UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

VINICIUS WOLOSZYN

Tell Me Why: an architecture to provide rich review explanations

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Prof. PhD Dante Augusto Couto Barone

Porto Alegre
July 2015

CIP — CATALOGING-IN-PUBLICATION

Woloszyn, Vinicius

Tell Me Why: an architecture to provide rich review explanations / Vinicius Woloszyn. – Porto Alegre: PPGC da UFRGS, 2015.

101 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2015. Advisor: Dante Augusto Couto Barone.

1. Opinion Mining. 2. Data Mining. 3. Natural Language Processing. 4. Natural Language Generation. 5. Big Data. I. Augusto Couto Barone, Dante. II. Título.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Prof. Dante Couto Barone, for all his comments and helps for this research. He always supported me and gave me a lot of inspiration of research. I also want to thank Sihem Amer-Yahia, François Portet and Cyril Labbe from Grenoble Informatics Laboratory, whose have invested their effort in guiding me within the purpose of the development of the Tell Me Why. Without their considerate guidance, this thesis could not be done.

Gostaria de agradecer aos meus pais Cláudio Woloszyn e Elizabeth Valim por todo o carinho e dedicação ao longo dos anos. Tais pessoas são um exemplo de caráter e simplicidade que eu quero seguir. Vocês estão sempre em meus pensamentos. Muito Obrigado.

Existe também uma pessoa a qual eu tenho muito carinho que me apoiou muito durante o desenvolvimento desse trabalho. Esse carinho e dedicação, mesmo que as vezes a distância, me ajudou ao cumprimento dos meus objetivos. Bianca Lagasse, eu te amo.

The scientific man does not aim at an immediate result. He does not expect that his advanced ideas will be readily taken up. His work is like that of the planter - for the future. His duty is to lay the foundation for those who are to come, and point the way.

— NIKOLA TESLA

ABSTRACT

What other people think has been always an important part of the process of decision-making. For instance, people usually consult their friends to get an opinion about a book, or a movie or a restaurant. Nowadays, users publish their opinions on collaborative reviewing sites such as IMDB for movies, Yelp for restaurants and TripAdvisor for hotels. Over the time, these sites have built a massive database that connects users, items and opinions expressed by a numeric rating and a free text review that explain *why* they like or dislike a specific item. But this vast amount of data can hamper the user to get an opinion.

Several related work provide a review interpretations to the users. They offer different advantages for various types of summaries. However, they all have the same limitation: *they do not provide personalized summaries nor contrasting reviews written by different segments of reviewers*. Understanding and contrast reviews written by different segments of reviewers is still an open research problem.

Our work proposes a new architecture, called Tell Me Why, which is a project developed at Grenoble Informatics Laboratory in cooperation with Federal University of Rio Grande do Sul to provide users a better understanding of reviews. We propose a combination of text analysis from reviews with mining structured data resulting from crossing reviewer and item dimensions. Additionally, this work performs an investigation of summarization methods utilized in review domain.

The output of our architecture consists of personalized statement using Natural Language Generation that explain people's opinion about a particular item. The evaluation reveal that it is a promising approach and useful in user's opinion.

Keywords: Opinion Mining. Data Mining. Natural Language Processing. Natural Language Generation. Big Data.

Tell My Why: Uma arquitetura para fornecer explicações ricas sobre revisões.

RESUMO

O que as outras pessoas pensam sempre foi uma parte importante do processo de tomada de decisão. Por exemplo, as pessoas costumam consultar seus amigos para obter um parecer sobre um livro ou um filme ou um restaurante. Hoje em dia, os usuários publicam suas opiniões em sites de revisão colaborativa, como IMDB para filmes, Yelp para restaurantes e TripAdiviser para hotéis. Ao longo do tempo, esses sites têm construído um enorme banco de dados que conecta usuários, artigos e opiniões expressas por uma classificação numérica e um comentário de texto livre que explicam por que eles gostam ou não gostam de um item. Mas essa vasta quantidade de dados pode prejudicar o usuário a obter uma opinião.

Muitos trabalhos relacionados fornecem uma interpretações de revisões para os usuários. Eles oferecem vantagens diferentes para vários tipos de resumos. No entanto, todos eles têm a mesma limitação: *eles não fornecem resumos personalizados nem contrastantes comentários escritos por diferentes segmentos de colaboradores*.

Compreeder e contrastar comentários escritos por diferentes segmentos de revisores ainda é um problema de pesquisa em aberto. Assim, nosso trabalho propõe uma nova arquitetura, chamado Tell Me Why. TMW é um projeto desenvolvido no Laboratório de Informática Grenoble em cooperação com a Universidade Federal do Rio Grande do Sul para fornecer aos usuários uma melhor compreensão dos comentários. Propomos uma combinação de análise de texto a partir de comentários com a mineração de dados estruturado resultante do cruzamento de dimensões do avaliador e item. Além disso, este trabalho realiza uma investigação sobre métodos de sumarização utilizados na revisão de produtos.

A saída de nossa arquitetura consiste em declarações personalizadas de texto usando Geração de Linguagem Natural composto por atributos de itens e comentários resumidos que explicam a opinião das pessoas sobre um determinado assunto. Os resultados obtidos a partir de uma avaliação comparativa com a Revisão Mais Útil da Amazon revelam que é uma abordagem promissora e útil na opinião do usuário.

Palavras-chave: Mineração de Opinião, Mineração de Dados, Processamento de Linguagem Natural, Geração de Linguagem Natural,Volumes Massivos de Dados.

LIST OF ABBREVIATIONS AND ACRONYMS

TMW	Tell Me Why
LIG	Grenoble Informatics Laboratory
NLP	Natural Language Processing
POS	Part of speech
WSD	Word Sense Disambiguation
NLG	Natural Language Generation
API	Application Programming Interface
ATS	Automatic text summarization
TF-IDF	Term Frequency - Inverse Document Frequency
DUC	Document Understanding Conference
ANN	Artificial Neural Network
LSA	Latent Semantic Analysis
SVD	Singular Value Decomposition
PCA	Principal Component Analysis
QA	Question Answering
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
KDD	Knowledge Discovery in Databases
BBN	Bayesian Belief Networks
DAG	Directed Acyclic Graph
AMHR	Amazong Most Helpful Review

LIST OF FIGURES

CONTENTS

1 INTRODUCTION

Getting information about what other people think has always been involved in the process of decision-making. For instance, people usually consult their friends to get an opinion about a book, or a movie or a restaurant. Nowadays, users publish their opinions on collaborative reviewing sites such as IMDB[1] for movies, Yelp[2] for restaurants and TripAdvisor[3] for hotels. These sites allow users to write comments about a determined product and also to see what other users wrote about it before decide buying something or watching a movie. Over the time, these sites have built a massive database that connects users, items and opinions expressed by a numeric rating and a free text review that explain *why* they like or dislike an item.

The review also plays an important part in purchasing decisions of the products over the Internet. Empirical studies of the consumers' behavior show the impact of positive reviews into the consumers purchasing decision about a product (CHATTERJEE, 2001; CHEVALIER; MAYZLIN, 2006; KIM; SRIVASTAVA, 2007). Unfortunately, to get a suitable opinion from this mass of data, users have to deal with a vast amount of textual information. Precisely, they have to deal with the problem of a huge mass of divergent reviews about an item and its aspects, in order to extract a suitable opinion from them. Aspects generally refer to features that are relevant to all products. From the exposed problem, we derive four major sub-problems: (i) divergence of opinions in which a significant part of the users approve while the other part disapproves the same item; (ii) user's personal relevance of the aspect of an item to getting an opinion. Some aspects such as "price" and "service" can be more relevant than others for different users at the time they seek for an opinion about an item; (iii) large number of highly redundant opinions; and (iv) too many information about an item, such as a thousand of reviews.

For instance, Figure 1.1 presents two reviews extracted from Yelp about the restaurant called "The Wolseley," located in London. Looking at the text review, we can observe more than just the reasons for the different ratings. It presents a contradictory opinion about same aspects, such as Service "excellent" and "dismissive"; food is "amazing" and "unimpressive". Additionally, the reviews reveal contextual information about the conditions that an item or its feature was observed. For instance, in the sentence "I booked for brunch just to be sure", the context "brunch" would explain a negative opinion about food. In the sentence "My favorite spot for breakfast in London", the context "breakfast" would explain the positive opinion about the food. Such kind of contextual information gives valuable information about the user conditions

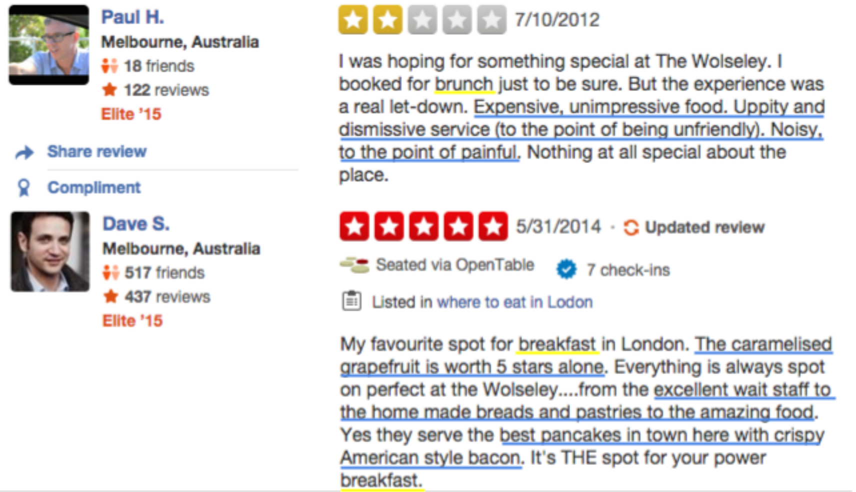
[1] http://www.imdb.com/

[2] http://www.yelp.com/

[3] http://www.TripAdvisor.com/

of the item at the time when it was observed.

Figure 1.1 – Two reviews about the restaurant "The Wolseley" extracted from Yelp. The blue underline marks feature opinions, and the yellow underline marks contextual information



Paul H.
Melbourne, Australia
18 friends
122 reviews
Elite '15

★★☆☆☆ 7/10/2012

I was hoping for something special at The Wolseley. I booked for brunch just to be sure. But the experience was a real let-down. Expensive, unimpressive food. Uppity and dismissive service (to the point of being unfriendly). Noisy, to the point of painful. Nothing at all special about the place.

Dave S.
Melbourne, Australia
517 friends
437 reviews
Elite '15

★★★★★ 5/31/2014 · Updated review

Seated via OpenTable
7 check-ins
Listed in where to eat in Lodon

My favourite spot for breakfast in London. The caramelised grapefruit is worth 5 stars alone. Everything is always spot on perfect at the Wolseley....from the excellent wait staff to the home made breads and pastries to the amazing food. Yes they serve the best pancakes in town here with crispy American style bacon. It's THE spot for your power breakfast.

Source: Developed by Author

In the sequence, Section 1.1 presents the problem that we are addressing and the limitations of related work. In Section 1.2, we present the proposed solution and the expected contributions. Finally, Section 1.3 shows the structure of this dissertation.

1.1 Problem Statement and Related Work Limitation

Given the importance of the reviews and the problems that users have in order to extract a suitable opinion from them, this proposed research focuses on digesting mix of reviews and provides a personalized summary. Regarding Review Summarization domain, several related work are involved to provide review interpretations. All the previous approaches offer different advantages for various types of summaries. These related studies are detailed below.

Abstractive summarization builds an internal semantic representation from a source document(s) to create a totally new text. Ganesan, Zhai and Han (2010) proposed the use of textual graphs representation to generate abstractive summaries (GANESAN; ZHAI; HAN, 2010). Additionally, the Opinion Digger uses free text review, numeric rating and a set of predefined aspects to provide comparisons of different aspects of products (MOGHADDAM; ESTER, 2010).

In the other hand, extractive summaries are built by extracting the most relevant sentences based on statistical analysis of the words frequency or of the occurrence of some features. The RnR system (LABBE et al., 2010) captures and summarizes the key rationale for positive and negative opinions expressed in a corpus of reviews and highlights the negative and positives features and vice versa. It provides a summary of the text review, a performance chart and gives details of each positive and negative aspects of reviews. Additionally, WANG (2014) propose a submodular function-based framework for query-focused opinion summarization. In the approach, the relevance ordering is produced by a statistical ranking, and information coverage with respect to topic distribution and diverse viewpoints are both encoded as submodular functions.

However, they all have the same limitation: *they do not provide personalized summaries nor contrast reviews written by different segments of reviewers*. Understanding and contrasting reviews written by different segments of reviewers is still an open research problem (DAS et al., 2011). Indeed, most of the richness of an opinion is contained in free-text comments whose information is still challenging to be extracted automatically. New solutions must be found to digest the information carried in individual reviews and create personalized summaries. Unfortunately, the ability to explain user reviews is an under-explored research question(LABBE et al., 2010; GANU; ELHADAD; MARIAN, 2009). It requires combining sophisticated text analysis from reviews with mining structured data resulting from crossing reviewer and item dimensions (DAS et al., 2011; REITER; DALE; FENG, 2000).

1.2 Proposed Solution and Contributions

Given the described problem and the related work limitation, the present research leverages data mining techniques and Natural Language Processing to contrast reviews written by different segments of reviewers. Precisely, we want to analyze item, user and review dimensions to extract useful patterns, such as: *this hotel is liked by young travelers* or *middle-aged Frenchs enjoy this restaurant*. After, we want to contrast them with sentences extracted from amount of review that would explain these patterns, such as: "*the room is clean and noiseless*" and "*The food was okay, not great, not bad.[...]Our favorite part, was the service!*". Finally, we use these extracted data to build a short explanatory statement using Natural Language Generation (REITER; DALE; FENG, 2000) that could explain why someone like or dislike an item.

Thus, our approach is composed of three main steps: (i) extracting relevant patterns from the structured data using data mining techniques; (ii) analyzing the text in order to extract sentences that would explain the cited patterns; and (iii) building a short explanatory text using Natural Language Generation. The main advantage of this approach is that it uses structured and unstructured data, which are found into reviews, to build a personalized summary of opinions.

This work makes four primary contributions:

- Development of a new Architecture to provide personalized summaries from reviews;
- Investigation of Automatic Text Summarization methods to summarize reviews;
- Providing comparison between methods to summarize reviews; and
- Provide better comprehension of Automatic Review Summarization;

1.3 Thesis Structure

In this Section, we present the structure of this thesis, as follows: in Chapter 2 we perform a review of theory in order to identify the most appropriate methodology for each desired task. Chapter 3 presents related work their strengths and limitations. Chapter 4 describes our architecture to handle with vast amount divergent reviews to provide a helpful opinion to the users. Chapter 5 describes our experiments and its parameters. Chapter 6 presents our results and discussion. Finally, in Chapter 7, we present the conclusions derived from the two experiments performed. Additionally, future work related to our conclusions are presented.

2 BACKGROUND

To conduct this research, was necessary perform an examination of the theory in order to identify the most appropriate methodology for each desired task. Thus, Section 2.1 presents some pre-processing concepts about Natural Language Processing. Section 2.2 presents the Natural Language Generation. Section 2.3 presents methods and features commonly used in Text Summarization. Section 2.4 presents the Evaluation Methods used in Automatic Text Summarization. Finally, Section 2.5 provides an overview of Machine Learning techniques related to Review Summarization.

2.1 Natural Language Processing

Natural Language Processing (NLP) is a field of research that exploits how computers can understand and handle a text or a speech (PALMER, 2010). The NLP techniques are widely utilized in many areas of studies, such as Machine Translation, Summarization, Information Retrieval and speech recognition. To identify the most suitable techniques to our problem, this section aims perform a review of primary NLP techniques utilized in this process.

Naturally, the first task performed on the text is the data preparation. The actions realized in the preprocessing step intended to prepare them to undergo some indexing algorithm or data mining (PALMER, 2010). In general, the preprocessing step aims to improve the quality of data already available. However, there is no single technique with satisfactory results in all the fields. Thus, to identify the most suitable method to our problem, the main preprocessing methods are listed bellow:

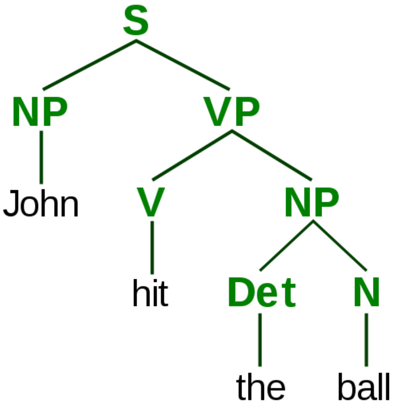
- **Stop Word Removal.** The words with low relevance to express the meaning are called Stop Words. Their presence in a text decreases the algorithm performance in the categorization task. Thus, the Stop Word Removal task aims to exclude these irrelevant words (SILVA; RIBEIRO, 2003). A list of stop words consists of connective words, prepositions, pronouns, articles and words with a very high incidence in a collection of documents. For instance, the English stop words list includes words such as *the, is, at, which, a, able, about, across, after, all, almost, also, am, among,* and *an*.
- **Stemming.** The Stemming refers to the crude heuristic process of cut off the ends of words in order to reveal the word root (MANNING; RAGHAVAN; SCHÜTZE, 2008). After the stemming process, all the words is represented by its stem. It represents an

important step used in text categorization. For example, "inform" could be the lemma of "information" or "inform".

- **Lemmatization.** This task refers the use of a vocabulary and the morphological analysis to determination of the base (lemma) of the verb and nouns to the singular form (MANNING; RAGHAVAN; SCHÜTZE, 2008). For example, in English, the verb "to walk" may appear as "walk", "walked", "walks", "walking".
- **Sentence Segmentation.** To extract and analyze the relevance of the sentences, we first need handle with the problem of split the amount of text into sentences (MANNING; RAGHAVAN; SCHÜTZE, 2008). The use of punctuation, particularly the full stop character is a good approximation. However, it is not trivial due to its utilization for abbreviations, which may not terminate a sentence. For instance, "Mr." is not a sentence in the sentence "Mr. Brow went to the club in Jones Street.". Thus, several methods have used machine-learning techniques (CHUANG; YANG, 2000).
- **Tokenization.** Given a sentence, the tokenization process perform the separation of the elements into pieces, called tokens (MANNING; RAGHAVAN; SCHÜTZE, 2008). In other words, tokens is an instance of a sequence of characters in some particular document. Typically, the tokenization process occurs at the word level, not including punctuation nor white space For example, the sentence "*Friends, Romans, Countrymen, lend me your ears*", the resulted tokens are: *"friends", "Romans", "Countrymen", "lend", "me", "your" and "ears".*
- **Part-of-speech tagging (POS).** POS tagging is the process of labeling a word of a sentence based on its grammatical classification such as nouns, verbs, adjectives, and adverbs (TOUTANOVA; MANNING, 2000). It is a relevant task in the process of identifying aspects and the opinions in the sentence, given the identification of nouns and adjectives. However, considering the high percentage of ambiguous words with more than one morphological classification, the tagging process is not a trivial task. For example, in the sentence "*The sailor dogs the hatch*" the word "*dogs*", that is usually classified as plural noun, can also be a considered as verb.
- **Word Sense Disambiguation (WSD).** The WSD aim to solve the ambiguity problem in the meaning of single words or phrases. An example is *bank* which may have among others the senses *financial institution* or the *border of a river*. A disambiguation process requires a dictionary to specify the senses that are to be disambiguation and a corpus of language data to be disambiguation. Additionally, the machine learning methods requires training corpus of language examples.

- **Parsing.** It refers to the task performed of understanding the exact meaning of a sentence using the utilization of a formal grammar. The result is a representation of relations between words in a sentence (MANNING; SCHÜTZE, 1999). Figure 2.1 shows the dependency trees of the sentence "John hit the ball". Where *S*, represents the sentence. *NP* is abbreviation to noun phrase. *VP* to verbs phrase, which is the predicate. *V* to a verb, in this case, it's a transitive verb "hit". *Det* means a determiner. *N* is used as noun.

Figure 2.1 – A dependency tree



Source: (MANNING; SCHÜTZE, 1999)

2.2 Natural Language Generation

Natural Language Generation (NLG) is a field of Natural Language Processing and linguistics that aims in understand how a computer can build texts automatically. The NLG system uses knowledge about language and the domain to produces automatically documents, reports, explanations and other types of text (REITER; DALE; FENG, 2000). At our work, we aim to build short explanatory statements based on data extracted from reviews. Thus, we present the SimpleNLG system, a tool to build Natural Language.

2.2.1 SimpleNLG

SimpleNLG (GATT; REITER, 2009) is a text realization engine for English language, which aims to provide simple and robust interface to generate syntactic structures and linearize them as a text. Albert Gatt and Ehud Reiter developed SimpleNLG at the University of Aberdeen's Department of Computing Science in 2009. It provides a direct control over the way phrases are built. It defines a set of lexical and phrasal types corresponding to the major grammatical categories, as well as a simple way of combining these and setting a lot of feature values.

2.3 Automatic Text Summarization

The Automatic Text Summarization (ATS) aims to reduce an amount of text into a shorter version preserving the core information content (PALMER, 2010). ATS has become one important field of research motivated by the abundance of text available on the internet and the difficult for human beings manually to summarize them. Review summarization is one objective addressed by this work. To identify the most suitable methods to our problem, we present the most common methods used in ATS, as well as its pros and cons.

To better understanding of ATS methods, some relevant concepts must be made to clarify and avoid misunderstanding. Regarding the utilization that the summary is supposed to serve, the summary can be categorized into indicative or informative summary.

- **Indicative Summary**. This type of summary does not claim any role in substituting the source document. Its purpose is merely to alert the reader in relation to the original contents. It is usually employed to allow users to decide which part of the original text should read (NENKOVA; MASKEY; LIU, 2011).
- **Informative Summary.** This type of summary can be read in place of the original document (NENKOVA; MASKEY; LIU, 2011). It includes the relevant facts reported in the original text. The purpose of this type of summary is to substitute the original document as far as coverage of the information is concerned.

Considering the relation between the summary and the original text, it can be an extractive or abstractive summary (NENKOVA; MASKEY; LIU, 2011). The goal of extractive summaries is to select most informative sentences from the original text. In other hand, the abstractive summarization uses the whole sentences to create a new text using the author's concepts (HAHN; MANI, 2000).

- **Extractive summary.** It avoids any efforts on text understanding to generate a summary. It selects a couple of relevant sentences from an original document(s) in order to use them in a summary. Extractive summaries (NENKOVA; MASKEY; LIU, 2011) are built by extracting the most appropriate sentences based on statistical analysis of the words frequency or the features.

- **Abstractive summary.** It attempts to understand the central concepts of the text and express those ideas in a new text using Natural Language Generation (HAHN; MANI, 2000; CHEUNG, 2008). The biggest challenge for abstractive summarization is the knowledge representation. Systems capabilities are constrained by the richness of their representations, and they cannot summarize what their representations cannot capture (NENKOVA; MASKEY; LIU, 2011). In limited domains, it may be easy to create appropriate structures, but a general-purpose solution depends on a semantic analysis.

