Comparing Summarisation Techniques for Informal Online Reviews

Mhairi McNeill[1, 2], Robert Raeside[1], Martin Graham[1] and Isaac Roseboom[2]
[1]*Edinburgh Napier University, Edinburgh, Scotland, U.K.*
[2]*deltaDNA, 25 Greenside Place, Edinburgh Scotland, U.K.*

Keywords: Latent Dirichlet Allocation, Natural Language Processing, Opinion Extraction, Review Summarisation, Sentiment Analysis, Text Mining.

Abstract: In this paper we evaluate three methods for summarising game reviews written in a casual style. This was done in order to create a review summarisation system to be used by clients of deltaDNA. We look at one well-known method based on natural language processing, and describe two statistical methods that could be used for summarisation: one based on TF-IDF scores another using supervised latent Dirichlet allocation. We find, due to the informality of these online reviews, that natural language based techniques work less well than they do on other types of reviews, and we recommend using techniques based on the statistical properties of the words' frequencies. In particular, we decided to use a TF-IDF score based system in the final system.

1 INTRODUCTION

In this paper three approaches for building an online games review summariser are described and compared. The systems were developed at deltaDNA, a small company who provide a games consulting platform specialising in providing services for developers of mobile and PC games. The company wished to provide a tool which allows analysis of external text reviews, found on game stores such as the iTunes store, Google Play and the Steam Store. For a given game, these reviews are numerous, typically short and have a numeric rating provided with them. Because of the volume of the reviews the area of text mining is explored to utilise ideas of classification as advanced by Foreman (2003), Pang and Lee (20020 and He et al (2012) and of summarisation, (see Hu and Liu, 2004).

We wanted to create a system where our clients can quickly get an overview of how well, or how poorly, their game and aspects of it are being reviewed. This system is somewhat different from many of the review summarisation systems described in the literature (Labbé and Poreit 2012, Zhuang et. al. 2006, Mahajan et. al. 2007). The reviews found on online game store are very informal. These reviews are characterised by non-traditional spelling, non-standard grammar, use of emoticons and relatively short reviews. Another way the system is different is because we are providing a summary for the sellers of the product, rather than the buyers of the product.

A state-of-the art system was not a priority for the company. The company wanted a system which is implementable and maintainable with current, mainstream NLP tools. We want to find a summarisation that balances being complete with being concise. In this paper, we also evaluate whether the system works on this data as we expected.

There is a need for the system to work in a reasonable amount of time. We wanted a review summarisation system which will batch summarise new and existing reviews once a day and we need it to do this for hundreds of thousands of games. We do not need 'real-time' summarisation i.e. preparing a summarisation when a user asks for it; this would require a summarisation that works within milliseconds. However, we do want to limit the processing time and power required, to ensure scalability.

Review summarisation has become a popular topic within text mining. He et al (2012) applied text mining to cluster and classify and Bing and Lui's paper used summarizing digital camera reviews as an example. Review summarization also been applied to tourism (Labbé and Poreit 2012), movie reviews (Zhuang et al. 2006), and restaurant reviews

McNeill, M., Raeside, R., Graham, M. and Roseboom, I..
Comparing Summarisation Techniques for Informal Online Reviews.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 322-329
ISBN: 978-989-758-158-8

(Mahajan et al. 2007). He et al (2013) show that using text mining to analyse customer generated content in the form of reviews given in social media can give companies a competitive advantage. Mostafa (2013) demonstrated that text analysis of tweets is useful to marketers in undertaking brand analysis.

Chevalier and Mayzlin (2006) showed that online consumer reviews were an important predictor of sales for books. Furthermore, they showed that the actual text of the reviews, not just the average rating given by users, was considered by purchasers. Chatterjee (2001) reported that online reviews are particularly important in situations when users did not know the brand or did not know the online retailer they were buying from.

Most work in review summarisation is based on the work of Hu and Lui, as described in their 2004 paper 'Mining and Summarizing Customer Reviews'. Hu and Lui took a natural language processing based technique to extract noun phrases and associated adjectives. This technique did not take account of the numeric rating which is often provided alongside the text of the review. We will implement a simple version of this technique in this paper, and compare it to other techniques. Important for this paper is the extension of this work by Titov and Macdonald (2008) who presented a joint model to combine text and aspect ratings for sentiment analysis. They found this method to be an accurate and efficient method of segmentation.

