
6	+mode, multiplayer, +player, +fun, online	This topic clearly groups terms related to mode of play like multiplayer, online etc
7	metal, +gear, +solid, +open, +world,	This topic clear groups terms related to metal gear solid game.

Table3: Distribution and Explanation of 'Text Clusters'

RULE BASED MODEL

Methodology

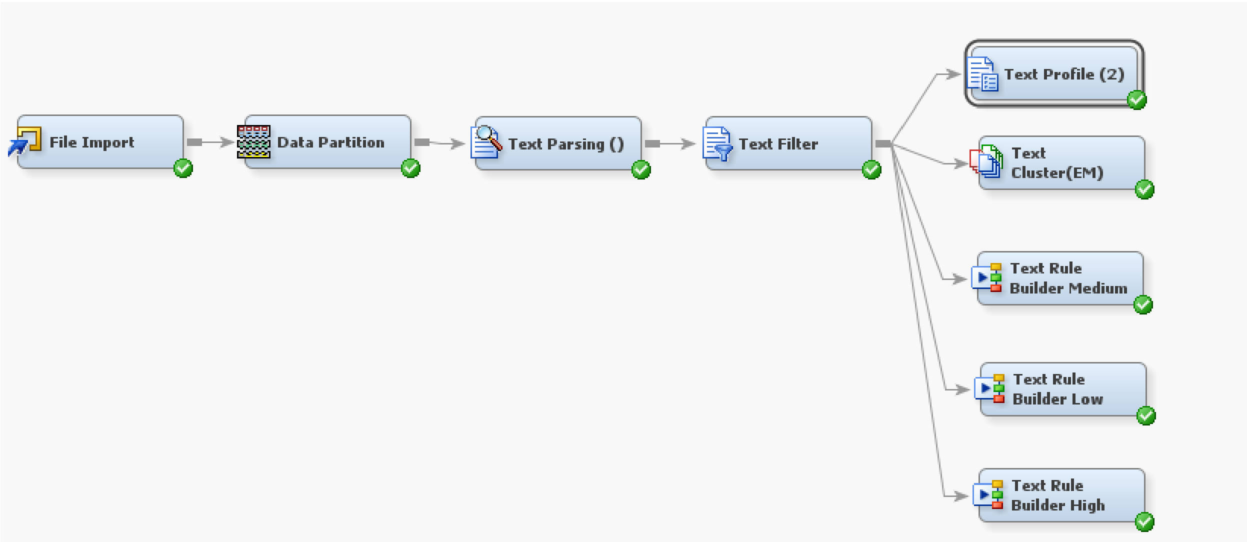

Figure12: Methodology for Rule based model

We have a data set with all the 10,000 reviews and the target variable coded as 1 for 'positive' and 0 for 'negative'. We first use the data partition node to set 70% of the observations as training and the rest 30% as validation. Then the text parsing and text filter nodes are added similar to before. All the properties of the text parsing and text filter node are set the same way as we did before building the clusters. Next we added the text rule builder node with different combination of settings in the properties panel.

The text rule builder node is run with low, medium and high settings for the generalization error, purity of rules and exhaustiveness settings. Amongst these, we found that the text rule builder with the high setting was the best model with the lowest misclassification rate. The misclassification rate for the validation data is 21.98%. To further improve the model accuracy we used the 'change target values' property to manually check if any review was classified incorrectly. An example is shown in figure 13.

Change Target Values-WORK.TRCHANGE

Text	Data Partition	Target Variable	Original Target	Predicted Target	Why Classified	Posterior Probability	Assigned Target
Metal Gear Solid 5 Review (Spoiler Free) - \n \nEver since I've seen the first trailer of MGS 5, i haven't been able to contain my hype. Waiting for the long awaited GOTY 2015 has been hard and finall	Validate	Rating	0	1	kojima	100.0%	0

Figure13: Target values edited

The review clearly shows that it is negative however it was originally classified as positive (1). The model predicted it correctly as negative (0). Hence using our judgement we went ahead and changed the value of the assigned target from positive (1) to negative (0). After making a few more changes the model was run again and now the misclassification rate for the validation data fell to 21.81%. The fit statistics of the model after the manual changes can be seen in figure 18.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Rating	Rating	_ASE_	Average Squar...	0.055304	0.056949	0.066459
Rating	Rating	_DIV_	Divisor for ASE	1320	330	500
Rating	Rating	_MAX_	Maximum Abs...	0.510618	0.510618	0.510618
Rating	Rating	_NOBS_	Sum of Freque...	660	165	250
Rating	Rating	_RASE_	Root Average ...	0.235169	0.238641	0.257796
Rating	Rating	_SSE_	Sum of Square...	73.00193	18.79327	33.2293
Rating	Rating	_DISF_	Frequency of C...	660	165	250
Rating	Rating	_MISC_	Misclassificati...	0.165152	0.218182	0.216
Rating	Rating	_WRONG_	Number of Wro...	109	36	54

Figure14: Rule based model fit statistics

Now to understand what terms were used to categorize the review as good or bad we will look at the rules that govern them. The rules for reviews are seen in figure 19.