Additionally, regarding the number of input documents used to build summary, the summarization can be categorized into *single document* or *Multi-documet* summarization.

- **Multi-document.** It can digest information from multiple sources into one single synthesized document. The significant challenges involving this approach are the repetition of the information, identification of relevant information from all the documents and the creation of summary coherent and non-redundant summary (NENKOVA; MASKEY; LIU, 2011).
- **Single Document.** In other hand, single-document summarization extracts sentences from a single document (NENKOVA; MASKEY; LIU, 2011).

The summarizer can be a monolingual, multilingual or even cross-lingual. It refers to the ability of the summarizer to generate summaries in more than one idiom (MCKEOWN et al., 2002; EVANS; KLAVANS, 2003). In the case of **monolingual**, the output language is the same as the input text. **Multilingual** summarization output language is the same as the text of the entry, but it can work with more than one idiom. A **cross-lingual** summarizer can accept a source text in a particular language and build the summary in another language.

2.3.1 Extractive summarization methods

As presented before, extractive summaries are made by extracting the most relevant sentences based on several methods. Thus, in order to identify the most appropriate method to our problem, we present an overview about the methods used for Extractive summarization, as follow:

- **Graph-Based.** The document(s) is represented by a graph, where nodes represent the sentences and edges represent the similarity between pairs of sentences (ERKAN; RADEV, 2004). Some sentences are more similar to each other while some others may share only a little information with the rest of the sentences. The sentences that are similar to many of the other sentences in a graph are more central or salient to the topic. There are two points to clarify in this definition of centrality. First is how to define *similarity* between two sentences. Second is how to compute the overall *centrality* of a sentence given its similarity to other sentences. We have utilized *LexRank algorithm* (ERKAN; RADEV, 2004) in our experiment to compute the sentence importance. To define *similarity*, LexRank use bag-of-words model to represent each sentence as an N-dimensional vector, where N is the number of all possible words in the target language. For each word in a sentence, the value of the corresponding dimension of the vector representation of the sentence is the number of occurrences of the word in the sentence times the idf of the word. Equation 2.1 show how LexRank compute the similarity between two sentences.

$$idf-modified-cosine(x,y) = \frac{\sum_{w\in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum_{x_i \in x}(tf_{x_i,x} idf_{x_i})^2} \times \sqrt{\sum_{y_i \in y}(tf_{y_i,y} idf_{y_i})^2}} \tag{2.1}$$

where $tf_{w,s}$ is the number of occurrences of the word w in the sentence s.
Algorithm 2.3.1 shows how LexRank computes the overall *centrality* of each sentence. All the values obtained from the algorithm are normalized, giving values from 0 to 1. The sentence with the highest score represents the most informative sentence in a topic.

- **Hypergeometric distribution.** In statistics, the hypergeometric distribution is userd to compute the statistical significance of having drawn a particular k successes out of n total draws from a sample population (LABBE; LABBE, 2005). The test is often used to identify which sub-populations are over-represented or under-represented in a sample. In the ATS, Hypergeometric distribution computes the statistical significance of the sentence based on underused and overused words (LABBE; LABBE, 2005). To better understanding, some necessary definitions must be made:

  U lexical universe of the sentence. Set of tokens or words in each sentence.

  C set of tokens in the entire corpus.

  N_c length of tokens in the whole corpus.

  N_u length of tokens on sentence.

  F_ic and F_iu absolute frequencies of a word i in the whole corpus C.

  F_iu and F_iu absolute frequencies of a word i in the sentence U.

Algorithm 2 Input An array S of n sentences, cosine threshold t output: An array L of LexRank scores

```
$ArrayCosineMatrix[n][n]$
$ArrayDegree[n];$
$ArrayL[n];$
for $i \leftarrow 1$ $to$ $n$ do
  for $j \leftarrow 1$ $to$ $n$ do
    $CosineMatrix[i][j] = idf - modified - cosine(S[i], S[j]);$
    if $CosineMatrix[i][j]$ > t then
      $CosineMatrix[i][j] = 1;$
      $Degree[i] + +;$
    else
      $CosineMatrix[i][j] = 0;$
    end if
  end for
end for
for $i \leftarrow 1$ $to$ $n$ do
  for $j \leftarrow 1$ $to$ $n$ do
    $CosineMatrix[i][j] = CosineMatrix[i][j]/Degree[i];$
  end for
end for
```

The mathematical expectancy $E_i(u)$ of a word i occur in a sample of size N_u tokens drawn out of C can be computed by the equation 2.2.

$$E_i(u) = F_i c * \frac{N_u}{N_c} \tag{2.2}$$

When a token is drawn out of set C, it is not replaced in it. The sample of N_u tokens extracted out of C will be always less or equal to its frequency in C, that is $E_i(u) \leq F_i c$. We must consider that, even in large corpora, the low-frequency words are small. As a precaution, we will utilize a hypergeometric distribution instead a binomial one(LABBE; LABBE, 2005). If the expected value of $E_i(u)$ = $F_i c$ we can say that a word is "neutral". But when two values differ we can say that a word is overutilized or underutilized. In this case, we are interested in finding sentences that contain the less frequent words. To solve this problem, one must consider the probability of the observed value $F_i u$ resulting from the combination of two events:

The number of different possibilities in choosing N_u tokens within C_c:

$$C_c^u = \frac{N_c!}{N_u!(N_c - N_c)!} = \frac{\lceil N_c \rceil}{\lfloor N_u \rfloor} \tag{2.3}$$

The number of differents options in choosing F_iu tokens within F_ic:

$$C_{f_{ic}}^{F_{iu}} = \frac{F_ic!}{F_iu!(F_ic - F_iu)!} = \frac{\lceil F_ic \rceil}{\lfloor F_iu \rfloor} \tag{2.4}$$

The joint probability of these two events follows a hypergeometric law with the parameters F_ic,F_iu,N_u,N_c:

$$P(X = F_{iu}) = \frac{\binom{F_{ic}}{F_{iu}}\binom{N_c - F_{ic}}{N_u - F_{iu}}}{\binom{N_c}{N_u}} \tag{2.5}$$

Where, N_c is the number of the token in the corpus C, N_u the number of token of the sub-corpus U.

F_{iu} and F_{ic} are the absolute frequency of word i in corpus U and C respectively.

When the probability is too high compared to its expectation ($E = F_{ic}\frac{N_u}{N_C}$), then the word is over-used in the sub-corpus. It provides the statistical significance of the sentence based on the proportion of underutilized in the corpus and overutilized in the sub-corpus.

- **Word probability.** Word probability is the simplest way to estimate the sentence relevance based on the probability to occur in a sentence (NENKOVA; MASKEY; LIU, 2011). Equation 2.7 show the probability of word w occur in a single or multi-document.

$$p(w) = \frac{c(w)}{N} \tag{2.6}$$

Were, the likelihood of a word w occur in a document, is the number of occurrences of a word $c(w)$ by the number of all words in the document(s) N.

Given the probability, the likelihood of a summary can be computed based on a multinomial distribution given in the Equation 2.7.

$$L[sum] = \frac{M!}{n! \dots n!} p(w_1)^{n_1} \dots p(w_r)^{n_r} \tag{2.7}$$

Where, M is the number of words in the summary, $n_1 + \cdots + n_r = M$ and for each i, n_i is the number of times word w_i appears in the summary and $p(w_i)$ is the probability of wi appearing in the summary estimated from the input documents.

- **Term Frequency-Inverse Document Frequency (TF-IDF).** The problem of using word frequency to measure the importance of the sentences is that there are few words appearing very often and many other words that appear infrequently. The most frequent words in the sentence include determiners, prepositions and auxiliary verbs or familiar words of the domain. Considering that often words do not indicate a relevant sentence. To solve this problem, TF-IDF (LUHN, 1957; SALTON; BUCKLEY, 1988) weights increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. That fact helps to adjust for the fact that some words appear more frequently in general (NENKOVA; MASKEY; LIU, 2011). The TF-IDF weights also are easy and fast to compute and good indicators of importance. These properties explain why TF-IDF is one of the most commonly used features for extractive summarization (NENKOVA; MASKEY; LIU, 2011). Equation 2.8 show the computation of TF-IDF weights, where $c(w)$ is the term frequency, D is the total of the documents and $d(w)$ is the number of documents that contain the word w.

$$TF - IDF_w = c(w) \times \log \frac{D}{d(w)} \tag{2.8}$$

- **Cluster-based.** The cluster-based method groups similar sentences based on its TF-IDF scores into multiple clusters(NENKOVA; MASKEY; LIU, 2011). Each cluster consists of a group of similar sentences representing a topic or theme in the document. It identifies the themes from the input set of the documents to be summarized and selects the most representative sentences from the appropriate clusters to create the summary.
- **Artificial Neural Networks.** The Artificial Neural Networks (ANN) is a method to estimate the sentence relevance. It is performed by a training process that consist of showing a set of positives and negatives examples of the sentences that should be included in the summary. The learning process is supervised, and the neural network learns about the features presented in the sentences. It can be done by using a Feed-forward neural network, which has been proven to be a universal function approximator (RUSSELL; NORVIG, 1995). The learning process involves presentation of a set of pairs $(x, y), x \in X, y \in Y$ aiming to find a function $f : X \to Y$. Where, x are the set of features of the sentence and y is the binary label that indicate if the sentence is included or not in the summary. The cost function is the relation of mismatches between our mapping and the data and it implicitly contains prior knowledge about the problem domain (NENKOVA; MASKEY; LIU, 2011).
- **Latent Semantic Analysis.** The Latent Semantic Analysis (LSA) is a robust unsupervised method derived from Singular Value Decomposition (SVD). SVD is a useful mathematical method to find principal orthogonal dimensions of multidimensional data. The documents are represented by a matrix A, where each row corresponds to a word that appears in the text and each column corresponds to a sentence in the text. Each entry a_{ij} of the matrix corresponds to the TF-IDF weight of word i in sentence j. If the sentence

does not contain the word, the weight is zero. Standard techniques for SVD from linear algebra are applied to the matrix A in order groups documents that are semantically related to each other, even when they do not share common words. In order to create an extractive summary, they consecutively consider each row of the matrix A to select the sentence with the highest value until the desired summary length is reached (NENKOVA; MASKEY; LIU, 2011).

- **Query-focused.** This method is generally employed on Question Answering (QA) systems. Some sites such ask.fm[1] and Yahoo! Answers[2] allow people to ask and answer questions about diverse subjects as health, politics and software. However, digesting the large amount of information on long topics can be difficult to humans. Given a subject expressed in a short paragraph statement, Query-focused methods can build summary highlights the related sentences in the same theme (WANG et al., 2014).

2.3.2 Features for extractive summarization

As exposed, the central process of extractive summarization is the selection of the most relevant sentences from the original document to be included in the summary. Such process of selection can be guided based on some features present in the sentence (KYOOMARSI et al., 2008; GUPTA; LEHAL, 2010). Thus, to identify the most appropriate method to our problem, we presents an overview of the main features used in extractive summarization, as follow:

- **Semantic Similarity.** This feature measures the semantic similarity between two words. Lesk (LESK, 1986) proposed that the relatedness of two words is proportional to the extent of overlaps of their dictionary definitions. Banerjee and Pedersen (BANERJEE; PEDERSEN, 2002) extended this notion to use WordNet as for the word definitions. WordNet. WordNet in the version 1.7, there has 107.930 noun arranged in 74.448 synsets, 10.860 verbs in 12,754 synsets, 21.365 adjectives in 18,523 synsets, ans 4.583 adverbs in 3.618 synsets. Each synset has an associated definition or gloss. It consists of a short entry explaining the meaning of the concept represented by the synset. Synsets are connected to each other through a variety of semantic relations.
- **Keywords.** Keywords are words that represent the main topics discussed in the text. One way to extract the most relevant sentences in the text is checking the presence of keywords (NENKOVA; MASKEY; LIU, 2011). However, the automatic determination

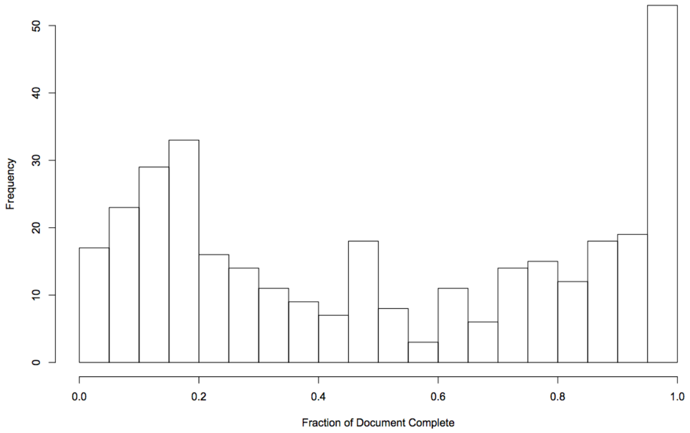
[1] http://www.ask.fm

[2] http://answers.yahoo.com/

of the keywords is a challenging due to the intricate complexities of natural language.

- **Title Word.** Considering that the main idea of the text should be addressed to the title, the words present in the title should be an indicative of importance. The sentences that contain these words would have more chances to express the central idea of the text (NENKOVA; MASKEY; LIU, 2011).
- **Sentence location.** Studies have shown that the sentences from the first and last paragraphs are more relevant. It means that they have more chances to express the main topics of the text (KUPIEC et al., 1998). Figure 2.2 shows the histogram of the location of the extracted sentences. There are two peaks in this plot, being one early in the document while the other is being in the final five percent of the text. Based on it, the sentences that are in the beginning and the end have more chance to be relevant.

Figure 2.2 – Quotation location within document



Source: (KUPIEC et al., 1998)

- **Sentence Length.** Very short sentences have a small probability to be an informative sentence (CHEN; HAN; CHEN, 2002). In other hand, the long sentence has many words to express an idea. The work (CHEN; HAN; CHEN, 2002) considers the sentence relevance being the inverse proportion to the square root of its length. In the work of (RADEV et al., 2004) shows a use of fixed threshold of 5 words.
- **Proper Noun.** A Proper noun is a name of a person, place and concept. The sentences containing proper nouns have greater chances to be included in the summary (ZHUANG; JING; ZHU, 2006).

- **Cue-Phrase Feature** This feature was introduced by Edmundson (1969) to indicate sentence importance by the identification of the significant words. Using this approach, “in summary”, “in conclusion”, and superlatives such as “the best”, “the most important” can be useful indicators of significant relevance.
- **Sentence-to-Centroid Cohesion.** Its measures the cohesion in relation to a centroid that represents the fundamental concept of the text (HABOUSH et al., 2012). Thus, it uses bag-of-words model to represent each sentence as an N-dimensional vector, where N is the number of all possible words in the original document. For each word in a sentence, the value of the corresponding dimension in the vector representation of the sentence is the number of occurrences of the word in the sentence. The similarity between the centroid and the sentence is obtained by the cosine value between centroid to the related sentence. The normalized value for a sentence s is obtained by computing the ratio of the cosine value over the largest value among all sentences in the document. Sentences with feature values closer to 1 have a greater degree of cohesion regarding the centroid of the original document and should represent the fundamental concept of the text.
- **Sentence-to-Sentence Cohesion** Similar to Sentence-to-Centroid, this feature compute the similarity between s and other sentence s' (HABOUSH et al., 2012). The normalized value for a sentence s is obtained by computing the ratio of the value for s over the largest value among all sentences in the document. Values closer to 1 indicate that the sentences have hight cohesion.
- **Occurrence of non-essential information.** The occurrence of words “because”, “furthermore”, and “additionally” can indicate a non-essential information (NENKOVA; MASKEY; LIU, 2011). Informative summaries should avoid non-informative sentences.
- **Discourse analysis.** This feature determines the overall discourse structure of the text in order to remove all the peripheral sentences. It allows a creation of coherent summaries containing just central topics of the original document (CHAN et al., 2000).

2.4 Evaluation methods of Automatic Text Summarization

According to Uman et al. (2015), a good summary should contain main topics of the original document (**coverage**) while keeping the redundancy to a minimum (**high diversity**) and smooth connection among sentences (**high coherence**). Several methods have been proposed to evaluate the summary quality, and the better comprehension of these methods is essential to perform a correct evaluation of the system. In this Section, we provide an overview of different methods for evaluating automatic summarization systems, its implications, as well as present the pros and cons involved in this choice.

Methods for evaluating ATS can be broadly classified into two categories (NENKOVA; MASKEY; LIU, 2011). The first, the *Intrinsic* methods are frequently employed soliciting human judgments on the goodness and utility of a given summary, or by a comparison of the summary with a human-authored gold-standard. The second, the *Extrinsic* methods that aim to measure on how it affects the completion of some other task.

2.4.1 Extrinsic Evaluation

Extrinsic evaluations measure the impact of summarization on tasks of relevance assessment and reading comprehension. Relevance assessment can be performed by showing the judges a document (summary or source) and a topic, then asking them to determine whether the document is relevant to the topic. On average, if the choices for the document and the summary are the same, then the summary get a high score on relevance. The evaluation of comprehension, judges receive a document (the original document or the summary) and response a set of questions. Their answers to these questions determine their understanding of the text. If the answers based on the summary and the original document are similar, then the summary is positively evaluated (NENKOVA; MASKEY; LIU, 2011).

2.4.2 Intrinsic Evaluation

The intrinsic evaluations methods have assessed mainly the coherence and informativeness of summaries. One way to perform an intrinsic evaluation is comparing the summary generated by the system with a *gold-standard summary*[3] written by humans. In this case, evaluation can be quantitative and measure aspects such as precision and recall. The major problem with this approach is to make the people agree on what constitutes a "gold" summary (NENKOVA; MASKEY; LIU, 2011). The most relevant intrinsic evaluation methods are:

- **Precision and Recall.** A human is asked to choose sentences that seem to communicate the meaning from the original document to be summarized. Then the sentences selected

[3] In the literature review, the human summaries against which other summaries are compared are also called *models, references or models*

by the summarizer are evaluated against the human selections. In such settings, the typical information retrieval metrics of precision and recall can be used to evaluating a new summary (NENKOVA; MASKEY; LIU, 2011). Recall, showed at equation 2.9, is the relation of sentences are chosen by the person that were also correctly identified by the system. Precision, showed at equation 2.10, is the fraction of system sentences that were correct (NENKOVA; MASKEY; LIU, 2011).

$$Recall = \frac{|system - human\ choice\ overlap|}{|sentences\ chosen\ by\ human|} \tag{2.9}$$

$$Precision = \frac{|system - human\ choice\ overlap|}{|sentences\ chosen\ by\ system|} \tag{2.10}$$

The use of precision and recall as evaluation metrics can be used many times to evaluate automatically by a mere comparison of the sentence. However, there are several problems with these measures, as follow:

Human Variation. Different people tend to choose different sentences when asked to construct an extractive summary. Research as early (RATH; RESNICK; SAVAGE, 1961) reported that extracts selected by six different human judges for ten articles from Scientific American had only 8% overlap on average. It has been shown (DONAWAY; DRUMMEY; MATHER, 2000) that the same summary can obtain a recall score with between 25% and 50% difference depending on which of two available human gold summary are used for evaluation. Thus, a system can extract good sentences, but still be penalized in precision and recall assessment methods. It also seems that is more beneficial to concentrate on recall rather than precision (NENKOVA, 2006a). Precision is overly strict because some of the sentences chosen by the system might be good, even if the gold standard creator has not chosen them. Recall, on the other hand, measures the overlap over the already observed sentence choices.

Granularity. Another problem with the precision and recall measures is the fact that sentences are not the best granularity for measuring content (NENKOVA; MASKEY; LIU, 2011). Sentences are different in word length and convey different amounts of information. Selecting a longer and more informative sentence can be more desirable than choosing a short sentence. For instance, a gold standard summary containing the sentences: (1) "We need urgent help." and (2) "Fires have spread in the nearby forest, and threaten several villages in this remote area.”. Considering two system, each choosing only one sentence of the gold standard, one choosing sentence (1), and other one selection the sentence (2). Both summaries will have the same precision and recall score, but they are not equally informative (NENKOVA, 2006a).

Semantic Equivalence. Another problem with using sentences as the core of evaluation is that two distinct sentences can express the same meaning (NENKOVA; MASKEY; LIU, 2011). It can occur in summarization of news, and it is very frequent in multi-document summarization, in which the source documents consists of many articles on the same subject. Again, a human would select one of the equivalent sentences, but a system will be penalized for the utilization of an alternate sentence that expresses the same meaning (JING et al., 1998). Few alternative evaluation measures were designed to address the issues that were raised regarding precision and recall measures. Thus, it has been suggested to use multiple models rather than a single person's judgment (JING et al., 1998). Smaller, more semantically oriented units of analysis have been proposed, and more emphasis has been given on recall (NENKOVA; MASKEY; LIU, 2011).

- **Relative Utility.** Relative utility (RADEV; TAM, 2003) is a way to solve the human variation and redundancy problem using precision and recall as metrics. Like some other evaluation metrics, it compares sentence selection between machine and "gold standard" summary created by a human. Relative utility approach uses multiple judges to score each sentence in the source document using a scale from 0 to 10. Being a score of 10 indicate that a sentence is central to the topic of the cluster while a score of 0 marks an entirely irrelevant sentence. The judges also explicitly mark which sentences are mutually substitutable in function of the semantic equivalence. Thus, the summaries having sentences semantic equivalent to the gold standard are better evaluated.
- **ROUGE.** Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (LIN, 2004) is inspired by the success of the Bilingual Evaluation Understudy (BLEU) (PAPINENI et al., 2002) that uses n-gram overlap as a based on the measure. It is based on the comparison of n-grams between the summary to be evaluated and the "gold standard". ROUGE has been preferred for summarization because it is recall-oriented, unlike BLEU, which emphasizes precision (PAPINENI et al., 2002). ROUGE is one of most used measures of content selection quality to be fast and easy. Many researchers usually use the ROUGE with supplementary manual evaluation such as Pyramid (POIBEAU et al., 2012).
- **DUC Manual Evaluation.** The Document Understanding Conference (DUC) and Text

Analysis Conference (TAC) has been woking on in assessments methods to the summarization systems. They developed cheap methods to perform a comparation between "gold standards" to summarization system. Thus, they have created summary to be used as the gold standard. The DUC evaluations methods used between 2001 and 2004 are based on a single human model. After that, in order to mitigate the bias coming from the use of one person, different annotators created models for various subsets of the test data (HARMAN; OVER, 2004). Other improvements were the adoption of elementary discourse units (EDU) to perform more fine-grained analysis than the sentence level. The evaluations of the summary were conducted analyzing which they cover each EDU in the model. The overall score, called coverage, was the average score across all EDU in the model. DUC also started using human-generated abstracts as gold standards, but this evaluation method is expensive, requiring more human involvement. Nowadays, the DUC evaluation protocol was no longer used, giving rise to the Pyramid manual evaluation.