Another technique we use is based on TF-IDF scores, supplemented with numeric ratings. TF-IDF scores are a very popular technique in information extraction and are used as part of several review summary techniques (for example Chen and Chue 2005).

The final technique we try is based topic modelling using on latent Dirichlet allocation (LDA). This statistical model does not seem to be commonly applied to the field of review summarisation. One notable exception to this is Liu and Wang's 2012 paper which used it as a feature extraction method. In this paper we will used supervised LDA (Blei, McAuliffe 2008), which allows us to consider also the numeric ratings in the topics extracted. Supervised LDA does not seem to have been applied to review summarisation so far.

As well as considering techniques that have not been commonly applied to text summarisation we are applying these methods to a relatively unstudied type of review. We have currently seen no papers applying review summarisation techniques to game reviews.

2 METHODS

The reviews were collected from the iTunes store, using the iTunes reviews API on 06/07/2015. For this paper we will be looking at reviews for the popular game 'Candy Crush'. The API returned 1,450 reviews from four iTunes stores: the British store, the American store, the Australian store and the New Zealand store. At the time the reviews were collected the game has 1,874 reviews in the British store alone, making it a good candidate for review summarisation. Each review was returned with a title, the text of the review and the rating out of five that the reviewer gave. For the purpose of this analysis we treated the review as being made up of the review text and the title joined together.

2.1 Natural Language Processing

The first method investigated is based on the work of Hu and Lui 2004. This approach used natural language processing to find 'features' of the product being reviewed and then finds opinions of those features. We have made two major simplifications to Hu and Lui's method. Firstly, we have not carried out any pruning of noun phrases. Secondly, we have missed out steps for finding rare features. For the purposes of this paper we are only looking at a small number of top features, making those two steps less relevant.

We used the Python programming language for implementing this technique. The NLTK library (Bird et. al. 2009) was used to implement the natural language processing techniques. The method used for this paper is as follows:

1. Split the reviews into sentences. For this we used NLTK's sentence tokenizer.

2. Tag the parts of speech in the reviews. To do this we used NLTK's word tokenizer and part of speech tagger. We attempted to use the Stanford 2.0 tagger, as it is currently considered one of the state-of-the-art taggers (ACL Wiki Contributors, 2015). However, we found this tagger far too slow for our application. Tagging this dataset took over two hours with the Stanford tagger.

3. Find the most commonly used noun phrases. Noun phrases were detected using regular expression parser as described in Kim et al. (2010). For each noun phrase, we counted the number of sentences the phrase appeared in. Even if the phrase appeared twice in one sentence, this would only count as one use of the noun phrase.

4. Find the adjectives associated with those noun phrases. For each noun phrase we;
   1. took all sentences that contained that noun phrase;
   2. found the adjectives that were also in that sentence;
   3. counted how often each adjective was used with each noun phrase;
   4. found the sentiment of those adjectives;

Each adjective included in the summary had a sentiment attached to it. This was done by simply comparing the words to a list of words of known sentiment.

The final summary consists of a table of: noun phrases, how frequently that noun phrase occurred, adjectives associated with the noun phrase, frequency of that adjective and the sentiment of that adjective.

2.2 TF-IDF Scores and Average Rating

The second and third methods were implemented in R (R Core, 2015), using the 'tm' library (Feinerer, Hornik 2015). Several cleaning steps are common to both: all text was converted to lowercase, punctuation was removed. Words were stemmed using the Snowball stemmer. Stopwords on the SMART stopword list were removed (Buckley 1985). Also, sparse words were removed. These were words that did not occur very frequently - in particular did not occur across many documents. Many of these words were misspellings and were not useful for summarisation.

The second method used is similar to the natural language processing method described above in that in that the search is for summary words and trying to find the sentiment of those words. However, this method is focusing on phrases and is not differentiating between nouns and verbs. Rather, a metric called term frequency - inverse document frequency (TF-IDF) is used to extract the most important terms. The sentiment is ascertained, not by comparing with words of known sentiment, but by using the ratings that reviewers leave along with their review.

Term frequency - inverse document frequency is the ratio of two terms: the term frequency and the inverse of the document frequency. A word will have a high TF-IDF score for a document if it appears frequently in the document while simultaneously not appearing very often in the documents as a whole. We can think of the words with high TF-IDF scores in a document being the most important words in that document, or somehow describing the document.