Target Value	Rule # ▲	Rule	Precision
0	1	money	94.64%
0	2	repetitive & ~great	94.34%
0	3	worst & ~metal	94.59%
0	4	bungie	93.90%
0	5	boring	92.23%
0	6	ball	92.23%
0	7	bad & ~dog	90.52%
0	8	shoot	90.68%
1	9	naughty	98.08%
1	10	kojima	98.78%
1	11	uncharted	98.26%
1	12	great & ~money & ~repetitive & ~disappointment & ~worst	91.94%
1	13	metal	91.22%
1	14	stealth	90.79%
1	15	good & ~boring & ~worst & ~bungie	86.67%

Figure15: Rules to classify Positive/Negative reviews

The most important rule is if it costs more money, then the review is classified into a negative review.
The rule 12 states that if the review contains great as a term and there is no money and if the game is not repetitive and not disappointing and also if it is not the worst game then the review is classified as positive.

Now we will use the model built to score the data with 1000 observations having 500 positive and 500 negative. The data used to score already has a target variable coded as ‘1’ for positive and ‘0’ for negative. This can be used to check how many positive and negative reviews were correctly scored by the model.

CONCLUSION

- Game reviews give an insight of what people expect from a game, this can be used by the developers to come up with games that can satisfy and reach the expectations of the people.
- A Score node can be used to test new reviews. They can be classified into positive and negative reviews with the help of the text rule builder.
- We can also get an insight of what the users want from a game, like graphics and gameplay seem to be most important here, and if the game is repetitive they feel it is boring, this can be seen from the concept link of the term repetitive.
- It can also be seen from the concept link of the term repetitive that fighting games and shooting games are most repetitive and if so, can be really boring.

LIMITATIONS AND FUTURE WORK

- The study has several imitations which provides scope for further research and exploration.
- We couldn’t include many reviews because they were not in English. Additional linguistic research is needed.
- Advanced analysis could have been done with proper domain expertise.
- Deeper analysis on sentiments of people based on different consoles such as XBOX, PLAYSTATION, and PC is what we hope to achieve in the future

REFERENCES

1) Learning Word Vectors for Sentiment Analysis by Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).
2) Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS® by Goutam Chakraborty, Murali Pagolu, Satish Garla.
3) Sentiment Analysis and Opinion Mining by Bing Liu (May 2012).
4) SAS Institute Inc 2014. Getting Started with SAS® Text Miner 13.2. Cary, NC: SAS Institute Inc.
5) Huayi Li, Arjun Mukherjee, Jianfeng Si and Bing Liu. Extracting Verb Expressions Implying Negative Opinions. Proceedings of Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15). 2015.

ACKNOWLEDGEMENT

- I heartily thank Dr. Miriam Mcgaugh, Professor at Oklahoma State University for her constant inputs and guidance without which I couldn’t have done this research project.
- I would also like to express my gratitude to Dr. Goutam Chakraborty, Professor, Department of Marketing and founder of SAS and OSU Data Mining Certificate program, Director MS in Business Analytics – Oklahoma State University for reviewing the research work done and for providing valuable suggestions.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Mukesh Kumar Singh
Oklahoma State University
Email: mukesh.singh@okstate.edu
Mukesh Kumar Singh is a graduate student currently pursuing Masters in Business Analytics at the Spears School of Business, Oklahoma State University. He is currently working as a Graduate Teaching assistant at Oklahoma State University for marketing department under Professor Dr.Ted Matherly. He is SAS® Certified Base programmer, SAS® Certified Statistical Business Analyst and also SAS certified Predictive Modeler.

Nikhil Batti
Oklahoma State University
Email: nikhil.batti@okstate.edu
Nikhil Batti is a graduate student currently pursuing Masters in Business Analytics at the Spears School of Business, Oklahoma State University. He is currently working as Development assistant at KOSU-Uniquely Oklahoma. He is SAS® Certified programmer and SAS® Certified Statistical Business Analyst.

Dr. Miriam Mcgaugh
Oklahoma State University
Email: miriam.mcgaugh@okstate.edu
Dr. Miriam Mcgaugh professor of marketing at Oklahoma State University. She has many years of experience in using SAS® for data analysis.