- **Pyramid Method.** The Pyramid evaluation method (NENKOVA; PASSONNEAU; MCKEOWN, 2007) uses multiple human abstracts to derive a gold standard. In this process, information with the same meaning, even when expressed using different words, is marked as showing the same summary content unit (SCU). Each SCU is assigned a weight equal to the number of human summaries who expressed the SCU in their summaries. SCU analysis shows that summaries that have different content can be equally useful and assigns a stable score. A disadvantage of the approach is that the process is very labor intensive, even the fact that an annotation tool called DUCView5 were developed to facilitate the process. The approach was primary developed to evaluate abstractive summaries, and not efficient to extractive summaries.
- **Linguistic Quality Evaluation.** While researchers are working on improving system content selection, most automatic summaries have poor readability aspects such as coherence and referential clarity (NENKOVA, 2006b). Recent interest in sentence ordering and referential cohesion have led to a proposal for automatic evaluation of cohesion (LAPATA; BARZILAY, 2005) have been improving automatic summary readability.
  Human assessments on a scale, usually from 1 to 5, are probably the fastest and cheapest way to evaluate the readability aspects. It does not require the collection of gold standard summaries, nor any annotation or manual analysis from the assessors in order to come up with a summary quality score. Because of these properties, this evaluation approach is rather attractive especially when many systems have to be compared on many inputs (NENKOVA; MASKEY; LIU, 2011).

2.5 Machine Learning

Most of the review summarization strategies relies on Machine Learning methods to mining product features that have been commented on by customers. Since we aim to discover why people like and dislike about specific product, discovering product features is a crucial task. Then, we present next, in Section 2.5.1, an overview of machine learning. Finally, in Section 2.5.2, we introduce the machine learning approaches commonly used in review summarization.

2.5.1 Overview

Machine Learning (MITCHELL, 1999) is the study of algorithms that can learn from an experience to make predictions on data. Regarding the learning task, machine-learning techniques are categorized as *supervised* or *unsupervised*. Supervised methods are based on training data set with examples of inputs and their desired outputs. It aims to learn a general rule that maps the inputs to outputs. On the other hand, *unsupervised* techniques data set without labels are given to the algorithm, leaving it on its own to find structure in its input. This type of learning is used to discovering hidden patterns in data.

Regarding our problem, we want to analyze item, user and review dimensions in order to extract useful patterns, such as: *this hotel is liked by young travelers* or *middle-aged Frenchs enjoy this restaurant*. Thus, this section presents an overview of the machine learning techniques utilized to find these patterns.

2.5.2 Machine Learning techniques

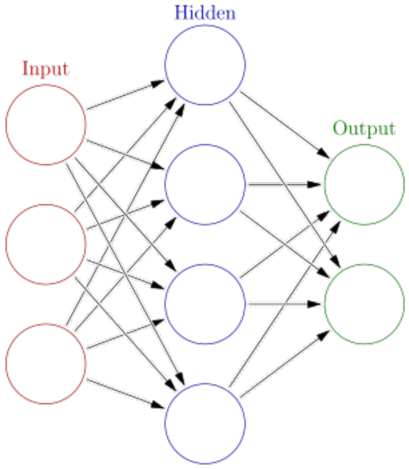
Machine Learning is composed by a couple of techniques used for different purposes, each method offers advantages and disadvantages depending the use. Such techniques are:

- **Association Rules Learning.** It is a statistical correlation between the occurrences of certain attributes in large datasets. It is formally defined as: Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the database. Each transaction in D contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items (for short itemsets) X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively (RUSSELL; NORVIG, 1995). For example, the association rule $\{\text{onions}, \text{potatoes}\} \Rightarrow \{\text{hamburger}\}$ present in the sales dataset of a supermarket would indicate that if a customer buys onions and potatoes together, they frequently also buy hamburger.

FP-growth algorithm is one of the currently most fast and most modern algorithms for association rule learning. It is based on a prefix tree representation of the given database of transactions, and this approach can save a considerable amount of memory for storing the transactions (HAN; PEI; YIN, 2000). The basic idea of FP-growth algorithm is: In the first pass, the algorithm counts the occurrence of items (attribute-value pairs) in the dataset. In the second pass, it builds the FP-tree structure by inserting instances. Items in each case have to be sorted in descending order of their frequency in the dataset so that the tree can be processed quickly. Items in each instance that do not meet minimum coverage threshold are discarded. If many cases share most frequent items, FP-tree provides high compression close to the tree root. Recursive processing of this compressed version of primary dataset grows large itemsets directly, instead of generating candidate items and testing them against the entire database. Growth starts from the bottom of the header table, by finding all instances matching given condition. The new tree is created, with counts projected from the original tree corresponding to the set of cases that are conditional on the attribute, with each node getting the sum of its children counts. Recursive growth ends when no individual items dependent on the attribute meet minimum support threshold, and processing continues the remaining header elements of the original FP-tree. Once the recursive process has completed, all large itemsets with minimum coverage have been found, and association rule creation begins.

- **Artificial Neural Networks.** Artificial neural networks (ANNs) is part of the statistical learning family algorithms inspired in a neural network of a brain. It is often used to approximate functions that can depend on a large number of inputs. Like in the human brain, the strength of neuron interconnections can change when a stimulus occurs, which enables the ANN to learn. For example, a neural network for character recognition is defined by a set of input neurons wich read the pixels of an input image. After being weighted and transformed by a function, the activations of these neurons are then passed on to other neurons. This process is repeated continuously until the learning of pattern is reached. It determines which character was read. Figure 2.3 shows a system having three layers. The first layer has input neurons that send data via synapses to the second layer of neurons, and then via more synapses to the third layer of output neurons. More complex systems will have more layers of neurons with some having increased layers of input neurons and output neurons. The synapses store parameters called weights that manipulate the data in the calculations.

Figure 2.3 – Three-layer Artificial Neural Network

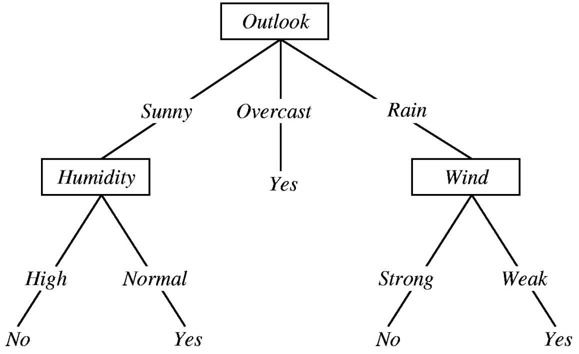


Source: Developed by the author

- **Decision Trees.** It is a predictive model that maps observations about an item to conclusions about the item's target value. Tree models, where the target variable can take a finite set of values, are called classification trees. In these tree structures, leaves represent class labels, and branches represent conjunctions of features that lead to a class. Decision trees, where the target variable can take continuous values, are called regression trees. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node or leaf. When a terminal node is reached, a decision is made. Decision trees can also be interpreted as a particular form of a rule set, characterized by their hierarchical organization of rules. Given the training set composed of the weather and the decision of play or not play tennis observations. Figure 2.4 represents a decision tree of how a decision is made.

- **Bayesian Belief Network.** Bayesian belief networks (BBN) is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). The nodes represent attribute variables, and the edges represent probabilistic dependencies between the attribute variables. Associated with each node are conditional probability distributions that describe the relationships between the node and its parents. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the BBN can be used to compute the probabilities of the presence of various diseases.

Figure 2.4 – Decision Tree of the play tennis problem



Source: (RUSSELL; NORVIG, 1995)

- **Genetic algorithms.** It is a search heuristic that are inspired by the principles observed in natural evolution. This heuristic is routinely used to generate solutions to optimization and search problems. The process starts with a population of candidate solutions called individuals. Each candidate solution has a set of properties called chromosomes or genotype that can be mutated and altered. Traditionally, solutions are represented in binary as strings. The evolution usually starts from a population of randomly generated individuals and is an iterative process, with the population in each iteration called a generation. In each generation, the fitness of each person in the population is evaluated. The fitness is usually the value of the objective function in the optimization problem being solved. The more fit individuals are stochastically selected from the current population, and each individual's genome is modified or recombined to form a new generation. The new generation of candidate solutions is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced or a satisfactory fitness level has been reached for the population.
- **Clustering.** Clustering is used for finding groups of with similar features. For example, given a data set of customers, identify subgroups of customers that have a similar buying behavior. It discovers groups and structures in the dataset that are in some way similar.

2.6 Final Remarks

In this Chapter, we presented an overview of concepts necessary to conduct this research. It provides a better understanding to the task of choose the most appropriate methodology for the problem that this research is addressed. Thus, we presented the Natural Language Processing techniques necessary to build short explanatory statements that could explain why some people like or dislike an item. After, we provide an overview of the Automatic Text Summarization techniques necessary to contrast these extracted patterns with relevant sentences of the reviews. Finally, we presented the most significant Machine Learning techniques necessary to perform the analysis of item, user and review dimensions to extract useful patterns from them.

3 RELATED WORK

People usually consult their friends to get an opinion about a book, or a movie or a restaurant. With the growth of the web over the last decade, users have generated opinions on collaborative reviewing sites. Unfortunately, the vast amount of such opinions becomes difficult to users. Thus, researchers on opinion summarization have introduced various techniques and paradigms to solve this relevant problem.

The related work in this area can be classified into those that rely on the presence of aspects, and those that do not require aspects. An *aspect* (also called feature) is an attribute or component of an item that has been commented on in a review. For instance, the 'soundtrack', 'art director' or 'the cast'. In this chapter, we present these two main classes of studies in review summarization, dividing the studies by *aspect-based* and *non-aspect-based* approaches. First, in Section 3.1, we provide an overview about Automatic Review Summarization. Next, in Section 3.2, presents approaches focuses on Aspect-based. In Section 3.3 show approaches focuses on Non-Aspect-based. Finally, in the Section 3.4, presents the synthesis of the related work.

3.1 Overview

Opinion and Review are usually addressed in the literature with the same sense. However, Review is a particular subtype of opinion, generally obtained in collaborative reviews sites like IMDb. A review can be viewed as a tuple <u,i,r,w,t>, that represents a text review w and rating r about an item i wrote by a user u, in a particular period t. The user u contains dimensions such as location, age, and gender. The item i also have dimensions such as director and actors for a movie. The rating r express a relative acceptance or rejection of a related item by a user. The users can also describe their reasons for acceptance or rejection of a particular item using a textual explanation w. For example, in a movie review the users may express their opinion about *aspects/features* of the movie.

Currently, there are three extensible surveys that are related to the study of opinion summarization (PANG; LEE, 2008; KIM et al., 2014). Our research aimed at treating overload of Reviews, but, of course, many works on opinions are related. First, Pang's book (PANG; LEE, 2008) covers various techniques in opinion mining and summarization. Second, Kim's survey (KIM et al., 2014) provides a comprehensive review of the work of this area. Finally, Chapter 11 of Liu's book (LIU, 2007) introduces basic concepts and definitions about opinion mining.

3.2 Aspect-based Opinion Summarization

The summary based on aspects provide a detailed overview of a particular item, dividing the input text into aspects (also called features). The segmentation of the text into aspects can provide quick interpretation of the summary. In divergent opinions, aspect-based opinion summarization can be useful because they expose separately the overall opinion of each aspect. It allows a direct comparison between positive against negative aspects. It is very popular and has been heavily explored over the last few years (LIU; HU; CHENG, 2005; HU; LIU, 2004; KU; LIANG; CHEN, 2006; MEI et al., 2007).

The aspect-based summarization involves the task of building summaries based a set of aspects (or features). Usually, aspect-based summarization is based on three distinct steps: (i) aspect/feature identification; (ii) sentiment prediction; and (iii) summary generation. First, feature identification task is used to discovery aspects within the text. Second, the sentiment prediction task is used to set the overall sentiment polarity (negative or positive) about the aspect. Finally, the summary generation step is used build the results of the previous step in an effective way. Various techniques have been proposed to solve the challenges in each of these steps. In the following three subsections, we will describe the common techniques used for this purpose.

3.2.1 Aspect identification

The Aspect identification is the task of discover the item’s aspects within the text. For instance, if we want create a summary about the movie "Donnie Darko", some of the expected aspects are ’soundtrack’, ’art director’ and ’cast’.

Natural Language Processing (NLP). It relies on Part-Of-Speech (POS) tagging method (see Section 2.1) to discover aspects candidate (POPESCU; ETZIONI, 2007; HU; LIU, 2004). It is a common starting point for feature discovery. However, the NLP techniques may not sufficient to discover all the aspects of the text. It because the features are not always nouns, and often they are not explicitly in the text. For instance, the sentence, "While light, it will not easily fit in pockets.", Implicitly the sentence refers to the ‘size’ aspect, but there is no mention of the word ‘size’ in the sentence. For solve this problem, some domain knowledge ontological word dictionary is required (KIM et al., 2014).

Probabilistic Latent Semantic Analysis (PLSA) has recently been applied to perform the aspect extraction (LU; ZHAI; SUNDARESAN, 2009; HENNIG; LABOR, 2009). To identify multiple topics in the text, it uses a Mixture Model to define the parameters of the multiple word distributions. The result had shown that is a promising approach to discover hidden aspects within the text.

Query-Based opinion summarization generates a summary based on the users' question. It tries to answer the question recovering the aspects that the user needs (JIN; HUANG; ZHU, 2009). Several modules are needed to perform it: (i) question analysis and query reformulation module; (ii) latent semantic analysis for topic detection; (iii) sentence polarity detection module; and (iv) redundancy removal module. For instance, the question "*What is free at Bestwestern Inn, San Francisco?*" the summary generated is "*Free wine reception in evening. Free coffee and biscotti and wine.*". The work (WANG et al., 2014) uses a submodular function-based to produce a summary based on aspects the user need. It uses a statistical ranker to rank the sentences by their relevance. The results have shown that the approach provides a high coverage of the information and little redundancy in their summaries.

Mining Techniques for Feature Discovery. Data mining techniques can be used way to discover features into the text (HU; LIU, 2004). Association Rules is the most common method used to identify frequent itemsets. In this context, item sets are frequent words that could indicate an aspect. It requires a manual tagging of words that represent aspects in the text. With the segmented and tagged data set, Association Rule Mining is performed. It learns rules in the form $A_1, A_2 \ldots A_n \rightarrow [aspect]$, where A_n is the remaining words in the text. After the learning process, an aspect can be discovered without the presence of the word that represent the aspect.

3.2.2 Sentiment Orientation

After the aspect discovery, the next step is often the sentiment prediction. It predicts the people's sentiment orientation (positive or negative) about an aspect. For instance, some people can think that the '*soundtrack*' of the movie '*Donnie Darko*' is '*positive*', while other persons think that the 'cast' is '*negative*'. Thus, the results of aspect-based sentiment predictions would help users to digest the general sentiments on the aspect.

Learning-based Methods can be applied to discover the sentiment orientation of the aspect. The problem can be viewed as text classification problem. Machine learning method can capture context to predict the sentiment orientation. The work (PANG; LEE; VAITHYANATHAN,

2002) compare the efficiency of Naive Bayes, Maximum Entropy and Support Vector Machines (SVM) techniques to predict the sentiment orientation of the text. The study showed that SVM provides better results to predict sentiment orientation.

Lexicon Methods relies on a word dictionary to predict sentiment orientation. The dictionary contains a list of positive and negative words used to match the sentence's word to discover the overall opinion orientation of the sentence. For instance, if the sentence has many words classified as positive and few negative, it is classified as a positive sentence. The work (KAMPS et al., 2004) have used WordNet[1] to measuring the semantic distance (or similarity) between sentence's words to the keywords "good" and "bad". Additionally, the study of (TURNEY, 2002) have used pointwise mutual information (PMI) as the semantic distance between two words. However, the work (TURNEY; LITTMAN, 2003) show better results by the using the cosine as the distance measure on latent semantic analysis (LSA) method.

3.2.3 Summary Generation

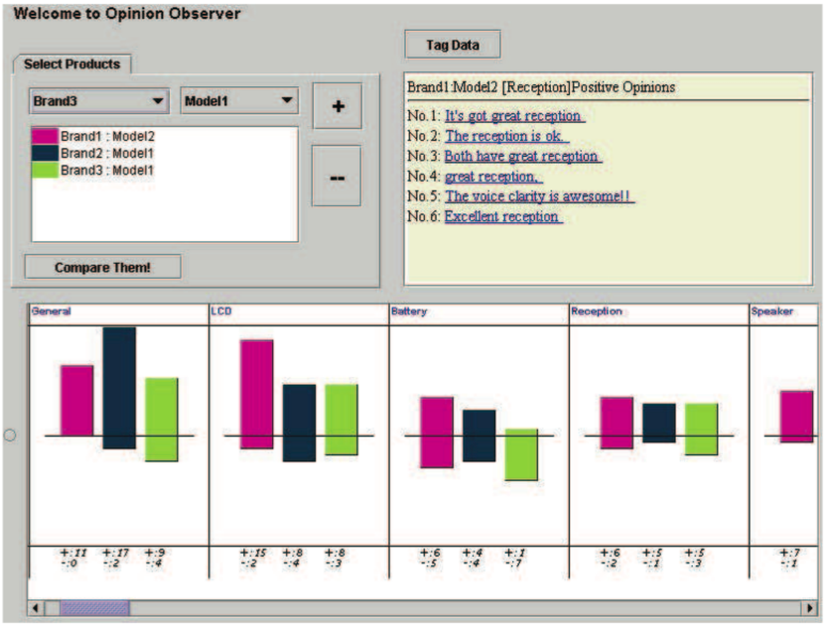
Given the aspects and the sentiment about them, the next task is to use them to present a useful summary to the user. It involves aggregation of the aspects and the sentiment into the concise summary. Several methods have been employed in this way, as follows:

Statistical Summary. It is an easy way to present aspects and sentiment about them. It gives the statistics about the number of the persons that have provided a negative or positive opinion in a format of the list. It allows to the user a quick interpretation of general opinion in each aspect. The work of (LIU; HU; CHENG, 2005) presents a graphic summary of product's feature with multiples bars. Figure 3.1 is a graphical interpretation of the opinion. The portion of a bar projecting above the centered "horizon" line represents the number of positive opinions. The part of the bar below the line represents the number of negative opinions. It allows an easily simultaneously comparative between negatives against positives.

Text Selection. Many approaches extract a couple of sentences that expose the reasons for the positive or negative opinion. However, selecting a sentence that describe an opinion is not a trivial task. Many of recent studies (TITOV; MCDONALD, 2008; POPESCU; ETZIONI, 2007; GANESAN; ZHAI; HAN, 2010) tries to extract from the text smaller pieces of text to contextualize the opinions. Thus, they use different granularities of text, such as word and sentence as level granularities.

[1] http://wordnet.princeton.edu

Figure 3.1 – Visualization of an aspect summary



Brand1:Model2 [Reception]Positive Opinions

No. 1: It's got great reception
No. 2: The reception is ok.
No. 3: Both have great reception
No. 4: great reception,
No. 5: The voice clarity is awesome!!
No. 6: Excellent reception

Source: (LIU; HU; CHENG, 2005)

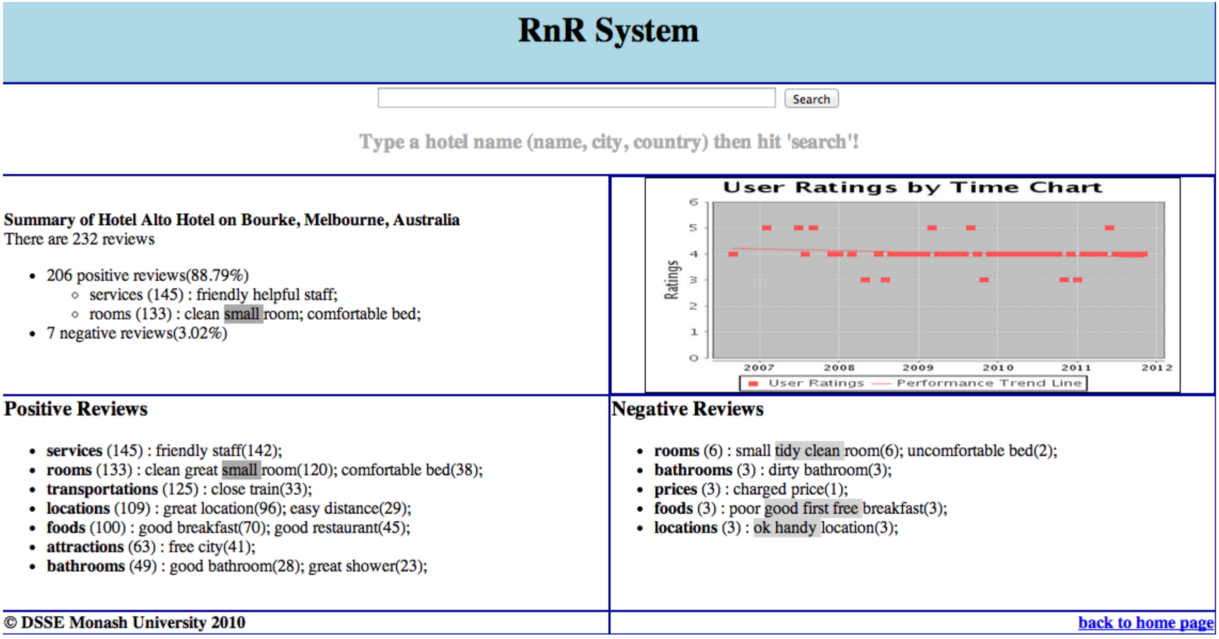
Rated aspect is a way to provide aspect summary that combines statistical and sentence extraction methods (LU; ZHAI; SUNDARESAN, 2009). The system combines aspect, rating and text selection to form a summary. For each aspect, it provides a rating average of cited reviews and a sentence that justify that rating.

Timeline summary provides a graphical interpretation of the opinion, showing opinion trends over a timeline (MEI et al., 2007; KU; LIANG; CHEN, 2006). While other approaches show the opinion based on current data, the timeline allows an analysis over the time. It provides us a analyzing the events that happened at the drastic opinion change. For instance, in a political analysis we can quickly identify that there is a radical opinion changing after a political debate.

Contrastive Opinion Summarization. While, general methods of opinion summarization present negative and positive sentences, aspect-based methods divides the sentences according to its aspects (LIU; HU; CHENG, 2005; HU; LIU, 2004; KU; LIANG; CHEN, 2006; MEI et al., 2007). Frequently, there are sentences with mixed orientation. For instance, the sentences, 'The battery life is long when we rarely use buttons' and 'The battery life is short when we use buttons a lot' would be classified as positive and negative respectively, but they are actually saying the same fact. Opinion Digger (MOGHADDAM; ESTER, 2010) creates abstractive summary based on predefined set of aspects to provide a comparison between them.

Unified method. While others approaches provide one type of summary, the Unified method provides an extensible summary that covers multiples methods. The RnR system (LABBE et al., 2010) captures and summarizes the key rationale for positive and negative opinions expressed in a corpus of reviews and highlights the negative features of positive reviews. Figure 3.2 shows the summary of a hotel in a four-quarter screen. In the left top, we can see statistics about the overall opinion of the hotel. The top right contains performance timeline of their ratings. The bottom gives details of each positive and negative aspects identified in the reviews.