The term frequency (TF) of term in a document is simply how often that term is used divided by the number of terms in the document.

$$TF_{t,d} = \frac{frequency\ of\ term\ in\ document}{total\ terms\ in\ document}$$

The document frequency (DF) is the fraction of documents that contain the term. The inverse document frequency (IDF) is the reciprocal of this:

$$IDF_t = \frac{total\ number\ of\ documents}{number\ of\ documents\ containing\ term}$$

We can combine these two measures to find TF-IDF (term frequency - inverse document frequency). To do this we multiply the term frequency with the log of the inverse document frequency:

$$TF - IDF_{t,d} = TF_{t,d} \times log_e(IDF_t)$$

To summarise a game's reviews we took all reviews of that game to be one document. We then compared that set of reviews with other sets of reviews for different games. This way we found a set of words that most uniquely described the game we were trying to summarise.

We wanted to find the sentiment associated with each word. To do this we found the 'average rating' of that word. For each word we have a count of how often that word appears in our n reviews. For a given word we can write this as a vector$(c_1, c_2, \ldots, c_n)$. We also have the rating that each reviewer gave along with their review, which we will write as$(r_1, r_2, \ldots, r_n)$. The word sentiment is the scalar product of these two vectors divided by how often the word appears i.e.

$$(c_1 \times r_1 + c_2 \times r + \ldots + c_n \times r_n) / \sum_{i=1}^{n} c_i r_i .$$

This technique will return a list of words, their TF-IDF scores and a sentiment score for each of those words.

2.3 Supervised LDA

The third review summarisation method used differs substantially from the first two. However, like the previous method, it does not use any natural language processing, rather a statistical process to find summary information.

It is proposed to use a variant of latent Dirichlet allocation (LDA), (see Blei et al 2003 and 2012) as a review summarisation technique. LDA is an example of a topic model, a statistical model for documents. The latent Dirichlet allocation model assumes documents are made up of up of a series of topics, which are in turn made up of a series of words. Each topic has words that are likely to be used in that topic. These words are the only information we have about the topic. The topics are said to be latent properties of the document (they are properties of the document we have no information about but are using in the model). It is assumed that topics are allocated to documents though a Dirichlet process. The latent topics, and use of the Dirichlet distribution, together give latent Dirichlet allocation its name. The hope with using topic modelling as a summarisation technique is that the topics should be aspects of the game that features repeatedly in the reviews.

In this situation, where we have the ratings as well as the reviews, we decided to use a variation of LDA called supervised latent Dirichlet allocation (sLDA). We will be using the average rating as a supervising variable, which will help us allocate topics to documents. This way we will be trying to find a meaningful structure in the documents which best explains the different ratings. Furthermore, each topic will be associated with an estimated rating - so we can see which topics are associated with positive reviews and which topics are associated with negative reviews. sLDA was developed for creating features for the predictive modelling of documents. The nature of topics which best predict an external variable has interesting interpretations and so we can use sLDA for review summarisation.

In both LDA and sLDA, we assume a generative process for the data and find parameters that best fit that process given the data. The generative process for LDA is:

1. Draw from a Dirichlet distribution to select θ, the topics of our document. The number of topics, K, is known:
$$\theta \sim Dir(\alpha)$$
2. For each word w_n:
   a. Chose a topic, z_n that word will come from:
   $$z_n \sim Multinomial(\theta)$$
   b. Chose a word from the multinomial probability distribution of our topic:
   $$w_n \sim Multinomial(\beta_{z_n})$$

In this way a document is made out of words, each of which have a topic associated with them.

sLDA is the same but with a third step. In this third step we draw the response variable from a normal distribution. The distribution from which the response variable is drawn depends on the average topic of the document:

3. For the document draw a response variable y:
$$y \sim N(\eta^T \underline{z}, \sigma^2)$$

where $\underline{z}$ is defined as:

$$\underline{z} = \frac{1}{N}\sum_{n=1}^{N} z_n$$

Both LDA and sLDA estimate the parameters $\alpha, \beta_{1:K}, \eta, \sigma^2$using approximate maximum-likelihood estimation using variation expectation-maximization. For more information of this procedure see see McAuliffe and Blei (2007) and Blei, et al. (2012). Applications of LDA are presented by Ramage (2009). To implement sLDA the 'lda' package in R (Chang 2012) was used.