Figure 3.2 – The RnR system output



RnR System

Type a hotel name (name, city, country) then hit 'search'!

Summary of Hotel Alto Hotel on Bourke, Melbourne, Australia
There are 232 reviews

- 206 positive reviews(88.79%)
  - services (145) : friendly helpful staff;
  - rooms (133) : clean small room; comfortable bed;
- 7 negative reviews(3.02%)

Positive Reviews

- **services** (145) : friendly staff(142);
- **rooms** (133) : clean great small room(120); comfortable bed(38);
- **transportations** (125) : close train(33);
- **locations** (109) : great location(96); easy distance(29);
- **foods** (100) : good breakfast(70); good restaurant(45);
- **attractions** (63) : free city(41);
- **bathrooms** (49) : good bathroom(28); great shower(23);

Negative Reviews

- **rooms** (6) : small tidy clean room(6); uncomfortable bed(2);
- **bathrooms** (3) : dirty bathroom(3);
- **prices** (3) : charged price(1);
- **foods** (3) : poor good first free breakfast(3);
- **locations** (3) : ok handy location(3);

© DSSE Monash University 2010

back to home page

Source: (LABBE et al., 2010)

3.3 Non-aspect-based Opinion Summarization

There are many studies that not rely on aspects to provide a summary (RADEV et al., 2004; LIU; HU; CHENG, 2005; HU; LIU, 2004; KU; LIANG; CHEN, 2006; MEI et al., 2007). Additionally, they provide different types of interpretation. We can categorize them into basic sentiment summarization, advanced text summarization, visualization, and entity-based summarization (KIM et al., 2014).

3.3.1 Basic Sentiment Summarization

Given the sentiment orientation obtained in Sentiment prediction (see Section 3.2.2), a basic summary can be built from them. A simple summary can be generated based on the count of how many positive/negative opinions have an item. This type of summary can show overall opinion distribution of input text without a sophisticated aspect analysis.

3.3.2 Text Summarization

Naturally, the task of review summarization is related to many other works on multi-document summarization. ATS approaches are broadly utilized for the general purpose, from News to scientific paper summarization. In the opinion summarization domain, they are employed to digest the vast amount of text and extract the most relevant sentences that can explain user opinion. Regarding how the summary was generated, it can be categorized in *Abstractive* and *Extractive* summarization. Considering how it exposes the sentiment orientation, it can provide *Constrastive Opinion Summary* (COS).

Abstractive Text Summarization is the task of creating a new text using the author's concepts (see Section 2.3). The biggest challenge for abstractive summarization is the knowledge representation. Systems capabilities are constrained by the richness of their representations, and they cannot summarize what their representations cannot capture (NENKOVA; MASKEY; LIU, 2011). In limited domains, it may be easy to create appropriate structures, but a general-purpose solution depends on a semantic analysis. The Opinosis (GANESAN; ZHAI; HAN, 2010) receives as input a set of free text reviews, and it performs an extraction of the aspects based on the graph representation of the source documents. This approach was primarily used to summarize opinions, But the authors also related their use on blogs and news articles. For instance, considering the food at Holiday Inn, London" the summary generated by the opinions is "*The food was excellent, good and delicious. Excellent selection of food*".

Extractive text summarization avoids any efforts on text understanding to generate a summary. It selects a couple of relevant sentences from an original document(s) in order to use them in a summary (see Section 2.3). The MEAD framework (RADEV et al., 2004) implements multiple algorithms such as position-based, centroid-based, largest common subsequence, and keywords. Additionally, it provides a tool to support evaluating tasks. Open Text Summarizer[2](ROTEM, 2012) is another example of Automatic Text Summarization. It reads a text and

[2] OTS available at http://libots.sourceforge.net/

decides which sentences are relevant based on its frequency. However, there is no way to the summarizer filter irrelevant information because it is based on the frequency of words. Its use in the review domain can result in irrelevant and biased summary, such as: “booking was done at very last minute...”, “I did a lot of research...”.

Multilingual opinion summarization refers to the ability of the summarizer to generate summaries in more than one language (MCKEOWN et al., 2002; EVANS; KLAVANS, 2003) (see Section 2.3). NewsBlaster (MCKEOWN et al., 2002) is one example of multilingual summarizer. It addresses the problem of users have on browsing news from multiple languages on sites on the internet (MCKEOWN et al., 2002). The system automatically collects, organizes, and summarizes news in multiple source languages, allowing the user to browse news topics with English summaries. Additionally, it provides a comparison of the perspectives from different countries on the same topics. SimFinderML (EVANS; KLAVANS, 2003) is another example of multilingual summarization. It identifies similar pieces of text by computing similarity over multiple features. Given a set of documents about a specific topic, the goal is to identify text units about the same event. Such identification is performed in five steps: (i) identifying primitives in each language; (ii) translating primitives between languages; (iii) computing feature values across primitives and translations of primitives; (iv) merging feature values into a single similarity value; and (v) clustering text units based on the computed similarity values.

3.3.3 Visualization

The methods for visualization also plays an important part in the interpretation of the review information. While many approaches have used a table-based statistical summary with relevant text to the user (LIU; HU; CHENG, 2005; HU; LIU, 2004; MEI et al., 2007), others provides an interpretation of the opinions by a histogram (LIU; HU; CHENG, 2005). Additionally, the work (CHEN et al., 2006) with different graph structures to analyze the features that have influenced the user’s opinion about a particular item.

3.4 Final Remarks

To the best of our knowledge *there is no work that provides personalized summaries to contrast reviews written by different segments of reviewers*. Indeed, most of the richness of an opinion is contained in free-text comments whose information is still challenging to extract automatically. New solutions must be found to digest the information carried in individual reviews and create personalized summaries. Unfortunately, the ability to explain user reviews is an under-explored research question (LABBE et al., 2010; GANU; ELHADAD; MARIAN, 2009).

To better understanding, the nex table shows a summary of the related work presented. The First column shows the authors name and publication year. The second column describes the input formats. The third column gives information about the domain. The fourth column describes the methods of the system. Finally, the last column provides information about the evaluation.

Table 3.1: Related work on opinion summarization

SYSTEM	*Input*	*Domain*	*Features*	*Evaluation*
Opinosis, (GANESAN; ZHAI; HAN, 2010), 2010	-Multi-Document -Monolingual (English)	-Specific (Opinion)	-Abstractive -Graph-Based -Avoids redundant text	-Intrinsic (ROUGE)
Minqing Hu and Bing Liu, (HOVY; LIN, 1998), 2004	-Multi-Document -Monolingual (English)	-Specific (Opinion)	-Abstractive -POS tagging -Frequent feature identification -Feature Pruning -Sentiment Analysis	Extrinsic
CBSEAS, (BOSSARD; GÉNÉREUX; POIBEAU, 2009), 2009	-Multi-Document -Monolingual (English)	-Specific (Opinion)	-Extractive -Clustering-Based (k-means) -Detect redundancy -Sentence ranking based on tf/idf -Avoids redundant text	intrinsic (PYRAMID)
QOS, (JIN; HUANG; ZHU, 2009), 2008	-Multi-Document -Monolingual (English)	-Specific (Opinion)	-Extractive -Sentiment Analisys -Detect redundancy -POS tagging	-Extrinsic (NIST)

LIPN, (BOSSARD; GÉNÉREUX; POIBEAU, 2008), 2008	-Multi-Document -Monolingual (English)	-Specific (Opinion)	-Extractive -Cluster-Based (MEAD) -Sentiment Analysis -Avoids redundant text	Intrinsic (PYRAMID)
RnR System, (LABBE et al., 2010), 2010	-Multi-Document -Monolingual (English)	-Specific (Opinion)	-Extractive -Sentiment Analisys -Semantic analysis -POS tagging -Graphical output	Intrinsic and Extrinsic
Query-Focused, (WANG et al., 2014), 2014	-Multi-Document -Monolingual (English)	-Specific (Question Answering)	-Extractive -Submodular Opinion Summarization based on Relevance, Coverage, Authorship Coverage, Polarity, Content Coverage Dispersion Function based on Lexical Dissimilarity, Semantic Dissimilarity, Topical Dissimilarity	Extrinsic

Open Text Summarizer, (ROTEM, 2012), 2007	-Multi-Document -Multi-lingual	-Generic	-Extractive -Open Source - Dictionary of synonyms -Cue terms -Stemming and Parsing	Extrinsic
Mead (RADEV et al., 2004), 2001	-Multi-Document -Multi-lingual	-Generic	-Extractive based on centroid score, position and overlap with the first sentence. - Avoids redundant text. -Graph-Based -Open Source	intrinsic and extrinsic

4 TELL ME WHY ARCHITECTURE

Given the related work and their limitation, we now introduce our architecture. In this Chapter we describe Tell Me Why, a new architecture to provide a better understanding of reviews. Thus, Section 4.1 provides an Architecture overview. Next, Section 4.2 presents the Data Retrieval. Section 4.3 describes the Abstraction module. Section 4.4 provides details about Macro Planning. Section 4.5 present Sentence Extraction. Finally, Section 4.6 Surface Realization shows information how handle the data into text.

4.1 Architecture Overview

We have developed an architecture to provide a helpful opinion to the users. Tell Me Why (TMW) analyze item, user and reviews dimensions to extract frequents item sets. In this context, item sets refers to the often features found into dataset dimensions. Next, it retrieves a couple of sentences to contrast these frequent itemsets. Finally, we build a short explanatory statement using Natural Language Generation. Such summary could explain why someone like or dislike an item and provide a useful opinion. Thus, we design our architecture to answer the following questions about movies:

1. Which product a group of user appreciates?
2. Who appreciates this product?
3. What a user group appreciate about a specific product?

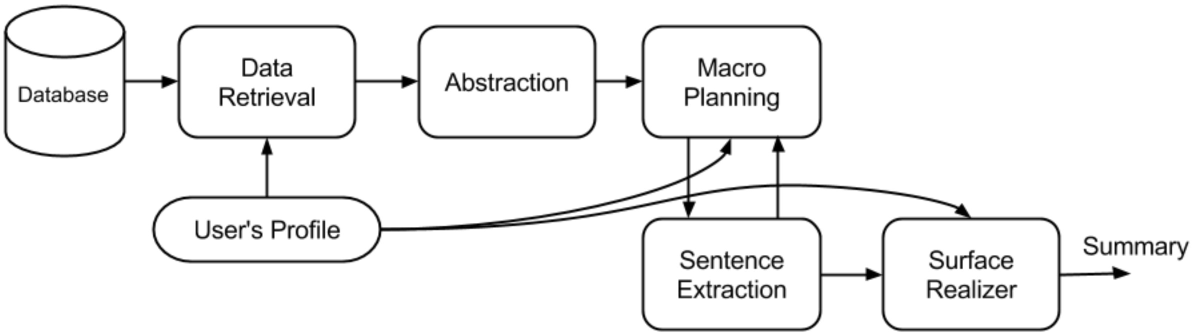
To answer these questions, the system generates two different types of summary, such as:

- **Non-personalized summary**. In this mode, the system can summarize characteristics about a product and extract sentences that explain what the general people's preference about an item.
- **Personalized summary**. In this mode, we added personalization. With this feature, we can provide a summarization for a specific people's characteristic.

Figure 4.1 presents the TMW1's architecture and its modules. The modules represents a black box responsible to handle a distinct task in the whole workflow, as follow:

In the sequence, we will explain the different modules and how they are interconnected.

Figure 4.1 – TMW1 Architecture



Source: Developed by the author

4.2 Data retrieval

This module receives a request for a summary and is responsible to perform a query on the database to create a data structure to the next modules. When it receives a request to create a summary, a particular query on the database. The related query is responsible to create the data structures in the format of the tuples $< u, i, r, w >$. Such tuples represent a text review w and rating r about an item i gave by a user u.

For instance, let us suppose that a middle-aged woman wants to get an opinion about the *Pulp Fiction* movie. In this case, the Q_1 represent the request to the TMW and the P_1 the real input of the system. The Data Retrieval module receives the parameter P_1 and perform a query on the database. It provides a set of $< user, item, rating, review >$ as output.

Q_1**: What do middle-aged females think about Pulp Fiction movie?**

$P_1 : \{user : \{gender : female, age : middle - aged\}, movie : \{title : PulpFiction\}\}$

4.3 Abstraction

The abstraction task consists of discovery pairs of values that are related each other by using of heuristics (PORTET et al., 2007). It allows rewriting these similar values into a higher-order unit. The abstraction process can be compared to an information compression problem. When applied a higher level of abstraction on the data, necessarily it will lose details about the original information. The dilemma represents the problem of choosing the right level of abstraction necessary to make possible finding similar patterns in a data set, without losing essential details. The user's age are abstracted as follows: teenager, ages under 20; Twenties, ages from 20 to 29; Thirties, ages from 30 to 39; forties, ages from 40 to 49; Quinquagenarian, ages from 50 to 59; Senior, ages over 60. The rating grades would be abstracted into labels "like" or "dislike".

As output, we have a tuple of $< user, item, rating, review >$ abstracted data used as input to the next module, called Macro Planning.

4.4 Macro Planning

Among all the attributes available in the dataset, this module decides which one is the most important to express the idea into the text. It defines what kind of information should be communicated (REITER; DALE; FENG, 2000). Thus, Association Rules Learning is used in the Macro Planning process to extract a set of frequent itemsets. These extracted rules will be used to communicate a pattern found in the dataset. Formally, A rule is defined as an implication of the form $X \implies Y$ where $X, Y \subseteq I$ and $I = \{i_1, i_2, \ldots, i_n\}$. The sets of items X and Y are called antecedent and consequent of the respective rule. For instance, from the item, user and review data we can extract the follow rule: $\{middleAged, woman\} \Rightarrow \{drama, AmericanBeauty\}$. However, this process may produce a vast number of association rules, being necessary the use of criteria to select the best one. The best rules are selected according to three criterias, as follow:

- **Support:** $\mathrm{supp}(X)$ is defined as the proportion of transactions in the dataset that contains the itemset X. When the rules are ranked by support, the most frequent patterns found in the data set are selected.
- **Confidence:** $\mathrm{conf}(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)}$ and represents the proportion of transactions that contains X and Y itemsets. When the rules are ranked by confidence, the truest patterns found in the data set are selected.
- **Number of words:** This metric is a quick way to select the best rules based on the size of its corpus. This metric allows only rules that have better chances to answer why a determined group like or dislike a movie will be chosen.

As output, each rule extracted from the data contains a list of characteristics that frequently occurs together. The output of this process is a list of frequents patterns $\{O_1, O_1, O_1 ... O_n\}$, where $O = \{atr_1 : val_1, atr_2 : val_2, \ldots, atr_n : val_n\}$.

4.5 Sentence Extraction

Sentence Extraction is responsible to extract sentences from the mass of reviews that explains *why* people like or dislike an item. The corpus of a rule is compared to the entire corpus. The sentences that have underutilized words are classified as the most distinguished.(LABBE; LABBE, 2005). Section 5 provides details about the two methods employed in this task.

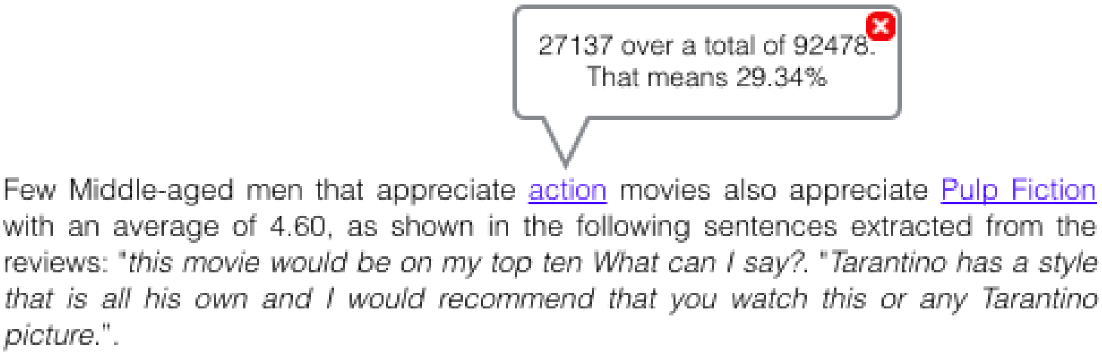
4.6 Surface Realization

This module renders the sentences and attributes as a string. The SimpleNLG (see 2.2) was utilized to put the features in a proper syntactic structure. To illustrate, the question Q_1 represents the question asked to TMW and I_1 : the related input into the Surface Realization module. Figure 4.2 show the output of Surface Realization related to the question Q_1.

Q_1: What do middle-aged females think about Pulp Fiction movie?

$$I_1 : \{user : \{gender : female, age : middleAged\}, item : \{genre : drama, title : PulpFiction\}, rating : \{avg : 4.11\}, metrics = \{support = 0.07, confidence = 0.3977\}, statistics : \{\ldots\}, sentences : \{s_1, s_2\}\}$$

Figure 4.2 – TMW Personalised Summary



27137 over a total of 92478.
That means 29.34%

Few Middle-aged men that appreciate action movies also appreciate Pulp Fiction with an average of 4.60, as shown in the following sentences extracted from the reviews: "*this movie would be on my top ten What can I say?.* "*Tarantino has a style that is all his own and I would recommend that you watch this or any Tarantino picture.*".

Source: Developed by author

5 EXPERIMENT DESIGN

As presented in Chapter 4, the Tell Me Why (TMW) architecture is composed of *Data retrieval, Abstraction, Macro Planning, Sentence Extraction and Surface Realization*. Additionally to conducting an evaluation of TMW, two methods were employed in *Sentences Extraction* module. In this Chapter, we describe the validation of TMW, as well as the experiments performed using Statistical-based and Graph-based methods in the Sentence Extraction. Thus, Section 5.1 describes the dataset utilized in our experiments. Section 5.2 presents two methods utilized into *Sentence Extraction* module. Section 5.3 presents the strategy utilized in the evaluation task. Finally, Section 5.4 describes the threats involved in the evaluation strategy.

5.1 Dataset

The essential requirement to opinion summarization is opinionated data. Thus, many researchers have created dataset by themselves by crawling review sites with specific queries (LIU; HU; CHENG, 2005; KIM; ZHAI, 2009; TITOV; MCDONALD, 2008; LU; ZHAI; SUNDARESAN, 2009). For example, IMDB for movies, Yelp for restaurants and TripAdvisor for hotels. Additionally, some standard data sets that are commonly used to evaluate the task of opinion summarization. These databases are TREC[1], NCTIR[2], and MPQA[3] that were initially designed for sentiment classification. However, considering that there is no standard data set specially designed for opinion summarization, just a few researchers have used these standard data sets.

Another way to perform an evaluation is using a gold standard as reference to performing a comparative assessment. However, to create a gold standard data is another issue in the evaluation. Most of previous opinion summarization research relies on multiple human annotations. It solves the problem of Human Variation, Granularity and Semantic Equivalence (see Section 2.4.2)(LIU; HU; CHENG, 2005; KU; LIANG; CHEN, 2006; LU; ZHAI; SUNDARESAN, 2009; POPESCU; ETZIONI, 2007; ZHUANG; JING; ZHU, 2006; GANESAN; ZHAI; HAN, 2010; STOYANOV; CARDIE, 2006). Most studies also have shown agreement rate among human assessors and ratings of reviews. Such agreement leads to the use of numeric classification as a gold standard (TITOV; MCDONALD, 2008; LU; ZHAI; SUNDARESAN,

[1] http://trec.nist.gov

[2] http://research.nii.ac.jp/ntcir/index-en.html

[3] ttp://nrrc.mitre.org/NRRC/publications.htm

2009). The output of another system also has been used in comparative assessments. The work (GANESAN; ZHAI; HAN, 2010) use MEAD [4] summarization system as gold standard.

In the movie domain, there are public databases that provide the necessary data to perform our experiments. In addition to Review data, they provide a *helpful* score that can be used as gold standard. Such related databases are:

- *Internet Movie Database (IMDb)*[5] It is website that aims to catalog every pertinent detail about a movie and provides an interface to people generate and browse reviews. For the research purpose, it provides 49 plain text files containing different characteristics about movies [6]. These files are continually updated, and until May 3th 2015 the movie list contains more than 3,200,000 titles. The database includes rich information about the movie and ranting, but neither information about user nor the text reviews. The full description of data available in this database is showed on table 5.1.
- *MovieLens Dataset.* MovieLens is a movie recommender system, developed by the Department of Computer Science and Engineering at the University of Minnesota. The MovieLens system implements a typical collaborative filtering system that collects users' preferences to suggest to user a movie. Two datasets with different sizes are available at for non-commercial, educational and research use [7]. The first one consists of 100,000 ratings for 1682 movies by 943 users. The second one consists of approximately 1 million ratings for 3883 movies by 6040 users. Unfortunately, the data set provides details only about users, ratings, and movies, but not text review. The full data description is presented on Table 5.1.
- *Stanford Web Data project* is a collection containing approximately 7 million records extracted from Amazon (MCAULEY; LESKOVEC, 2013). However, the location attribute, in the original data, comes directly from a text informed by the user, without any validation or revision. Such fact represents a significant problem. To solve this issue, use Google Maps API[8] to standardize the location attribute and get a location standard name from the text informed by the user. Table 5.1 presents the complete list of its attributes.

Considering the databases mentioned before, we have opted to perform our experiments in the movie review domain. Unfortunately, there is no unique database that provides all the necessary data. The architecture needs a database that fill the tuple <u,i,r,w> (described in the

[4] http://www.summarization.com/mead/

[5] http://imdb.com

[6] http://www.imdb.com/interfaces

[7] htp://grouplens.org/datasets/movielens/

[8] Google Maps API can be accessible at https://developers.google.com/maps/

Section 4). We thus decided create our database by building two new databases merging data from IMDb, MovieLens, and Amazon. Such databases are:

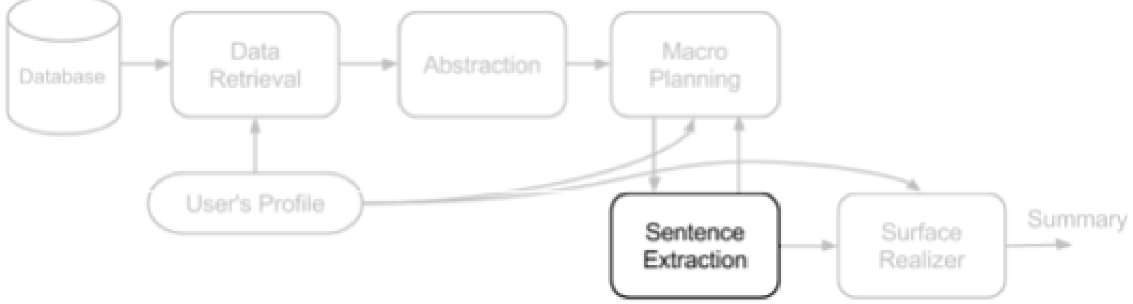
- $Dataset\ \alpha$*:* This database was created from merging Amazon, IMDb, and Amazon's databases. We have used it to create personalized summaries since it contains user's description. Note that, the criteria to get the text review from Amazon database is that the rating must be the most similar possible to each other. The complete list of attributes of the $Dataset\ \alpha$ is showed at table 5.1.
- $Dataset\ \beta$: This database was created from merging between MovieLens and IMDb's databases. We have used it to create non-personalized summaries since it does not contain user's description. It includes movie characteristics, user location, as well as the free text review. The complete list of attributes of the $Dataset\ \beta$ is showed at table 5.1.