3 RESULTS

The results of each summarisation technique are presented in tables. Only a sample of the table for each result will be presented, since the full tables are, in general, very large for presentation. The sample table gives an indication of the type of result obtained.

While these results do not provide evidence for one method performing better than another they do give a guide to how the results looked and how evaluations were made.

3.1 Natural Language Processing

Note that this table is truncated. It shows a summary of some of the results achievable with the natural language processing technique used. The most common nouns and the adjectives most commonly associated with them are shown in the table. We also have an estimate of the sentiment associated with those adjectives.

Table 1: Results for natural language processing.

Noun phrase	Noun count	Adjective	Adjective count	Adjective sentiment
game	479	addictive	51	Neutral
game	479	best	39	Positive
game	479	great	37	Positive
game	479	good	37	Positive
game	479	new	20	Neutral
love	216	addictive	17	Neutral
…	…	…	…	…

3.2 TF-IDF Scores and Average Rating

Note that this table is truncated. We are shown a list of common words, along with their TF-IDF scores. Each word also has the average rating out of five that is associated with it.

Table 2: Results for TF-IDF scores and average rating method.

Word	TF-IDF score	Ratings
fun	0.4495	4.5375
good	0.3217	4.4946
love	0.1716	4.5534
crush	0.1453	4.2688
great	0.1395	4.4974
time	0.1280	3.4619
update	0.1254	2.2222
levels	0.1196	3.7573
game	0.1166	3.9876
level	0.1053	2.9037
play	0.1036	3.3412
like	0.1034	3.9101
...	…	…

3.3 Supervised LDA

For supervised LDA we can present the whole table for 10 topics.

Table 3: Supervised LDA results.

Estimated Score	Top 5 words
1.0036	love crush saga though enough
0.9696	fun great super wish wait
0.9557	great game good cool win
0.9246	game like love play gold
0.9026	time good pass way still
0.8453	game never got wheel best
0.8286	game just will much get
0.5688	app can get level next
0.3494	screen reward claim need sugar
0.3256	fix level back scroll start

The score estimates the sentiment associated with each topic. Scores closer to one indicate a more positive topic. We have made no effort to describe each topic, as it would be difficult to do this as part of an automated system.

4 DISCUSSION

An objective method of evaluating these three techniques is not immediately apparent. Several people subjectively examined each summarisation to evaluate how well they met our needs for a summarisation system. It would have been possible to use many reviewers and calculate the precision and recall of each technique in a more objective way such as that documented in Lizhen et al. (2012) or Anwer et. al. (2010). However, this was considered too expensive and time consuming to be commercially viable. Furthermore, it was felt that this evaluation would be unable to adequately weigh-up the complex requirements we discussed in the introduction.

Each technique will be discussed in detail, with advantages and disadvantages of each. We used the following three questions as guide for our evaluation.

1. Does the method work as expected on this data?
2. Is the summarisation concise?
3. Is the summarisation complete, descriptive and accurate reflection of the content of the reviews?

We have read a random sample of the reviews used as data for these summarisations in order to best evaluate the final criteria. The most frequent comments that we hoped the summarisation would pick up on were:

- Positive reviews were in general quite similar. Mostly the reviewers were commenting that the game was fun, addictive and a good time waster.
- Negative reviews fell into two categories. There were many reviewers complaining about bugs, mainly the game freezing. The second set of negative reviews were unhappy about having to ask friends for lives or pay for items in the game.
- There were very few neutral reviews.

After this discussion of each technique's pros and cons, there will then be a brief discussion of the speed of each technique.

4.1 Natural Language Processing

4.1.1 Does the Method Work as Expected?

A major failing of natural language processing on these reviews is that type of speech tagging we used does not work well for this type of review. The lack of grammar and poor spelling makes the part of speech tagger mislabel both the nouns and the adjectives. Other part of speech tagging algorithms may be more effective, however, we also needed a reasonable speed in the part of speech tagging and we believe that any part of speech tagger which was not trained on an informal corpus would suffer similar problems.

There were two other sources of difficulty for Hu and Liu's method on this dataset. Firstly, it was difficult to split the reviews into sentences because full stops are used inconstantly. Secondly, while the algorithm was meant to pull out noun phrases, it almost exclusively pulled out single words.

4.1.2 Is the Summarisation Concise?

A further problem with this technique is that it returns a large table (five columns and many rows), which is difficult to evaluate. The size of the table depends on how many nouns and adjectives the user wishes to view. Many of the adjectives are repeated multiple times in the table making many of the elements of the table redundant.

4.1.3 Is the Summarisation Complete?

It was found that the natural language technique works surprisingly well, given the inaccuracy of the part of speech tagger. There was a clear indication that most users found the game fun, and addictive.

The 12th most common noun phrase extracted was 'please fix' with the associated adjectives: 'latest', 'bean', 'serial', 'impossible', 'next'. This does give some indication of the problems with bugs. However, this is quite deep into the summarisation and is far from clear what the problems actually are.

4.2 TF-IDF Scores and Average Rating

4.2.1 Does the Method Work as Expected?

In general this method works approximately as would be hoped. Words very common in the reviews do score highly for TF-IDF. However, many of these words are quite generic and would be shared across many games. Perhaps a larger corpus of comparison reviews would alleviate this problem somewhat.

In general, words that seem positive get positive average ratings; and more negative words get negative average ratings.

4.2.2 Is the Summarisation Concise?

The table retuned as two metrics for each word, which is reasonably concise. Many of the top words are quite similar so a relatively large number of words are needed to extract more unusual features of the reviews.

4.2.3 Is the Summarisation Complete?

Considering the top positively rated words and the top negatively rated words, one could get a reasonably good picture of the content of the reviews. However, it is difficult to understand the negative comments in detail because the summaries simply consist of single words without context.

4.3 Supervised LDA

4.3.1 Does the Method Work as Expected?

This method works on this data to some extent. Many of the middle topics seem to be associated with quite similar words, and the meaning of the topic cannot be ascertained. A different number of topics might have been a better summarisation. However, it is difficult to know the number of topics which best summarises a set of documents a priori.

It does seem that topics associated with a low estimated ratings are negative in nature; and vice versa for topics with a high estimated rating.

4.3.2 Is the Summarisation Concise?

This is the most concise of all the summarisation techniques and was the only technique for which we put the full table of results into the paper. Only one set of numbers is associated with the summary.

4.3.3 Is the Summarisation Complete?

This technique gives a relatively good description of the contents of the reviews. We can see many people found the game fun. The fifth topic includes 'time good pass way 'which is perhaps summarising the 'good way to pass the time' comments. However, clearly, the meaning is lost somewhat since we did not consider the order of words.

Some of the topics with low estimated ratings seem to be trying to describe the problems with bugs. One can also see quite a gap between the top 8 topics and the bottom 3, which is probably the result of most reviews being positive, with a small number of negative reviews; and very few balanced or neutral reviews.

It is important to note that, unlike the two other techniques in this paper, supervised LDA relies on simulation, and so gives different results on different runs of the algorithm.

4.4 Time Comparison

Table 4 gives the time taken to run each summarisation algorithm to get the results shown in this paper. These times include all data cleaning needed to get the final result (expect for TF-IDF, when the comparison corpus is simply loaded in clean). All times given are in seconds.

Table 4: Speed comparison results.

Technique	Time (s)
Natural language	27.04
TF-IDF	2.70
Supervised LDA	5.72

These times should only be used a rough guide, since we made no attempt at optimising the implementations used. However, note that the natural language technique takes significantly longer than the other two.

4.5 Conclusions and Future Work

Natural language processing based techniques are the most popular variety used for review processing currently. However, for this dataset they perform comparatively poorly against other, more statistical techniques. Furthermore, the natural language techniques as described do not take into account the rating given with the review. This can be an important and useful source of sentiment information.

Out of the other two techniques the most impressive results were from using supervised LDA. This technique is not currently being used as a statistical summary technique as far as we are aware and we encourage its use. But it does have three main failings which need to be overcome: it is not obvious how many topics to model, there is no accounting for word order or context and the model is quite a 'black box'; it is difficult to understand what the model is doing.

TF-IDF scores have the advantage of being transparent and easy to understand. The summary provided by them is not as concise or complete but can be useful under the right circumstances.

For the final product we ended up using a TF-IDF score based summarisation. A major advantage of this technique is that we were able to explain its implementation to our clients and stakeholders. Furthermore, this technique was easy to implement without depending on external libraries, which can add complexity to maintenance.

None of the techniques were entirely adequate. We speculate that a technique with a better understanding of the patterns of speech used in casual internet communication, combined with topic modelling, could give very good results for this data.