Given the database utilized, we now introduce our experiments and the methods employed in Sentence Extraction module.

5.2 Experiments

The *Sentence Extraction* module, showed at Figure 5.1, aims to analyses and extracts relevant sentences that explain *why* some group of people like or dislike a movie. We have employed two different methods on this module. Thus, Section 5.2.1 present the input and output data of Sentence Extraction module. Section 5.2.2 describes the Statistical-based method for sentence extraction. Section 5.2.3 presents Graph-based method for sentence extraction.

Figure 5.1 – Sentence Extraction module



Source: Developed by the author

5.2.1 Input and Output of Sentence Extraction module

Macro Planning extracts a set of Association Rule (ARs) of the review data (see Section 4). The Macro Planning's output is used as input to the *Sentence Extraction*. It is composed by a set of features F, a set of related sentences S and the number of sentences to be extracted N. The output of the *Sentence Extraction* module is a set of sentences S, and their score R.

5.2.2 Experiment 1: Statistical-based method for Sentence Extraction

This Section describes the statistical-based method, called as TMW1. This method uses *Hypergeometric Law* to identify the most distinguished sentences based on the likely of overutilized and underutilized words (LABBE; LABBE, 2005) (Section 2.3.1). Thus, we use the vocabulary U related to the input features F. Next, U is compared to the entire corpus of review C to highlight the words that are over utilized and underutilized. Equation 5.1 given Hypergeometric law used to compute this probability.

$$P(X = F_{iu}) = \frac{\binom{F_{ic}}{F_{iu}}\binom{N_c - F_{ic}}{N_u - F_{iu}}}{\binom{N_c}{N_u}} \tag{5.1}$$

Where N_c is the number of token in the corpus C, N_u the number of token of U. The F_{iu} and F_{ic} are the absolute frequency of word i in U and C respectively. When the probability is too high compared to its expectation ($E = F_{ic}\frac{N_u}{N_C}$), then the word is over-used in the sub-corpus. Thus, the sentences are ranked according to the proportion of underutilized in the corpus and overutilized in the sub-corpus (LABBE; LABBE, 2005).

5.2.3 Experiment 2: Graph-based method for Sentence Extraction

Based on results obtained in TMW1, were identified deficiencies and limitations that can hinder the understanding of the summary by the user. Thus, some modifications were proposed to handle with this cited, such as: (i) considering the size of the sentences to be included in the review; (ii) regarding the semantic similarity between words of sentences and the features found in the association rules; (iii) considering its centrality in the corpus (ERKAN; RADEV, 2004).

We propose the function-based method (WANG et al., 2014), to compute sentence relevance, called TMW2. Each sentence $s \subset r \in R$, where r is an individual text review wrote by user subset of the mass of reviews R, are evaluated individually. The computation of the scores of the sentences is performed by the weighted sum of a set of functions called sub-functions. Each sub-function aims to evaluate a different aspect of the sentence to provide a relevance score. The function-based developed in the TMW2 is based on the follow functions: (i) Sentence salience (ii) Sentence length; and (iii) Semantic Similarity. The function is give in Equation 5.3:

$$F(s) = \alpha c(s) + \beta l(s) + \omega w(s) \tag{5.2}$$

Where $c(S)$, $l(S)$, $w(S)$, are *Sentence Salience function*, *Lenght function* and *Semantic Similarity function* respectively. The coefficients α, β, ω are 0.5, 0.25, 0.25 for α, β , ω, respectively, as tuned on the development set.

$c(s)$ is stochastic graph-based method for computing the sentence relevance (ERKAN; RADEV, 2004). It uses the concept of sentence salience to identify the most important sentences in a set of documents (see section 2.3.1). We have utilized LexRank (ERKAN; RADEV, 2004) algorithm for computing sentence relevance based on the centrality of a graph representation.

$l(s)$ is a binary function, resulting 1 to sentences that are in the range of 5 to 50 words. Small sentences cannot entirely justify an opinion. On the other hand, large sentences use many words to express an opinion.

$w(s)$ computes the Semantic similarity between the features found in Association Rule to the words of the sentence (see Section 2.3.1). We have used *WordNet Similarity for Java (WS4J) API*[9] to compute the Semantic similarity between features and words. For instance, the normalized score between the features *middleAged, Woman, drama, American Beauty* and the sentence "*My wife loves this movie*" is 0.6334.

5.3 Evaluation Strategy

We have performed two types of evaluation. First, a comparative assessment, where human judges the goodness and utility of the TMW summary by a comparison with a human-authored gold standard. We used Amazon Most Helpful Review (AMHR) as the gold standard. The Amazon site allows people to say how much helpful is a review. The most voted is called

[9] Available at HTTP://code.google.com/p/ws4j/

The Most Helpful Review. Second, we performed an intrinsic evaluation using ROUGE to measure the content selection quality. Thus, Section 5.3.1 describes the comparative assessment. Section 5.3.2 describes the intrinsic evaluation.

5.3.1 Comparative Evaluation

To evaluate the TMW, a population was recruited to compare TMW with AMHR. In addition to the task of deciding which reviews are most helpful, the population also was invited to answer questions about the summary utility, as follow:

1. What do you prefer: AMHR or TMW summary?
2. What do you prefer: AMHR or TMW personalized summary?
3. Does the explanation is helpful?
4. Is it easy to understand?
5. Would you use this system to get an opinion?

To better understanding of the answers given by population, we asked if they usually read movie reviews. We use this information to categorize people into three classes, as follow: (i) who always read reviews to get an opinion; (ii) who sometimes read review; and (iii) who never use read reviews to get opinion; This segmentation provide us a better understanding of the people preference in each class. We can assume that people who always read reviews to get an opinion is an expert. In other hand, people who sometimes and never use to read reviews also provide us valuable feedback about the summarization process.

5.3.2 Intrinsic Evaluation

ROUGE-N is an *n-gram* recall between a candidate summary and a set of reference summaries. We use ROUGE (LIN, 2004) to quantitatively assess the agreement of TMW summaries with a human review. ROUGE is based on an *n-gram co-occurrence* between gold standard and system summary. It is a widely accepted standard for evaluation of summarization tasks (GANESAN; ZHAI; HAN, 2010). ROUGE-1 and ROUGE-2 have been shown to have the most correlation with human summaries (LIN; HOVY, 2003) and higher order ROUGE-N scores (N>1) estimate the fluency of summaries. We use five human reviews as reference in our evaluation since it can achieve better correlation with human judgment (LIN, 2004). ROUGE-N is computed as follow:

$$ROUGE-N=\frac{\sum\limits_{S\in\{ReferenceSummary\}}\sum\limits_{gram_n\in S}count_{match}(gram_n)}{\sum\limits_{S\in\{ReferenceSummary\}}\sum\limits_{gram_n\in S}count(gram_n)} \tag{5.3}$$

Where n stands for the length of the *n-gram, gramn*, and *Countmatch(gramn)* is the maximum number of *n-grams* co-occurring in a candidate summary and a set of reference summaries. We have chosen MEAD (RADEV et al., 2004) as baseline. MEAD is the state of the art extractive summarizer based on clustering. It aims to detect the most important words from the whole cluster of sentences to select the best one to be included in the final summary. The sentence selection is based on three parameters: *minimum sentence length*, *centroid*, and *position* in the text. The minimum sentence length by default is nine words. It means that sentences shorter than nine words is cutted off. The parameter position is related to the position of the sentence in the document. In Our experiments, we turn this parameter off since it's not considered in our methods. MEAD is the ideal baseline because a good summary in our case would be one that capture the essential information. It is what centroid-based summarize aims to produce.

5.4 Threats to Validity

This section presents the threats to validity our evaluation and the actions that were utilized to handle with those risks.

Movie bias

To avoid bias, the movies are selected according to two criteria. First, the selected films should cover as many different genres as possible. Second, the selected films should be familiar to most of population. According to criteria, we have selected eight movies from the top 250 list of IMDB. The selected movies are *Stand by Me, Pulp Fiction, The Good the bad and the ugly, Forrest Gump, The Silence of the Lamb, Forrest Gump, Schindler's List, American beauty*. For each movie, we have used 200 reviews. There are more than 25,000 sentences and more than 484,000 words in all the selected reviews.

Criteria of choice of gold standard review

To perform our evaluation is necessary select some reviews written by human as a gold-standard. To avoid an inadequate assessment of our system, we have used the AMHR as the gold standard.

Mobile devices

The results obtained in the TMW1 lead us to conclude that review summarization could be more useful in mobile devices. The size of the screen even the transmission rate can hamper users to access and read lengthy reviews. To validate this hypothesis, we conducted a brief survey to collect people opinion about read long text in devices mobile. We recruited a population composed of 21 men and women that self-declared as a user of mobile devices. The results show that 76% of the people said that felt uncomfortable to read a lengthy review on a mobile devices. Based on the results obtained, we conduct the TMW2 experiments on mobile devices.

Human choice between multiple options

It seems intuitively that the task of choosing between more than two options can be difficult for humans. Thus, in the comparative evaluation, we only show AMHR and TMW to the population.

Blind comparison between AMHR and TMW

It seems intuitively that any association with the Amazon trademark can influence the answers given by the user. Thus, we performed a blind comparison between TMW and AMHR. We do not provide any information about where come each review. It ensures a fair comparison.

Table 5.1 – DataSets Summary

Fild Name	*Description*	*IMDb*	*MovieLens*	*Amazon*	α	β
Id	Internal identifier code of the record.	✓	✓	✓	✓	✓
Title	Title of the movie.	✓	✓	✓	✓	✓
Genre	Genre of the movie rated, such as Drama or Action.	✓	✓		✓	✓
Release Date	The release date of the movie rated	✓			✓	✓
Actors	The names of the actors that starred the movie	✓			✓	✓
Directors	The name(s) of the movie director(s)	✓			✓	✓
Writers	The name(s) of the movie writer(s)	✓			✓	✓
Runtime	The length of the movie	✓			✓	✓
Rating	Numeric rating of the review	✓			✓	✓
Timestamp	The date of the review.	✓	✓		✓	✓
Actresses	Actresses of movies	✓			✓	✓
ReleaseDates	Dates of movie releases	✓	✓		✓	✓
Language	Languages of movies	✓			✓	✓
Country	Countries of movie	✓			✓	✓
Producers	Producers of movies	✓			✓	✓
UserId	Internal identifier code of the user		✓		✓	✓
Age	Age of the user.		✓		✓	
Gender	The gender of the user, such as Male or Female.	✓	✓			
Occupation	The User occupation.		✓		✓	
ZipCode	User's postal code.		✓		✓	
Location	User's location.			✓		✓
Review	The User's comments about the movie.			✓		✓
Helpfulness	A numeric rating giving by the users to mensure the Helpfulness of the review.			✓		✓

Source: Developed by the author

6 RESULTS

Given details of how the experiments were performed, in this Chapter we provide the results and discussion of the evaluation. Thus, Section 6.1 presents the results obtained using Statistical-based method for sentence extractions. Section 6.2 results using Graph-based method for sentence extractions.

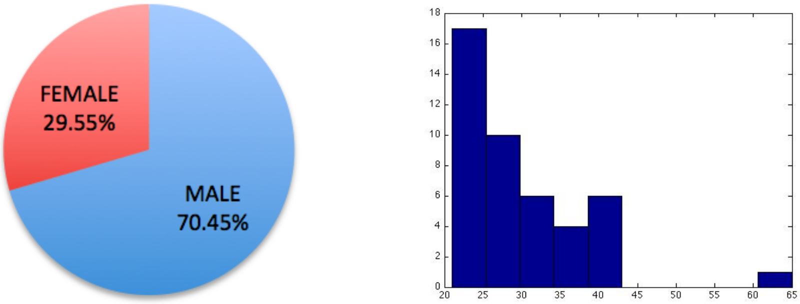
6.1 Experiment 1: Statistical-based method for sentence extraction

This section describes the results obtained from the Statistical-based method for sentence extraction. Thus, Section 6.1.1 describes comparative evaluation. Section 6.1.2 provides results of Intrinsic Evaluation. Finally, Section 6.1.3 provides discussion about the results obtained.

6.1.1 Comparative Evaluation

To perform a comparative evaluate, we have recruited a population composed of 44 men and women between 20 and 45 years of age. The figure 6.17 reveals detail about the age and the gender of the population.

Figure 6.1 – Gender distribution and Age histogram of the population.



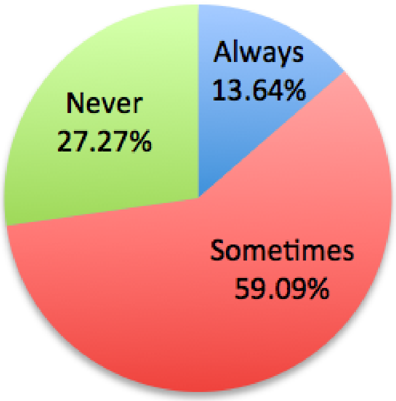
Source: Developed by the author

The populations were confronted with two options to choose the most helpful summary. One being the AMHR and other from TMW1 (The complete form and prototype used to this evaluation is available in Appendix B). Additionally, they were invited to answer a couple of questions, as follow:

1. **Do you usually browse reviews before watching a movie?**

   To better understanding of the answers given by population, we asked if the population read reviews before watching a movie. It provides a valuable information to carry out an analysis of the population's answers.

Figure 6.2 – Who reads reviews before watching a movie?

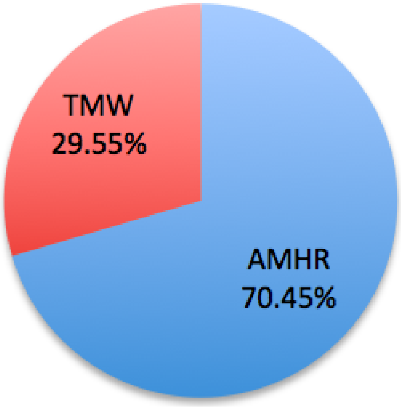


Source: Developed by the author

2. **What do you prefer: AMHR or TMW non-personalized summary?**

   As described in Section 4, TMW1 provides non- and personalized summaries. The TMW1 non-personalized summary just contains information about items and review's dimensions. The population were confronted to choose the most helpful summary between AMHR and TMW. The "Pan's Labyrinth" movie was utilized because the vast amount of reviews available. The summaries used to this evaluation is available on Appendix B. Figure 6.3 show that almost all the population think that AMHR is the most helpful. Additionally, Table 6.1 show that all the people that always read reviews chosen the AMHR as the most useful. Just a few people that sometimes and never use to read reviews accepted TMW1 as most helpful. A full discussion about results is provided on Section 6.1.3.

Figure 6.3 – Comparative evaluation between AMHR and TMW non-personalized summary



Source: Developed by the author

Table 6.1 – Comparative evaluation between AMHR and TMW non-personalized segmented by: a) People who always read review; b) People who sometimes read review; and c) People who never read review;

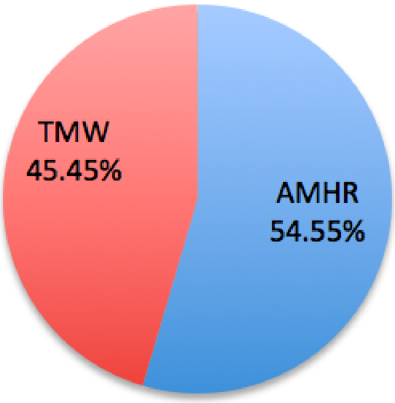
	a) Always	b) Sometimes	c) Never
TMW1	0%	34.62%	33.33%
AMHR	**100%**	**65.38%**	**66.67%**

Source: Developed by the author

3. **What do you prefer: AMHR or TMW personalized summary?**

   TMW1 personalized summary offer a summary that match a user's profile. To the task of evaluate TMW personalized summary, we asked people to assume that they are looking for a movie to a middle-aged female friend. They choose the most helpful summary between AMHR and TMW. Figure 6.4 shows that almost half of the population has chosen TMW's personalized summary as the most helpful review. Table 6.2 shows that almost all the people who always read reviews have preferred AMHR. In order hand, half of people who sometimes and never read reviews have preferred TMW1.

Figure 6.4 – Comparative evaluation between AMHR and TMW personalized summary



Source: Developed by the author

Table 6.2 – Comparative evaluation between AMHR and TMW non-personalized segmented by: a) People who always read review; b) People who sometimes read review; and c) People who never read review;

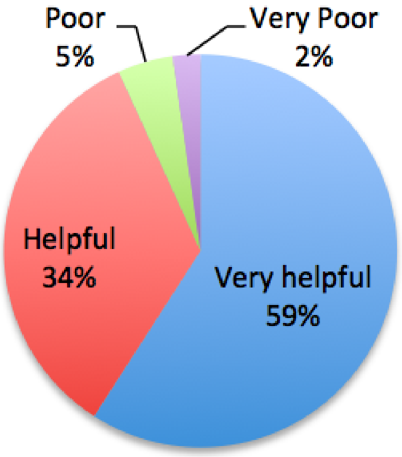
	a) Always	b) Sometimes	c) Never
TMW1	16.67%	**50%**	**50%**
AMHR	**83.33%**	**50%**	**50%**

Source: Developed by the author

4. **Does TMW1 summary is helpful?**

   To measure the quality of the TMW, we ask us for people to evaluate the usefulness of TMW summary. Figure 6.5 shows that overall people said that TMW1 is very helpful. Table 6.3 shows that most of the people that says that TMW1 is helpful belongs to the class of always read reviews. It means the same class of people who says that AMHR is better then TMW1 said that TMW1 was helpful. Section 6.1.3 provides a discussion about this fact.

Figure 6.5 – TMW personalized summary helpfulness



Source: Developed by the author

Table 6.3 – Comparative evaluation between AMHR and TMW non-personalized segmented by: a) People who always read review; b) People who sometimes read review; and c) People who never read review;

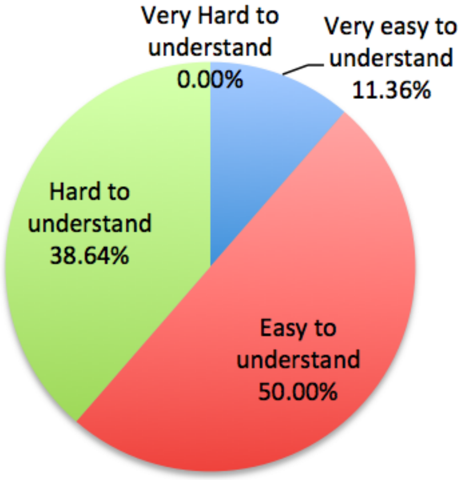
	a) Always	b) Sometimes	c) Never
Very Helpful	0%	3.85%	0%
Helpful	**66.67%**	**57.69%**	**50%**
Poor	33.33%	38.46%	33.33%
Very Poor	0%	0%	16.67%

Source: Developed by the author

5. **Is it easy to understand?**

   Assuming that a summary should be clear, we want to know how easy TMW1 is to understand. Figure 6.6 shows that most of the people think that the summary is easy and very easy to comprehend. Table 6.4 shows that all the classes of people has the same preference about understanding of TMW1 summary.

Figure 6.6 – Understanding of the TMW summary



Source: Developed by the author

Table 6.4 – Comparative evaluation between AMHR and TMW non-personalized segmented by: a) People who always read review; b) People who sometimes read review; and c) People who never read review;

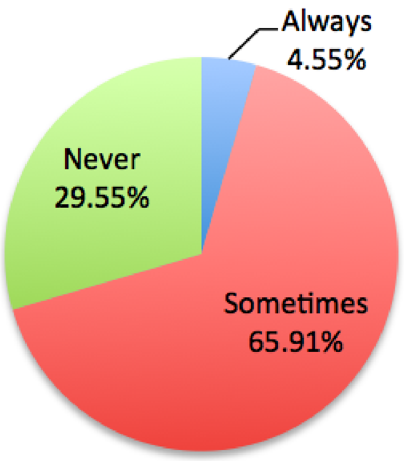
	a) Always	b) Sometimes	c) Never
Very Easy to understand	0%	11.54%	16.67%
Easy to understand	**50%**	**50%**	**50%**
Hard to understand	**50%**	38.46%	33.33%
Very Hard to understand	0%	0%	0%

Source: Developed by the author

6. **Who would like to use TMW to get an opinion**

   Finally, we asked who would like to use TMW to get an opinion. Figure 6.7 shows that almost all the people would use this system to get an opinion. Table 6.5 shows that specialist would use sometimes TMW1 to get an opinion.

Figure 6.7 – Who would use TMW



Source: Developed by the author

Table 6.5 – Comparative evaluation between AMHR and TMW non-personalized segmented by: a) People who always read review; b) People who sometimes read review; and c) People who never read review;

	a) Always	b) Sometimes	c) Never
Always	0%	3.85%	8.33%
Sometimes	**66.67%**	**69.23%**	**58.33%**
Never	33.33%	26.92%	33.33%

Source: Developed by the author

6.1.2 Intrinsic evaluation

As presented in Section 5, we use ROUGE-N to assess quantitatively the agreement of TMW summaries (LIN, 2004). ROUGE is based on an *n-gram co-occurrence* between gold standard summaries and system summaries. It is a widely accepted standard for evaluation of summarization tasks (GANESAN; ZHAI; HAN, 2010). To compare the results, we use MEAD as baseline. Table 6.6 shows the performance comparison between TMW statistical-based method and baseline. This comparative have shown that the baseline method has better results in the recall, precision, and f-measure scores. We have observed that the sentences chosen by baseline has 35 words on average against 16 of TMW1. Such characteristic could improve the baseline recall. Unfortunately, we cannot change the max length of the sentences to provide a fair comparative.

Table 6.6 – Performance comparison between TMW1 and Baseline.

Recall

	Rouge-1	Rouge-2	Avg Words
TMW1	0.03737	0.00719	16
Baseline	**0.07318**	**0.01500**	35

Precision

	Rouge-1	Rouge-2	Avg Words
TMW1	0.50000	0.08276	16
Baseline	**0.55429**	**0.23284**	35

F-measure

	Rouge-1	Rouge-2	Avg Words
TMW1	0.06954	0.01323	16
Baseline	**0.12929**	**0.02819**	35

Source: Developed by the author

6.1.3 Discussion

To evaluate the TMW1 were recruited a population composed of 44 men and women between 20 and 45 years of age to compare TMW with the AMHR. To evaluate non-personalized summarization, we chose "Pan's Labyrinth" movie because the vast amount of available reviews. The results showed that 70.45% of the people preferred AMHR instead TMW. The results showed that all the specialist selected AMHR as the most helpful. Such fact could say that TMW1 non-personalized summary has no advantage comparing with AMHR. There are two possible explanation of this negative result. First, the reviews used to generated the summary provide few information about the movie. Second, the method employed in the Sentence Extraction could not be the most suitable to capture the most relevant sentences in review summarization. To better understand this negative result, we performed an intrinsic evaluation in the sentence extraction. We used ROUGE-N to compare our results with MEAD. The results showed that Hypergeometric Law have no advantages comparing with MEAD. Both recall and support had lower precision and recall comparing with the baseline. It can explain the negative results.

To evaluate TMW1 personalized summary, the people were asked to assume they are looking for a movie to a middle-aged female friend. We presented AMHR and a personalized summary about "American Beauty" movie. Additionally to choosing the most helpful, they were inquired about the understanding, helpfulness and if they would use TMW1 in the future to get opinion. The results have shown that 45.45% of the recruited population has chosen TMW1 as the most helpful review against 54.55% of AMHR. Additionally, 50% of all people considered TMW1 easy to understand and 65.91% would use TMW1 to get an opinion.

TMW1 personalized summary achieved better results than TMW1 non-personalized summary. However, TMW1 is not even better than AMHR. A possible cause for this negative result is the fact that the sentence extraction does not select the most informative sentences. We also realize that TMW could be more useful in mobile devices. The size of the screen even the transmission rate can hamper users to access and read lengthy reviews. To validate this hypothesis, we propose conduct a survey to collect people opinion about read long text in devices mobile. The evaluation also have shown that the current limitations of the TMW1 are present on the quotations extraction process. The extracted sentences are not directly related to the features found in the association rule, and this may not explain why some people like or dislike a movie. However, the whole process of getting the information, abstract them, discover frequent itemsets, sentence extraction, and surface realization are functional. Most of the people also agree that the TMW can be helpful, easy to understand and would like to use this system to get an opinion.

As future works, we highlight the improvements in the quotation extraction. Considering the limitation of the TMW previously cited, three primary improvements were identified on the *Sentence Extraction* module: (i) considering the size of the sentences to be included in the review; (ii) regarding the semantic similarity between words of sentences and the features found into the itemsets. (iii) considering its centrality in the corpus (ERKAN; RADEV, 2004).

6.2 Experiment 2: Graph-based method for sentence extraction

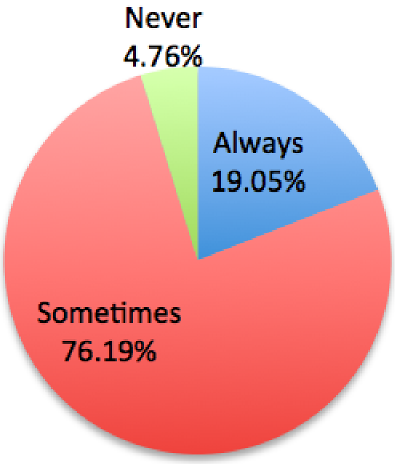
This section describes the results obtained from the Graph-based method for sentence extraction employed by TMW2. Thus, Section 6.2.1 performs comparative evaluation between AMHR and TMW2. Section 6.2.2 provides results of intrinsic evaluation of Graph-based method for sentence extraction. Finally, Section 6.2.3 presents discussion about results.

6.2.1 Comparative evaluation

To evaluate the TMW2 were recruited a population composed of 21 men and women between ages of 20 to 35 years to perform blind comparison evaluations between TMW and the AMHR (The complete form used to this assessment is available in Appendix B). In such configuration, the population was invited to answer the follow questions:

1. **Do you browse reviews before watching a movie?**
   To better understanding of the answers given by population, we asked if the population read reviews before watching a movie. It provides us a valuable information to carry out an analysis of the population's answers. Figure 6.8 provides the results.

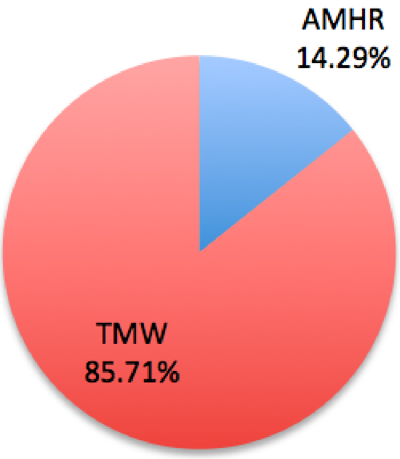
Figure 6.8 – Who reads reviews before watching a movie?



Source: Developed by the author

2. **What do you prefer, AMHR or TMW2?**
   We asked people to choose a best review between AMHR and TMW2. Figure 6.9 shows that almost all the population have preferred our approach then AMHR (The summary and the questions are provided on Appendix B). Table 6.7 shows that even the people who usually read reviews have preferred TMW2.

Figure 6.9 – People's preference between AMHR and TMW2



Source: Developed by the the author

Table 6.7 – Comparative evaluation between AMHR and TMW non-personalized segmented by: a) People who always read review; b) People who sometimes read review; and c) People who never read review;

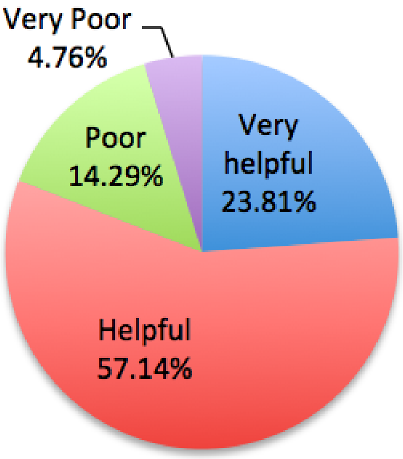
	a) Always	b) Sometimes	c) Never
TMW1	**100%**	**75%**	**100%**
AMHR	0%	25%	0%

Source: Developed by the author

3. **Does TMW2 summary is helpful?**

   We have asked for people how useful is TMW2 summary. Figure 6.10 shows that more than half of people said that TMW2 is very helpful and helpful. Table 6.8 shows that the specialist also agree that TMW2 is helpful.

Figure 6.10 – TMW2 summary helpfulness



Source: Developed by the author

Table 6.8 – Comparative evaluation between AMHR and TMW non-personalized segmented by: a) People who always read review; b) People who sometimes read review; and c) People who never read review;

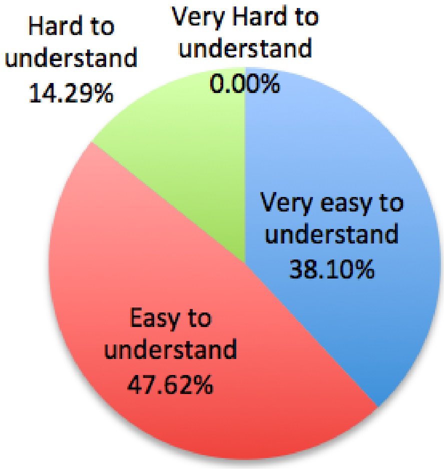
	a) Always	b) Sometimes	c) Never
Very Helpful	0%	25%	50%
Helpful	**66.67%**	**56.25%**	**50%**
Poor	33.33%	12.50%	0%
Very Poor	0%	6.25%	0%

Source: Developed by the author

4. **Does it easy to understand?**

   We want to assess how TMW2 is easy to comprehend. Figure 6.11 shows that almost all the people considered TMW2 summary *Very easy* and *Easy* to Understand. Table 6.9 shows that the majority of people who always read reviews before watching a movie said that it is Very easy to understand.

Figure 6.11 – Understanding of the TMW2 summary



Source: Developed by the author

Table 6.9 – Comparative evaluation between AMHR and TMW non-personalized segmented by: a) People who always read review; b) People who sometimes read review; and c) People who never read review;

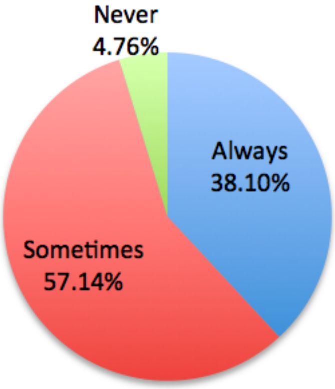
	a) Always	b) Sometimes	c) Never
Very Easy to understand	66.67%	35.71%	25%
Easy to understand	33.33%	50%	50%
Hard to understand	0%	14.29%	25%
Very Hard to understand	0%	0	0%

Source: Developed by the author

5. **Would you use this system to get an opinion?**

   We would like to know who want to use this system to obtain an opinion before seeing a movie. Figure 6.12 shows that more than half of people have said that they would sometimes use TMW2 to get a further opinion about a movie. Just a few people have said that won't use the TMW2. Additionally, Table 6.10 shows that all the experts agree to use TMW2.

Figure 6.12 – Who would like to use TMW2 to get an opinion.



Source: Developed by the author

Table 6.10 – Comparative evaluation between AMHR and TMW non-personalized segmented by: a) People who always read review; b) People who sometimes read review; and c) People who never read review;

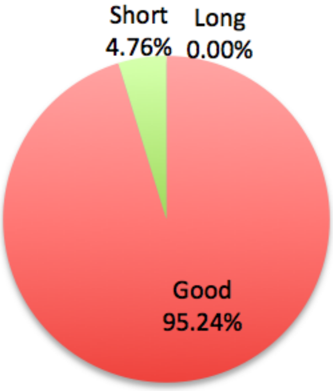
	a) Always	b) Sometimes	c) Never
Always	0%	**53.85%**	25%
Sometimes	**100%**	46.15%	**50%**
Never	0%	0%	25%

Source: Developed by the author

6. **Does the sentence length is enough to get an opinion?**

   TMW2 summary expose a couple sentence from the amount of reviews. We want to understand the people opinion about the sentence length. Figure 6.13 have shown that almost all the people considered the sentences length as "Good". Additionally, Table 6.11 show that the experts also considered length sufficient to get opinion.

Figure 6.13 – People's opinion about the length of the sentences



Source: Developed by the author

Table 6.11 – Comparative evaluation between AMHR and TMW non-personalized segmented by: a) People who always read review; b) People who sometimes read review; and c) People who never read review;

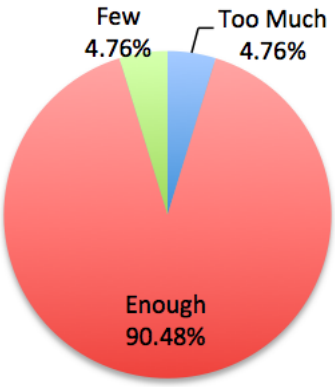
	a) Always	b) Sometimes	c) Never
Long	0%	0%	0%
Good	**100%**	**92.86%**	**100%**
Short	0%	7.14%	0%

Source: Developed by the author

7. **Does the number of sentences is sufficient to get an opinion?**

   The TMW2 extracts five sentences to each summary. Figure 6.14 shows that the almost all the people think that the number of the sentence are enough to get an opinion. Table 6.12 shows that the expert also agree that the number of sentences is sufficient to get an opinion.

Figure 6.14 – People's opinion about the number of sentences



Source: Developed by the author

Table 6.12 – Comparative evaluation between AMHR and TMW non-personalized segmented by: a) People who always read review; b) People who sometimes read review; and c) People who never read review;

	a) Always	b) Sometimes	c) Never
Too much	0%	7.14%	0%
Enough	**75%**	**92.86%**	**100%**
Few	25%	0%	0%

Source: Developed by the author

6.2.2 Intrinsic Evaluation

For intrinsic evaluation of the TMW2, we also have chosen the MEAD as baseline. By default, MEAD has three parameters: mining sentence length, centroid and position in the text. We have utilized with default parameters (Centroid, and Position are both 1. The default Length cutoff is 9). Table 6.14 shows the performance comparison between TMW2 and baseline. First, we can see that TMW2 has better results using ROUGE-2. This show that TMW's graph-based approach has good performance in the task of capture the essential information of the original text. Second, the recall scores of TMW summary using ROUGE-1 are slightly lower than baseline while the precision scores are higher. It also shows the effectiveness of the graph method into the summarization task.

Table 6.13 – Performance comparison between TMW2 and baseline.

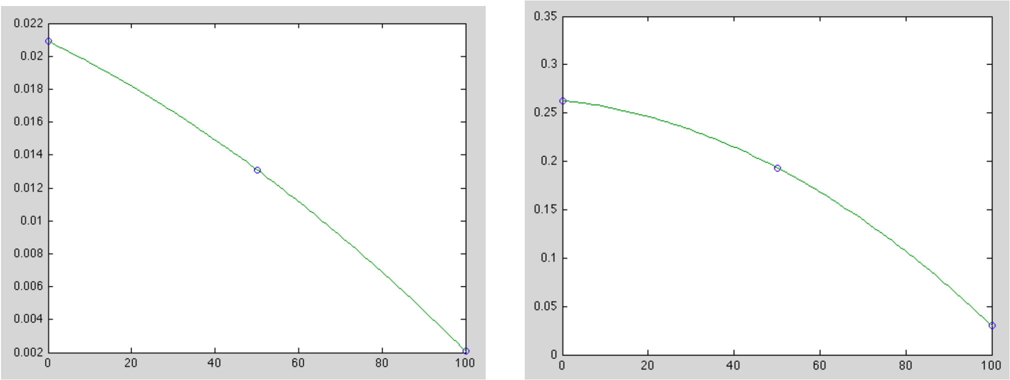
Recall			
	Rouge-1	Rouge-2	Avg Words
TMW2	0.05614	**0.02090**	15
Baseline	**0.07318**	0.01500	35
Precision			
	Rouge-1	Rouge-2	Avg Words
TMW2	**0.64865**	**0.26286**	15
Baseline	0.55429	0.23284	35
	Rouge-1	Rouge-2	Avg Words
F-measure			
TMW2	0.10334	**0.03872**	15
Baseline	**0.12929**	0.02819	35

Source: Developed by the author

The TMW2 is based on submodular function-based is given by $F(S) = \alpha c(S) + \beta l(S) + \omega w(S)$. Where $c(S)$, $l(S)$, $w(S)$, are *Sentence Salience function (Graph)*, *Length function* and *Semantic Similarity function* respectively and α, β and ω are the weights. Additionally to intrinsic evaluation, we performed experiments with different values of β and ω to better understanding of system results.

The semantic similarity function gives the score to the sentences based on the similarity between features found into itemset and words the words of the sentence. Figure 6.15 shows the Recall and Precision scores of ω weight from 0% to 100%. When ω increases, the recall and precision tend to decrease. It because the submodular function gives better scores to semantic related sentences.

Figure 6.15 – Precision and Recall



Source: Developed by the author

6.2.3 Discussion

Based on the prior results obtained in the experiments realized under TMW1, we propose a Graph-based method to sentence extraction, called TMW2. It uses sub-modular functions to compute the relevance of each sentence (WANG et al., 2014). We propose a combination of Length function, Semantic similarity function, and a Graph function. The results obtained on TMW1 also leads us to deduct that the TMW could be more useful in mobile devices. The size of the screen even the transmission rate can hamper users to access and read lengthy reviews. We conducted a brief interview to collect people opinion about read long text in devices mobile. The results show that 76% of the people said that felt uncomfortable to read long text on mobile devices and preferred short version instead. Based on it, we conduct the experiments on mobile devices context.

To perform a comparative evaluation, were recruited a population composed of 21 men and women between 20 and 35 years of age to evaluate a comparison between TMW2 and AMHR. As the result, most of the people chose TMW2 as the most helpful review and *Easy* to understand. They also said that the length and number of sentences are enough to get an opinion and they would use TMW2 to get an opinion.

Additionally to a comparative evaluation, an intrinsic evaluation was performed using

ROUGE-N. It shows that graph-based approach has good recall and precision in the task of select the most relevant sentences. We also use different values of β and ω to understand how it can impact on recall and precision (see Equation 5.2). It shows that when the semantic function receives a higher weight the recall and precision decrease.

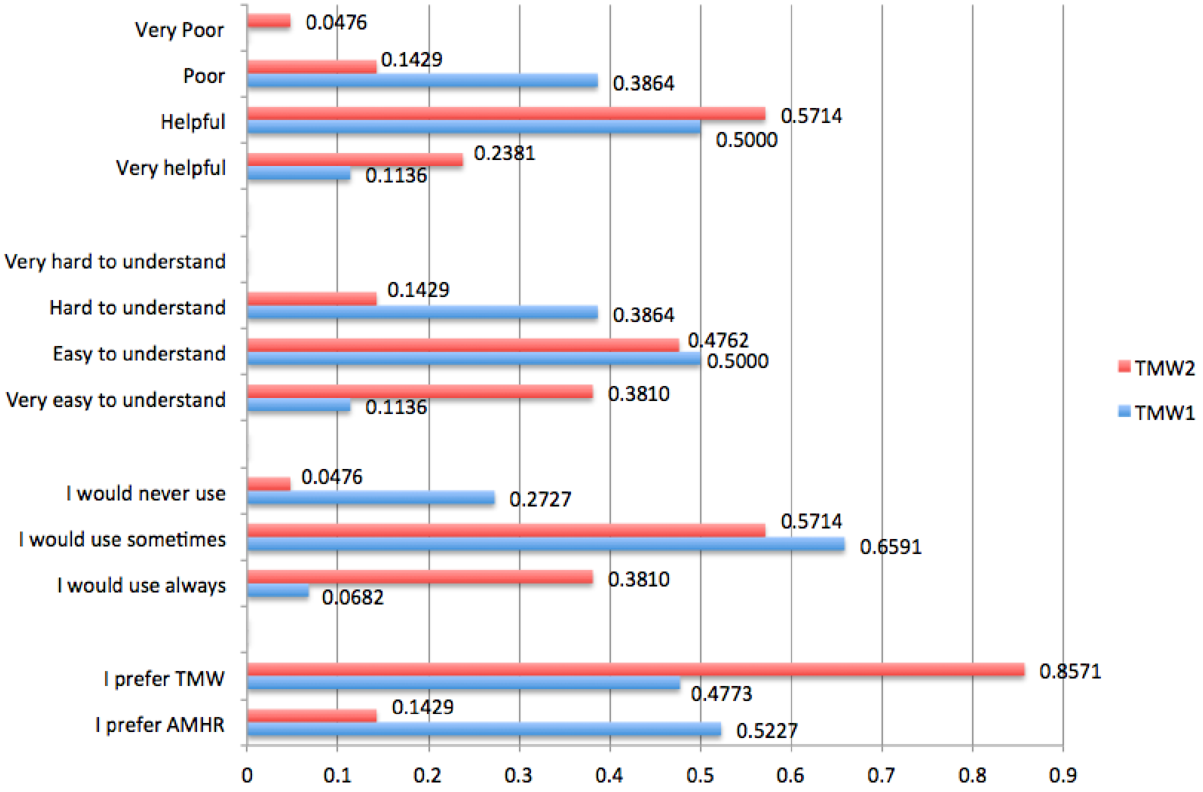
The results prior cited, show that the whole process of getting the information, abstract them, find patterns, execute quotation extraction and surface realization are functional and satisfactory. Most of the people agree that the TMW2 can be helpful, easy to understand, and they would use TMW2 to get an opinion. The TMW2 also provide good precision and recall in the sentence extraction.

6.3 Comparison between TMW1 and TMW2

Given the result of comparative evaluation of TMW1 and TMW2, this section provides a comparison between them. Thus, Figure 6.16 exposes their results obtained in the comparative assessment. The red bars represents the TMW2, and the blue one is the TMW1. The horizontal group of the axis represents four questions: 1) Does the explanation is helpful? 2) Is it easy to understand? 3) Would you use this system to get an opinion? And 4) What do you prefer: AMHR or TMW summary?

Some observation must be made: a) the evaluation were performed individually. TMW1 was not confronted directly against TMW2; b) The people that evaluate TMW1 were not the same that Evaluate TMW2; c) TMW1 evaluations were performed on the computers, while TMW2 evaluation were conducted on mobile device; Such observations prevent us from making a direct comparison between TMW1 and TMW2. However, we can we can observe that TMW2 obtained a slight advantage in the opinion of users.

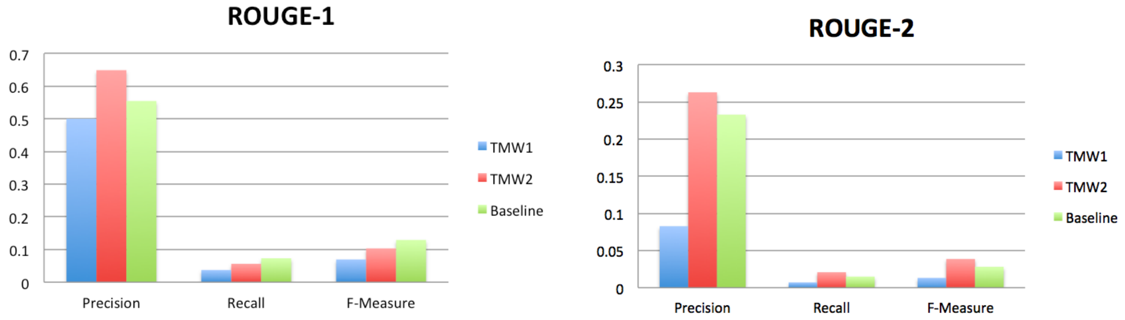
Figure 6.16 – Comparative evaluation of TMW1 and TMW2



Source: Developed by the author

Additionally, we perform a comparison of intrinsic results. Table 6.14 shows that TMW2 is better than TMW1 in all the measures. It shows that Graph-based method has better results than Hypergeometric, using ROUGE measure. Figure 6.17 provides visual interpretation of the results.

Figure 6.17 – Intrinsic evaluation between TMW1, TMW2, and baseline.



Source: Developed by the author

Table 6.14 – Performance comparison between TMW1 and baseline.

Recall			
	Rouge-1	Rouge-2	Avg Words
TMW1	0.03737	0.00719	16
TMW2	0.05614	**0.02090**	15
Baseline	**0.07318**	0.01500	35
Precision			
	Rouge-1	Rouge-2	Avg Words
TMW1	0.50000	0.08276	16
TMW2	**0.64865**	**0.26286**	15
Baseline	0.55429	0.23284	35
	Rouge-1	Rouge-2	Avg Words
F-measure			
TMW1	0.06954	0.01323	16
TMW2	0.10334	**0.03872**	15
Baseline	**0.12929**	0.02819	35

Source: Developed by author

7 CONCLUSION AND FUTURE WORK

Users increasingly rely on collaborative reviewing sites to achieve tasks such as purchasing a product or renting a movie. Over the time, these sites have built a rich database containing user demographic dimensions such as location, age, gender, item dimensions such as director and actors for a movie, ratings, and reviews. But this vast amount of data can hamper the user to get an unbiased opinion. The review also plays an important part in the purchasing decisions of the consumers(CHATTERJEE, 2001; CHEVALIER; MAYZLIN, 2006; KIM; SRIVASTAVA, 2007). However, to get a suitable opinion from this mass of data, the users have to deal with a vast amount of textual information. Precisely, they have to deal with the problem of the vast amount of divergent reviews about the item and its features, to extract a suitable opinion from them. Unfortunately, the ability to explain user reviews is an under-explored research question. It requires combining sophisticated text analysis with the exploration of an exponential search space that results from crossing reviewer and item dimensions.