REFERENCES

Anwer, N., Rashid, A., & Hassan, S. (2010, August). Feature based opinion mining of online free format customer reviews using frequency distribution and Bayesian statistics. In Networked Computing and Advanced Information Management (NCM), 2010 Sixth International Conference on (pp. 57-62). IEEE.

ACL Contributors. (2015). POS Tagging (State of the art). Available: http://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art). Last accessed 31th Jul 2015.

Blei, D. M., Ng, A. Y. and Jordan, M. I., (2003). Latent Dirichlrt allocation. Journal of Machine Learning Research, 3: 993-1022.

Blei, D.M., Ng, A.Y. & Jordan, M.I., 2012. Latent Dirichlet Allocation J. Lafferty, ed. *Journal of Machine Learning Research*, 3(4-5), pp.993–1022.

Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python.

Buckley, C. (1985). Implementation of the SMART information retrieval system. Cornell University.

Chang, J 2012. lda: Collapsed Gibbs sampling methods for topic models.. R package version 1.3.2. http://CRAN.R-project.org/package=lda.

Chatterjee, P., 2001. Online Reviews: Do Consumers Use Them? *Advances in Consumer Research*, 28, pp.129–134.

Chen L. and Chue W., 2005. Using Web structure and summarization techniques for web content mining, Information Processing and Management.

Chevalier, J. a & Mayzlin, D., 2003. The Effect of Word of Mouth on Sales : *National Bureau of Economic Research*, p.40.

Feinerer, I, Hornik K, and Meyer D (2008). Text Mining Infrastructure in R. Journal of Statistical Software 25(5):1-54. URL: http://www.jstatsoft.org/v25/i05/.

Forman, G., (2003). An extensive empirical study of feature selection metrics for text classification, journal of machine Learning Research, 3: 1289-1305.

He, W., Chee, T., Chong, D. Z. and Rasnick, E., (2012). Analysing the trends of E-marketing from 2001 to 2010 with use of biblometrics and text mining. International Journal of Online Marketing, 2(1), 16-24.

He, W., Zha, s. and Li, L., (2013). Social media competitive analysis: A case study in the pizza industry, International Journal of Information Management, 33: 464-472.

Hu, M. & Liu, B., 2004. Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*, 04, p.168.

Hu, M. & Liu, B., 2004. "Mining Opinion Features in Customer reviews". In Proceedings of Nineteenth National Conference on Artificial Intelligence (San Jose, California, USA, July 2-29, 2004). The AAAI Press, Menlo Park, CA, 755-760.

Labbé, C. & Portet, F., 2012. Towards an abstractive opinion summarisation of multiple reviews in the tourism domain. In *CEUR Workshop Proceedings*. pp. 87–94.

Lihui, C. & Chue, W.L., 2005. Using Web structure and summarisation techniques for Web content mining. *Information Processing and Management*, 41(5), pp.1225–1242.

Lizhen Liu; Wentao Wang; HangShi Wang, "Summarizing customer reviews based on product features," Image and Signal Processing (CISP), 2012 5th International Congress on.

Mcauliffe, J.D. & Blei, D.M., 2008. Supervised Topic Models. In *Advances in Neural Information Processing Systems*. pp. 121–128. Available at: http://papers.nips.cc/paper/3328-supervised-topic [Accessed July 13, 2015].

Mostafa, M., 92013). More than words: Social networks' text mining for consumer brand sentiments, Expert Systems with Applications, 40: 4241-4251.

Nguyen, P., Mahajan, M. & Zweig, G., 2007. *Summarization of Multiple User Reviews in the Restaurant Domain*, Available at: http://research.microsoft.com:8082/pubs/70488/tr-2007-126.pdf.

Pang, B. and Lee, L. (2002). Thumbs up? Sentiment Classification using Machine Learning, Proceedings of the conference on empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002: 79-86. Association for Computational Linguistics.

Ramage, D., Hall, D., Nallapati, R. and Manning, C. D., (2009). Labelled LDA: A supervised topic model for credit attribution in multi-labelled corpora, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 248-256., Singapore 6-7 August 2009.

Titov, I. and Macdonald, R., (2008). A joint model of text and aspect ratings for sentiment summarization, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 15-20, 2008, Ohio state University, Columbus, Ohio, USA.

Zhuang, L., Jing, F. & Zhu, X.-Y., 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06*.