Our work proposes a new architecture, called Tell Me Why. It is a project developed at Grenoble Informatics Laboratory in cooperation with the Federal University of Rio Grande do Sul to provide users a better understanding of reviews. This architecture provides a combination of text analysis from reviews with mining structured data resulting from crossing reviewer and item dimensions. Additionally, this work performs an investigation of summarization methods utilized Sentence Extraction module.

We have employed Hypergeometric Law on the Sentence Extraction to extract an explanation from the reviews. We call this method as TMW. To evaluate the TMW1 were recruited a population composed of 44 men and women between 20 and 45 years of age to compare TMW1 with the Amazon Most Helpful Review (AMHR). They were inquired about the understanding, helpfulness and if they would like to use TMW in the future. The results have shown that 45.45% of the recruited population has chosen TMW as the most helpful review against 54.55% of AMHR. 50% of all people considered TMW easy to understand and 65.91% said they would like to use this system in the future to get an opinion.

Based on the prior results obtained in the experiments realized under TMW1, we propose a new architecture called TMW2. This structure keeps the main features of the first approach, but some improvements were made to get better results. To evaluate the TMW2 were recruited a population composed of 21 men and women between 20 and 35 years of age to compare TMW2 against Amazon Most Helpful Review (AMHR). The results have shown that 85.71% of the recruited population has chosen TMW as the most helpful review against 14.29% of AMHR.

47.62% of all people considered TMW easy to understand and 38.10% said they would like to use this system in the future to get an opinion about a movie.

The results obtained in both evaluation show us that it is a promising approach and useful in user's opinion. However, TMW presented deficiencies in the summarization process. The current limitations of the TMW summarization are related to: (a) **Coverage**. TMW just consider just the main topics in the review, not covering all topics available about the related movie; (b) **Redundancy**. The extractive summarization process also presents a redundancy, allowing that multiples sentences about the same topic; and (c) **Coherence**. The TMW also showed a coherence problem between sentences, chosen sentences that have no semantic link each other. The summarization also ignores the position of the sentence in the text allowing introductory sentences be inserted after the concluding sentences. It can hinder the understanding of the summary to the user.

The evaluation process also reveals a deficiency to measure the quality of the summary. The questions asked to the populations don't cover all the relevant aspects of the summary. To solve this problem, we recommend the adoption of a standard evaluation framework. The DUC [1] framework is a standard and a widely used into evaluation process of the summarization system. The linguistic quality questions are targeted to assess how readable and fluent the summaries are. A comparative evaluation with other opinion summarization also is an important manner to measure the quality of the approach.

7.1 Contributions

Given the results obtained, we now list the main contributions of this work, as follow:

Development of a new architecture to provide a personalized summary of reviews

As specified in Chapter 1, the users have faced a vast amount of reviews to get an opinion about a product or a service, such as a restaurant, movie or a book. To provide users better understanding of reviews, we proposed a new architecture to review summarization called Tell Me Why. Such architecture are composed of five modules, which represents the common tasks in the review summarization process. The output of this architecture is a piece of text using Natural Language Generation to provide users an explanation of reviews. Therefore, one of our main contribution in this work is a new architecture to summarize reviews.

[1] http://www-nlpir.nist.gov/projects/duc/duc2007/quality-questions.txt

Comparative evaluation of Tell Me Why architecture

The Amazon site allows people to say how much helpful is a review. The most voted is called The Most Helpful Review (AMHR). It is a cheap and quick method to perform a comparative evaluation of the system. To the best of our knowledge, we were the first to use the AMHR as the gold standard to perform an evaluation. In this configuration, the people who have said that always read review before seeing a movie is identified as the specialist. The results obtained from comparative evaluation show that TMW is functional, useful and easy to understand in the user's opinion.

Investigation and Employment of Automatic Text Summarization methods

The Tell Me Why architecture is composed of *Data Retrieval, Abstraction, Macro Planning, Sentence Extraction* and *Surface Realization*. The *Sentence Extraction* module is responsible to extract a couple of relevant sentences. Thus, we have performed a theory review to identify the most suitable technique to be employed for sentence extraction. One contribution of this work is the employment of Statistical-based and Graph-based methods into review summarization.

Intrinsic comparison of methods employed on sentences extraction

Additionally to a comparative Evaluation, we perform an assessment of the methods used in Sentence Extraction module. We used ROUGE-N to compare the performance of the state of the art in extractive summarization (MEAD). The intrinsic evaluation showed that both Statistical-based and Graph-based has good performance on the task of summarization. However, the Statistical-based has worse results. One contribution of this work is a comparative between these methods and the baseline.

7.2 Future work

Our work proposes a new architecture to provide to the user a review explanation. Additionally to the architecture, this work performed a study of summarization methods employed in the Sentence Extraction module. The future work related to this work are listed below.

Large Scale Evaluation.

Our procedure of evaluation achieved relevant results to review summarization. To understand the system performance in a real application, we propose a performance evaluation in large datasets. This assessment provides results of the TMW's applicability in large-scale use, such as collaborative review sites.

Individual study of TMW modules.

In this work, we have performed an investigation and evaluation of the methods to the Sentence Extraction module. As future work, we propose an individual and focused study in each module. An individual study on the TMW modules will provide an improvement of the result.

Sentiment Analysis.

Currently, each sentence of the text review is classified based on their numeric rating. However, the results have shown the sentences found in the review have different sentiment orientation. The same review can contain positive and negative sentences. Thus, the use of sentiment analysis can identify the individual orientation of the sentence.

Temporal Extraction

Currently, TMW provides a summary based on overall opinion. However, such approach can hide an opinion changing. It can be relevant in domains of service. As future work, we propose generate summaries exposing the people opinion about an item along the time.

Aspect-based summarization

As future work, we proposed generate summaries based on aspects. We aim to recognize the users' aspects of interest in an item based on their prior reviews. Thus, we can provide them a personalized summary that explains the pros and cons. Precisely, we want to create useful summaries that cover their aspects of interest to provide a suitable opinion. From this broader problem, it can be divided it into two sub-problems: (i) How to automatically discover users' aspects of interest based reviews prior generated by them; and (ii) How to create a summary that explain the pros and cons taking into account the user's aspects of interest.

REFERENCES

BANERJEE, S.; PEDERSEN, T. An adapted lesk algorithm for word sense disambiguation using wordnet. In: INTERNATIONAL CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS. **Proceedings...** [S.l.]: Springer, 2002. p. 136–145.

BOSSARD, A.; GÉNÉREUX, M.; POIBEAU, T. Description of the lipn system at tac 2008: Summarizing information and opinions. In: INTERNATION CONFERENCE ON TEXT ANALYSIS, Gaithersburg, USA. **Proceedings...** [S.l.]: Nist, 2008. p. 282–291.

BOSSARD, A.; GÉNÉREUX, M.; POIBEAU, T. Cbseas, a summarization system integration of opinion mining techniques to summarize blogs. In: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, Aber Wrac'h, France. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2009. p. 5–8.

CHAN, S. W. et al. Mining discourse markers for chinese textual summarization. In: INTERNATION WORKSHOP ON AUTOMATIC SUMMARIZATION. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2000. p. 11–20.

CHATTERJEE, P. Online reviews: do consumers use them? **Journal of Consumer Research**, Oxford University Press, p. 129–134, 2001.

CHEN, C. et al. Visual analysis of conflicting opinions. In: IEEE SYMPOSIUM ON VISUAL ANALYTICS SCIENCE AND TECHNOLOGY. **Proceedings**. [S.l.]: IEEE, 2006. p. 59–66.

CHEN, F.; HAN, K.; CHEN, G. An approach to sentence-selection-based text summarization. In: INTERNATIONAL CONFERENCE ON COMPUTERS, COMMUNICATIONS, CONTROL AND POWER ENGINEERING. **Proceedings...** [S.l.]: IEEE, 2002. v. 1, p. 489–493.

CHEUNG, J. C. **Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection**. Thesis (PhD) — Citeseer, 2008.

CHEVALIER, J. A.; MAYZLIN, D. The effect of word of mouth on sales: Online book reviews. **Journal of marketing research**, American Marketing Association, v. 43, n. 3, p. 345–354, 2006.

CHUANG, W. T.; YANG, J. Extracting sentence segments for text summarization: a machine learning approach. In: INTERNATIONAL CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. **Proceedings...** [S.l.]: ACM, 2000. p. 152–159.

DAS, M. et al. Mri: Meaningful interpretations of collaborative ratings. **Journal of Very Large Databases**, Citeseer, v. 4, n. 11, 2011.

DONAWAY, R. L.; DRUMMEY, K. W.; MATHER, L. A. A comparison of rankings produced by summarization evaluation measures. In: INTERNATION WORKSHOP ON AUTOMATIC SUMMARIZATION. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2000. v. 4, p. 69–78.

EDMUNDSON, H. P. New methods in automatic extracting. **Journal of the ACM**, ACM, v. 16, n. 2, p. 264–285, 1969.

ERKAN, G.; RADEV, D. R. Lexrank: graph-based lexical centrality as salience in text summarization. **Journal of Artificial Intelligence Research**, p. 457–479, 2004.

EVANS, D. K.; KLAVANS, J. L. Simfinderml: multilingual text similarity identification, experimentation, and evaluation. **Journal of ACM Computer Documentation**, ACM, p. 1–2, 2003.

GANESAN, K.; ZHAI, C.; HAN, J. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2010. p. 340–348.

GANU, G.; ELHADAD, N.; MARIAN, A. Beyond the stars: Improving rating predictions using review text content. In: INTERNATIONAL WORKSHOP ON THE WEB AND DATABASES. **Proceedings...** [S.l.]: ACM, 2009. v. 9, p. 1–6.

GATT, A.; REITER, E. Simplenlg: A realisation engine for practical applications. In: EUROPEAN WORKSHOP ON NATURAL LANGUAGE GENERATION. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2009. p. 90–93.

GUPTA, V.; LEHAL, G. S. A survey of text summarization extractive techniques. **Journal of Emerging Technologies in Web Intelligence**, v. 2, n. 3, p. 258–268, 2010.

HABOUSH, A. et al. Arabic text summarization model using clustering techniques. **Journal of Computer Science and Information Technology**, p. 2221–0741, 2012.

HAHN, U.; MANI, I. The challenges of automatic summarization. **Journal of IEEE Computer**, IEEE, v. 33, n. 11, p. 29–36, 2000.

HAN, J.; PEI, J.; YIN, Y. Mining frequent patterns without candidate generation. In: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA. **Proceedings...** New York, NY, USA: ACM, 2000. v. 29, n. 2, p. 1–12. ISSN 0163-5808.

HARMAN, D.; OVER, P. The effects of human variation in duc summarization evaluation. In: INTERNATION WORKSHOP ON TEXT SUMMARIZATION BRANCHES OUT. **Proceedings...** [S.l.]: ACL, 2004. p. 10–17.

HENNIG, L.; LABOR, D. Topic-based multi-document summarization with probabilistic latent semantic analysis. In: INTERNATION CONFERENCE ON RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING. **Proceedings...** Borovets, Bulgaria: ACL, 2009. p. 144–149.

HOVY, E.; LIN, C.-Y. Automated text summarization and the summarist system. In: INTERNATION WORKSHOP ON HELD. **Proceedings...** [S.l.]: Association for Computational Linguistics, 1998. p. 197–214.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. **Proceedings...** [S.l.]: ACM, 2004. p. 168–177.

JIN, F.; HUANG, M.; ZHU, X. A query-specific opinion summarization system. In: INTERNATIONAL CONFERENCE ON COGNITIVE INFORMATICS. **Proceedings...** [S.l.]: IEEE, 2009. p. 428–433.

JING, H. et al. Summarization evaluation methods: Experiments and analysis. In: AAAI SYMPOSIUM ON INTELLIGENT SUMMARIZATION. **Proceedings...** [S.l.]: AAAI Press, 1998. p. 51–59.

KAMPS, J. et al. Using wordnet to measure semantic orientations of adjectives. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION. **Proceedings...** Lisbon, PT, 2004. p. 1115—-1118.

KIM, H. D. et al. Comprehensive review of opinion summarization. **Journal of Computer Engineering and Applications**, v. 4, 2014.

KIM, H. D.; ZHAI, C. Generating comparative summaries of contradictory opinions in text. In: INTERNATION CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT. **Proceedings...** [S.l.]: ACM, 2009. p. 385–394.

KIM, Y.; SRIVASTAVA, J. Impact of social influence in e-commerce decision making. In: INTERNATIONAL CONFERENCE ON ELECTRONIC COMMERCE. **Proceedings...** [S.l.]: ACM, 2007. p. 293–302.

KU, L.-W.; LIANG, Y.-T.; CHEN, H.-H. Opinion extraction, summarization and tracking in news and blog corpora. In: AAAI SPRING SYMPOSIUM ON COMPUTATIONAL APPROACHES TO ANALYZING WEBLOGS. **Proceedings...** [S.l.]: AAAI Press, 2006. v. 100107.

KUPIEC, J. M. et al. **Automatic method of generating feature probabilities for automatic extracting summarization**. [S.l.]: Google Patents, 1998. US Patent 5,778,397.

KYOOMARSI, F. et al. Optimizing text summarization based on fuzzy logic. In: INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION SCIENCE, 2008. **Proceedings...** [S.l.], 2008. p. 347–352.

LABBE, C.; LABBE, D. How to measure the meanings of words? amour in corneille's work. **Journal of Language Resources and Evaluation**, Springer Netherlands, v. 39, n. 4, p. 335–351, 2005.

LABBE, C. et al. Rnr: Extracting rationale from online reviews and ratings. In: INTERNATIONAL CONFERENCE ON DATA MINING. **Proceedings...** [S.l.]: IEEE, 2010. p. 358–368.

LAPATA, M.; BARZILAY, R. Automatic evaluation of text coherence: Models and representations. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. **Proceedings...** [S.l.], 2005. v. 5, p. 1085–1090.

LESK, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: INTERNATIONAL CONFERENCE ON SYSTEMS DOCUMENTATION. **Proceedings...** [S.l.]: ACM, 1986. p. 24–26.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: INTERNATION WORKSHOP ON TEXT SUMMARIZATION BRANCHES OUT. **Proceedings...** [S.l.], 2004. p. 74–81.

LIN, C.-Y.; HOVY, E. Automatic evaluation of summaries using n-gram co-occurrence statistics. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON HUMAN LANGUAGE TECHNOLOGY. **Proccedings...** [S.l.]: Association for Computational Linguistics, 2003. v. 1, p. 71–78.

LIU, B. **Web data mining**. [S.l.]: Springer Science & Business Media, 2007.

LIU, B.; HU, M.; CHENG, J. Opinion observer: analyzing and comparing opinions on the web. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. **Proceedings...** [S.l.]: ACM, 2005. p. 342–351.

LU, Y.; ZHAI, C.; SUNDARESAN, N. Rated aspect summarization of short comments. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. **Proceedings...** [S.l.]: ACM, 2009. p. 131–140.

LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information. **Journal IBM of research and development**, IBM, v. 1, n. 4, p. 309–317, 1957.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. [S.l.]: Cambridge university press Cambridge, 2008.

MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. [S.l.]: MIT press, 1999.

MCAULEY, J. J.; LESKOVEC, J. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. **Proccedings...** Republic and Canton of Geneva, Switzerland, 2013. p. 897–908.

MCKEOWN, K. R. et al. Tracking and summarizing news on a daily basis with columbia's newsblaster. In: INTERNATIONAL CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY RESEARCH. **Proceedings...** [S.l.]: Morgan Kaufmann Publishers Inc., 2002. p. 280–285.

MEI, Q. et al. Topic sentiment mixture: modeling facets and opinions in weblogs. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. **Proceedings...** [S.l.]: ACM, 2007. p. 171–180.

MITCHELL, T. M. Machine learning and data mining. **Journal of ACM Communications**, ACM, v. 42, n. 11, p. 30–36, 1999.

MOGHADDAM, S.; ESTER, M. Opinion digger: an unsupervised opinion miner from unstructured product reviews. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT. **Proceedings...** [S.l.]: ACM, 2010. p. 1825–1828.

NENKOVA, A. Summarization evaluation for text and speech: issues and approaches. In: INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING. **Proceedings...** [S.l.], 2006.

NENKOVA, A. **Understanding the process of multi-document summarization: content selection, rewriting and evaluation**. Thesis (PhD) — Columbia University, 2006.

NENKOVA, A.; MASKEY, S.; LIU, Y. Automatic summarization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2011. p. 3.

NENKOVA, A.; PASSONNEAU, R.; MCKEOWN, K. The pyramid method: Incorporating human content selection variation in summarization evaluation. **ACM Transactions on Speech and Language Processing**, ACM, v. 4, n. 2, p. 4, 2007.

PALMER, D. D. Text pre-processing. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). **Handbook of Natural Language Processing, Second Edition**. Boca Raton, FL: CRC Press, Taylor and Francis Group, 2010. ISBN 978-1420085921.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. **Journal of Foundations and trends in information retrieval**, Now Publishers, v. 2, n. 1-2, p. 1–135, 2008.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: INTERNATION CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2002. v. 10, p. 79–86.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2002. p. 311–318.

POIBEAU, T. et al. **Multi-source, Multilingual Information Extraction and Summarization**. [S.l.]: Springer Science & Business Media, 2012.

POPESCU, A.-M.; ETZIONI, O. Extracting product features and opinions from reviews. In: **Natural language processing and text mining**. [S.l.]: Springer, 2007. p. 9–28.

PORTET, F. et al. Automatic generation of textual summaries from neonatal intensive care data. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE IN MEDICINE. **Proccedings...** [S.l.], 2007. p. 227–236.

RADEV, D. et al. Mead-a platform for multidocument multilingual text summarization. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION. **Proceedings...** [S.l.], 2004.

RADEV, D. R.; TAM, D. Single-document and multi-document summary evaluation via relative utility. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT. **Proccedings...** [S.l.]: ACM, 2003.

RATH, G.; RESNICK, A.; SAVAGE, T. The formation of abstracts by the selection of sentences. **Journal of American Society for Information Science**, American Society for Information Science, v. 12, n. 2, p. 139–141, 1961.

REITER, E.; DALE, R.; FENG, Z. **Building natural language generation systems**. [S.l.]: MIT Press, 2000.

ROTEM, N. **Open Text Summarizer**. 2012. Http://libots.sourceforge.net.

RUSSELL, S.; NORVIG, P. **Artificial intelligence: a modern approach**. [S.l.]: Citeseer, 1995.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Journal of Information processing & management**, Elsevier, v. 24, n. 5, p. 513–523, 1988.

SILVA, C.; RIBEIRO, B. The importance of stop word removal on recall values in text categorization. In: INTERNATIONAL CONFERENCE ON NEURAL NETWORKS. **Proceedings...** [S.l.]: IEEE, 2003. v. 3, p. 1661–1666.

STOYANOV, V.; CARDIE, C. Toward opinion summarization: Linking the sources. In: INTERNATION WORKSHOP ON SENTIMENT AND SUBJECTIVITY IN TEXT. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2006. p. 9–14.

TITOV, I.; MCDONALD, R. Modeling online reviews with multi-grain topic models. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. **Proceedings...** [S.l.]: ACM, 2008. p. 111–120.

TOUTANOVA, K.; MANNING, C. D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: INTERNATION CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND VERY LARGE CORPORA. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2000. p. 63–70.

TURNEY, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2002. p. 417–424.

TURNEY, P. D.; LITTMAN, M. L. Measuring praise and criticism: Inference of semantic orientation from association. **Journal of ACM Transactions on Information Systems (TOIS)**, v. 21, n. 4, p. 315–346, 2003.

UMAM, K. et al. Coverage, diversity, and coherence optimization for multi-document summarization. **Jurnal Ilmu Komputer dan Informasi**, v. 8, n. 1, p. 1–16, 2015.

WANG, L. et al. Query-focused opinion summarization for user-generated content. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS. **Proceedings...** [S.l.]: Association for Computational Linguistics, 2014. p. 1660–1669.

ZHUANG, L.; JING, F.; ZHU, X.-Y. Movie review mining and summarization. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT. **Proceedings...** [S.l.]: ACM, 2006. p. 43–50.

APPENDIX A - COMPLEMENTARY RESULTS

This appendix presents complementary results obtained from the experiments performed in the TMW2. The table 7.1 presents the sentences extracted from the mass of review by our approach. The first column shows the movie title of the related sentence. The column labeled as Features shows the features extracted from the Association Rule, to guide the sentence selection. The column labeled Sentence shows the phrase extracted from the mass of reviews. The sentiment shows the rating of the review, where the related sentence were extracted. Next, the Support and Confidence of the Association Rule is given. The Column labeled as Score shows the final scored obtained by the sub-modular function-based described in Chapter 5

Table 7.2 shows the TMW2 summary to teenager female. The firs column describe the movie title. The second show the features of the personalized summary. The table also shows a positive and negative explanation containing a short explanatory text and five sentences.

Table 7.1: Extracted Sentences from reviews

Movie	*Features*	*Sentence*	*Sentiment*	*Support*	*Confidence*	*Score*
Pulp Fiction	Female, Teenager	This movie would be on my top ten What can I say?	+	0.2020	0.4040	0.7770
Pulp Fiction	Female, Twenties	One part of the movie I like was that they never showed you what was in the case.	+	0.30502	0.5	0.7660
Pulp Fiction	Female, Thirties	PULP FICTION is Quentin Tarantino's MASTERPIECE! the best film he ever made!	+	0.3680	0.5	0.7968
Pulp Fiction	Male, Teenager	This movie would be on my top ten What can I say?	+	0.2020	0.4040	0.7770
Pulp Fiction	Male, Twenties	What you might not know is that the each and every scene is a little movie in of itself.	+	0.207	0.414	0.7895
Pulp Fiction	Male, Thirties	If you haven't seen it you have missed one of the best films of the last 20 years.	+	0.0893	0.5	0.7545
Pulp Fiction	Female, Teenager	Most people that have the slightest problem with a film will rate it one star so at least I'm fair.	-	0.2020	0.4040	0.9217
Pulp Fiction	Female, Twenties	Also a little known fact is that Jonathan Winters was to have a cameo in the film, but became ill during the shooting.	-	0.2020	0.5	0.9401
Pulp Fiction	Female, Thirties	BTW, the fact there are all these "Kid's Reviews" of this movie here is disturbing.	-	0.3680	0.5	0.8893
Pulp Fiction	Male, Teenager	Probably my favorite part of the movie, it ends in the basement of a pawn-shop, with a sword and and a perverted cop named Zed.	-	0.1122	0.5	0.8945
Pulp Fiction	Male, Twenties	This is HIS movie, and his dialogue and mannerisms are so in tune with Tarantino's writing, you wonder if this is how Jackson really is in life (hope not).	-	0.207	0.414	0.9081
Pulp Fiction	Male, Thirties	When he dies they will refer to him first and foremost as the director of PULP FICTION.	-	0.0893	0.5	0.9051
Forrest Gump	Female, Teenager	Thanks, Tom! Tom Hanks gives an outstanding performance in one of the best movies I have ever seen, and this one is also one of my favorites!	+	0.1185	0.355	0.7817

Forrest Gump	Female, Twenties	This movie is one of the best movies that I have ever seen; Tom Hanks has got to be one of the best actors of our time.	+	0.11853	0.355	0.8265
Forrest Gump	Female, Thirties	There is nothing I can say to those millions of you who have seen this All I can do, is give you my review.	+	0.11853	0.3555	0.7564
Forrest Gump	Female, Teenager	I had to turn my T V up and it started to buzz.	-	0.1185	0.3555	0.9099
Forrest Gump	Female, Twenties	I hate Tom Hanks, but he is a good actor and this show is interesting, if highly unprobable.	-	0.1185	0.3555	0.9365
Forrest Gump	Female, Thirties	I can easily read the words on signs in the movie.	-	0.1185	0.3555	0.9261
The Silence of the Lambs	Female, Teenager	"Hannibal the Cannibal", a genius serial killer who ate parts of his victims.	+	0.0885	0.5	0.7460
The Silence of the Lambs	Female, Twenties	The story, as we all know by now, revolves around haunted FBI Agent Clarice Starling (Jodie Foster) trying to get information out of imprisoned cannibalistic killer Hannibal Lecter (Anthony Hopkins) on a serial killer called Buffalo Bill (Ted Levine).	+	0.1933	0.5	0.8014
The Silence of the Lambs	Female, Thirties	Clarice Starling is a FBI trainee who is assigned by her boss Jack Crawford, to help in the case of Buffalo Bill, a serial killer who skins his victoms.	+	0.1097	0.5	0.7574
The Silence of the Lambs	Male, Teenager	As good as the movie is, am I the only person in the world who thinks both actors have done better work in other films?	+	0.092	0.5	0.7799
The Silence of the Lambs	Male, Twenties	Jodie Foster stars as Clarice Starling, an FBI agent in training.	+	0.209	0.5	0.8008
The Silence of the Lambs	Male, Thirties	This is an excellent movie but if you're one of the very few people out there who have never seen it before then I'd recommend that you should read the book first, because it's just fantastic.	+	0.0953	0.5	0.7527
The Silence of the Lambs	Female, Teenager	"The Silence of the Lambs" is a great movie in almost every way.	-	0.0885	0.4040	0.7940
The Silence of the Lambs	Female, Twenties	The movie was predictable at times and its direction was shaky at the same time.	-	0.1933	0.5	0.8640
The Silence of the Lambs	Female, Thirties	This glaring inaccuracy throughout the movie leads me to conclude that the Starling character is nothing but window dressing.	-	0.1097	0.5	0.8016

The Silence of the Lambs	Male, Teenager	And Hopkins’s seething, malevolent Lecter makes the knife-wielding Norman Bates in "Psycho" look like a boy scout.	-	0.092	0.5	0.8433
The Silence of the Lambs	Male, Twenties	Nonetheless, I find it every bit as entertaining now as I did then but for different reasons.	-	0.209	0.5	0.9247
The Silence of the Lambs	Male, Thirties	Her judgment proves better than his as they pursue Buffalo Bill and he duly acknowledges that.	-	0.0953	0.5	0.8428

Table 7.2: Related opinion summarization systems

Movie	*Personalized summary to*	*Positive Explanation*	*Negative Explanation*

Pulp Fiction	Female, Teenager	Few (66%) women in their thirties enjoy Pulp Fiction movie with rating of 4.7 on average, as shown in the following sentences extracted from the reviews: 1)*PULP FICTION is Quentin Tarantino's MASTERPIECE! the best film he ever made!* 2)*What you might not know is that the each and every scene is a little movie in of itself.* 3)*If you want to see something like you have never seen before, give this movie a try, and if you are like me, it will give new meaning to the term "great movie".* 4)*The person I was most shocked to see in this movie was Bruce Willis!* 5)*He has a bunch of the actors talking about how great a director he is and how he approached the making of the movie, but I came away with the same questions about the movie as I did when I watched the Special Features.*	Few (34%) women in their thirties dislike Pulp Fiction movie with rating of 2.6 on average, as shown in the following sentences extracted from the reviews: 1)*BTW, the fact there are all these "Kid's Reviews" of this movie here is disturbing.* 2)*However, many of these serve merely as entertainment, and don't reveal much in the way of character or story.* 3)*I love all of the actors, Bruce Willis as a boxer running from a man he was paid to throw a fright, John Travolta, and Samuel Jackson with his overacting both are hit men, sent to kill Willis.* 4)*There are also a few cameos, one of which is Christopher Walken who has a comical 5 minutes telling an 8 year old boy how a watch was smuggled safely so Walken could give it to the boy.* 5)*I absolutely love this movie; it is one of the few films that would be I would catch on cable late at night and I would find myself staying up just to finish watching it.*

Forrest Gump	Female, Teenager	Almost all (76.9%) women in their twenties enjoy Forrest Gump movie with rating of 4.6 on average, as shown in the following sentences extracted from the reviews: 1)*This movie is one of the best movies that I have ever seen; Tom Hanks has got to be one of the best actors of our time.* 2)*Not only will it enthrall you, it may just make you a better person.* 3)*This movie is the only movie I can say that every single person in my family has seen and loved from young to old and I know that all of them own it* 4)*I am also a fan of Tom Hanks and I do think that this is his best movie.* 5)*Thanks, Tom! Tom Hanks gives an outstanding performance in one of the best movies I have ever seen, and this one is also one of my favorites!*	Few (23.1%) women in their twenties dislike Forrest Gump movie with rating of 2.4 on average, as shown in the following sentences extracted from the reviews: 1)*I ordered this movie and I was really looking forward to it and it came in the mail right on time.* 2)*I hate Tom Hanks, but he is a good actor and this show is interesting, if highly unprobable.* 3)*Some of the lines have become part of our national culture.* 4)*when I began to watch the movie, the sound quality was not real good.* 5)*It's neat seeing exactly how all of the special effects were done or how sound and makeup were used to enhance mood in particular scenes.*

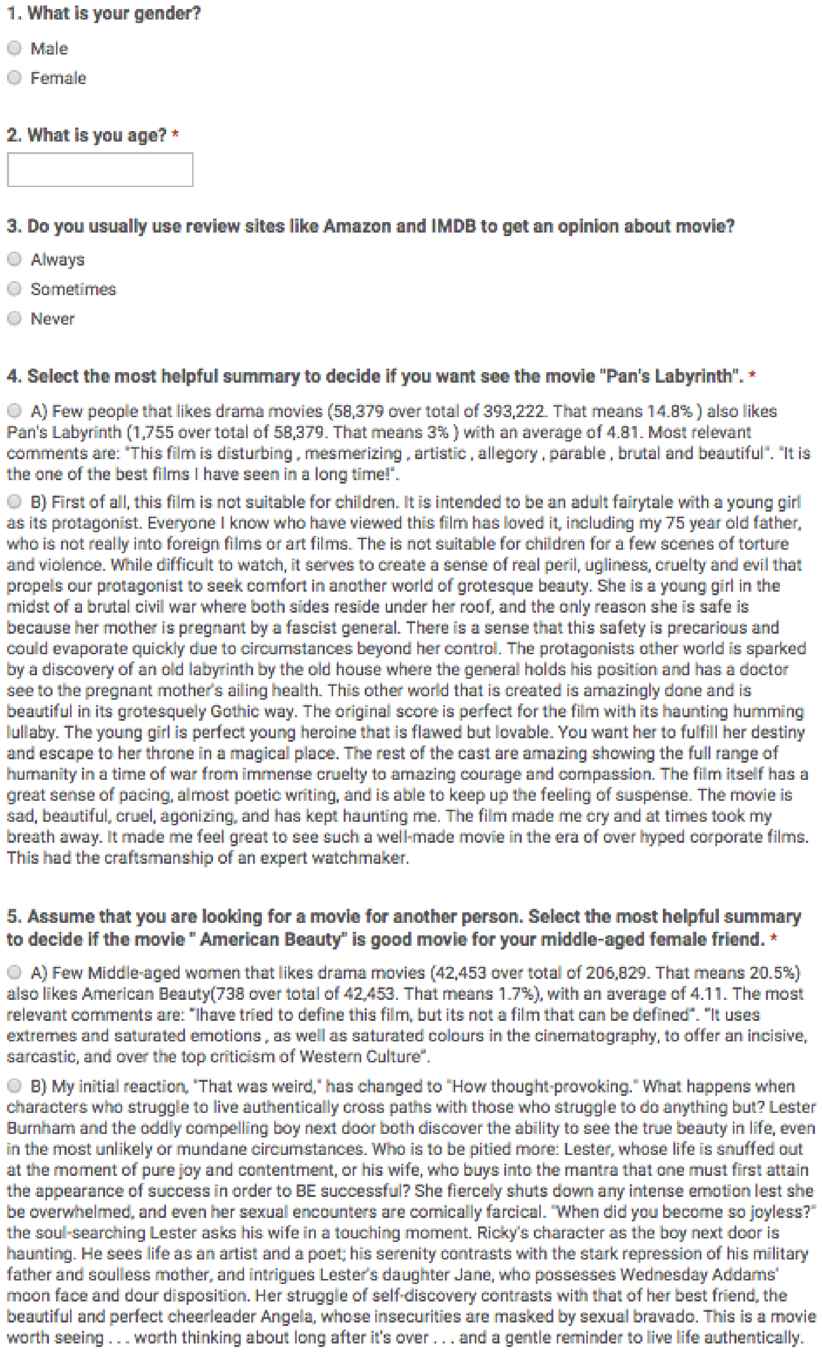
The Silence of the Lambs	Female, Teenager	Almost all (87.2%) women in their twenties enjoy The Silence of the Lambs movie with rating of 4.7 on average, as shown in the following sentences extracted from the reviews: 1) *Jodie Foster stars as Clarice Starling, an FBI agent in training.* 2)*The story, as we all know by now, revolves around haunted FBI Agent Clarice Starling (Jodie Foster) trying to get information out of imprisoned cannibalistic killer Hannibal Lecter (Anthony Hopkins) on a serial killer called Buffalo Bill (Ted Levine).* 3)*It's rare for a film to win Best Actor, Best Actress, Best Picture, Best Director and Best Screenplay Oscars, but The Silence of the Lambs did just that.* 4) *While I love this film - I think that it may be the best serial killer film ever - I do think that there were issues, of course.* 5) *The silence of the lambs is by far one of the best movies in movie history.*	Few (12.8%) women in their twenties dislike The Silence of the Lambs movie with rating of 2.6 on average, as shown in the following sentences extracted from the reviews: 1)*The movie was predictable at times and its direction was shaky at the same time.* 2)*This is one of the best movies to be released during the last twenty years.* 3)*It was a good film, no question about that but not a masterpiece like One Flew Over the Cuckoo's Nest.* 4)*I'm glad it won the five big Academy Awards (Best Picture, Best Actor, Best Actress, Best Director, Best Screenplay).* 5)*The commentary track, recorded 3 years after the movie came out, was overly analytical.*

APPENDIX B - FORMS

7.3 Experiment 1: Statistical-based method for sentence extraction

In this Section, we expose the form used to perform a comparative evaluation of the Statistical-based method for sentence extraction. Thus, Figures 7.1 and 7.2 show the form used in this task.

Figure 7.1 – TMW1 comparative evaluation form



1. What is your gender?

○ Male

○ Female

2. What is you age? *

3. Do you usually use review sites like Amazon and IMDB to get an opinion about movie?

○ Always

○ Sometimes

○ Never

4. Select the most helpful summary to decide if you want see the movie "Pan's Labyrinth". *

○ A) Few people that likes drama movies (58,379 over total of 393,222. That means 14.8%) also likes Pan's Labyrinth (1,755 over total of 58,379. That means 3%) with an average of 4.81. Most relevant comments are: "This film is disturbing , mesmerizing , artistic , allegory , parable , brutal and beautiful". "It is the one of the best films I have seen in a long time!".

○ B) First of all, this film is not suitable for children. It is intended to be an adult fairytale with a young girl as its protagonist. Everyone I know who have viewed this film has loved it, including my 75 year old father, who is not really into foreign films or art films. The is not suitable for children for a few scenes of torture and violence. While difficult to watch, it serves to create a sense of real peril, ugliness, cruelty and evil that propels our protagonist to seek comfort in another world of grotesque beauty. She is a young girl in the midst of a brutal civil war where both sides reside under her roof, and the only reason she is safe is because her mother is pregnant by a fascist general. There is a sense that this safety is precarious and could evaporate quickly due to circumstances beyond her control. The protagonists other world is sparked by a discovery of an old labyrinth by the old house where the general holds his position and has a doctor see to the pregnant mother's ailing health. This other world that is created is amazingly done and is beautiful in its grotesquely Gothic way. The original score is perfect for the film with its haunting humming lullaby. The young girl is perfect young heroine that is flawed but lovable. You want her to fulfill her destiny and escape to her throne in a magical place. The rest of the cast are amazing showing the full range of humanity in a time of war from immense cruelty to amazing courage and compassion. The film itself has a great sense of pacing, almost poetic writing, and is able to keep up the feeling of suspense. The movie is sad, beautiful, cruel, agonizing, and has kept haunting me. The film made me cry and at times took my breath away. It made me feel great to see such a well-made movie in the era of over hyped corporate films. This had the craftsmanship of an expert watchmaker.

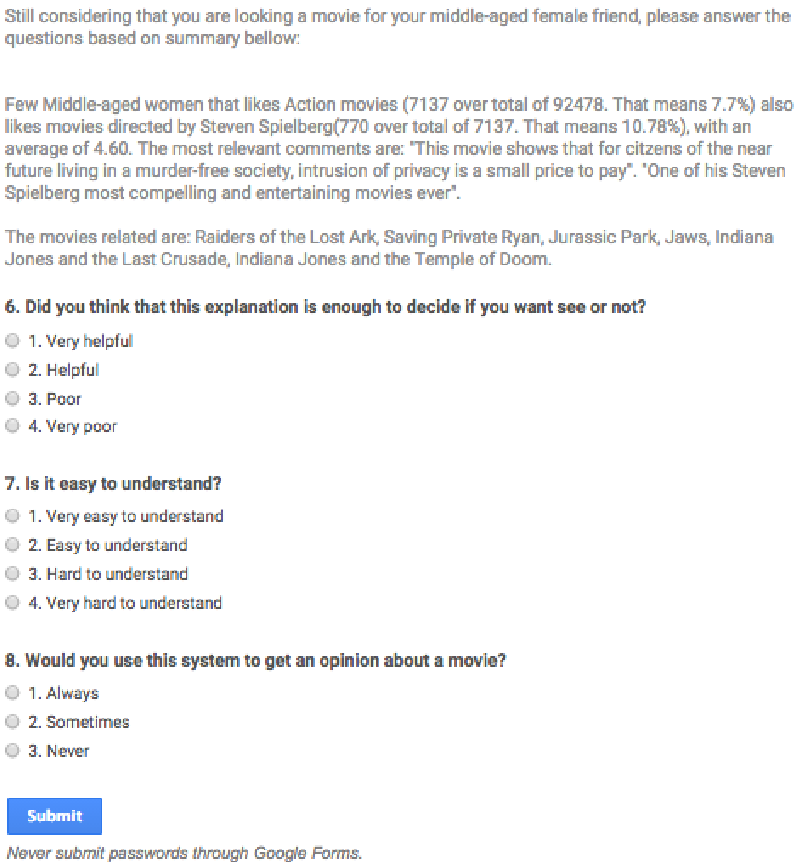
5. Assume that you are looking for a movie for another person. Select the most helpful summary to decide if the movie " American Beauty" is good movie for your middle-aged female friend. *

○ A) Few Middle-aged women that likes drama movies (42,453 over total of 206,829. That means 20.5%) also likes American Beauty(738 over total of 42,453. That means 1.7%), with an average of 4.11. The most relevant comments are: "Ihave tried to define this film, but its not a film that can be defined". "It uses extremes and saturated emotions , as well as saturated colours in the cinematography, to offer an incisive, sarcastic, and over the top criticism of Western Culture".

○ B) My initial reaction, "That was weird," has changed to "How thought-provoking." What happens when characters who struggle to live authentically cross paths with those who struggle to do anything but? Lester Burnham and the oddly compelling boy next door both discover the ability to see the true beauty in life, even in the most unlikely or mundane circumstances. Who is to be pitied more: Lester, whose life is snuffed out at the moment of pure joy and contentment, or his wife, who buys into the mantra that one must first attain the appearance of success in order to BE successful? She fiercely shuts down any intense emotion lest she be overwhelmed, and even her sexual encounters are comically farcical. "When did you become so joyless?" the soul-searching Lester asks his wife in a touching moment. Ricky's character as the boy next door is haunting. He sees life as an artist and a poet; his serenity contrasts with the stark repression of his military father and soulless mother, and intrigues Lester's daughter Jane, who possesses Wednesday Addams' moon face and dour disposition. Her struggle of self-discovery contrasts with that of her best friend, the beautiful and perfect cheerleader Angela, whose insecurities are masked by sexual bravado. This is a movie worth seeing . . . worth thinking about long after it's over . . . and a gentle reminder to live life authentically.

Source: Developed by author

Figure 7.2 – TMW1 comparative evaluation form



Still considering that you are looking a movie for your middle-aged female friend, please answer the questions based on summary bellow:

Few Middle-aged women that likes Action movies (7137 over total of 92478. That means 7.7%) also likes movies directed by Steven Spielberg(770 over total of 7137. That means 10.78%), with an average of 4.60. The most relevant comments are: "This movie shows that for citzens of the near future living in a murder-free society, intrusion of privacy is a small price to pay". "One of his Steven Spielberg most compelling and entertaining movies ever".

The movies related are: Raiders of the Lost Ark, Saving Private Ryan, Jurassic Park, Jaws, Indiana Jones and the Last Crusade, Indiana Jones and the Temple of Doom.

6. Did you think that this explanation is enough to decide if you want see or not?

○ 1. Very helpful

○ 2. Helpful

○ 3. Poor

○ 4. Very poor

7. Is it easy to understand?

○ 1. Very easy to understand

○ 2. Easy to understand

○ 3. Hard to understand

○ 4. Very hard to understand

8. Would you use this system to get an opinion about a movie?

○ 1. Always

○ 2. Sometimes

○ 3. Never

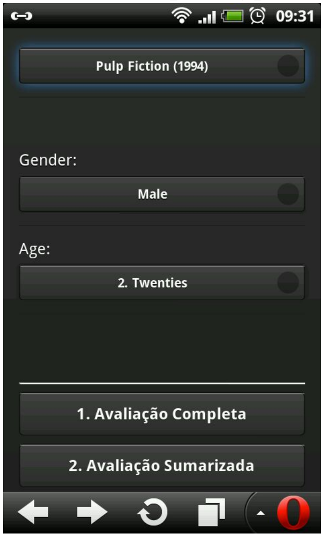
Never submit passwords through Google Forms.

Source: Developed by author

7.4 Experiment 2: Graph-based method for sentence extraction

To evaluate Graph-based method for sentence extraction, we built a prototype to collect the people opinion about reading a summary in a mobile device. Figure 7.3 show the prototype's main screen. At the top of the main screen, people were invited to choose a movie title from the movie list.

Figure 7.3 – The TMW prototype's main screen



Pulp Fiction (1994)

Gender:

Male

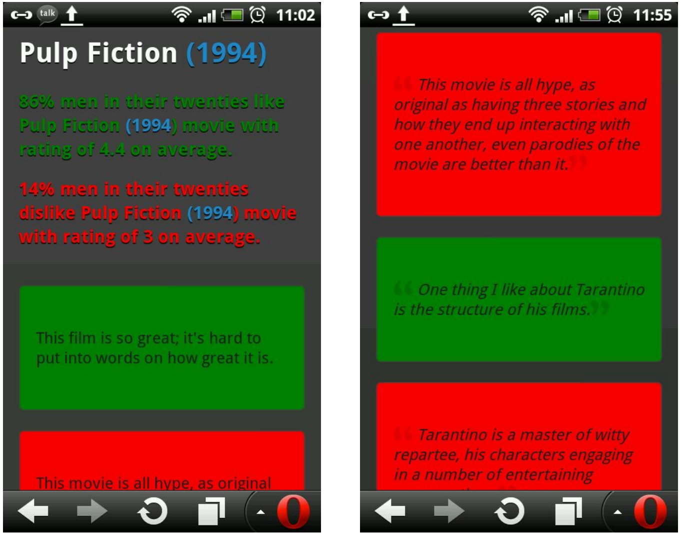
Age:

2. Twenties

Source: Developed by authors

After chose a movie title, enter their personal data, the people are invited to read two reviews from the film selected, being one from AMHR and other from TMW. Figure 7.6 shows the TMW's summary about the Pulp Fiction movie. The TMW2's summary are composed by: (i) The movie title and release date; (ii) A short personalized positive summary about the related movie, showed in green; (iii) A brief personalized negative summary about the related movie, showed in red; (iv) Five positive sentences, showed into green box; and (v) Five negative sentences, showed into red box.

Figure 7.4 – Prototype presenting the TMW2's summary



Pulp Fiction (1994)

86% men in their twenties like Pulp Fiction (1994) movie with rating of 4.4 on average.

14% men in their twenties dislike Pulp Fiction (1994) movie with rating of 3 on average.

This film is so great; it's hard to put into words on how great it is.

This movie is all hype, as original

“This movie is all hype, as original as having three stories and how they end up interacting with one another, even parodies of the movie are better than it.”

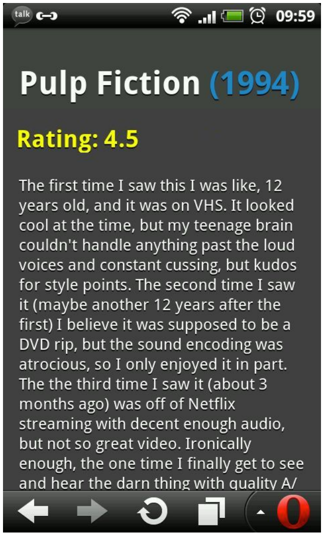
“One thing I like about Tarantino is the structure of his films.”

“Tarantino is a master of witty repartee, his characters engaging in a number of entertaining

Source: Developed by authors

Figure 7.5 presents the Amazon Most Helpful Review (AMHR) about the Pulp Fiction movie. Added to the text review, the screen also show the rating and the release movie date.

Figure 7.5 – The prototype showing the AMHR



Pulp Fiction (1994)

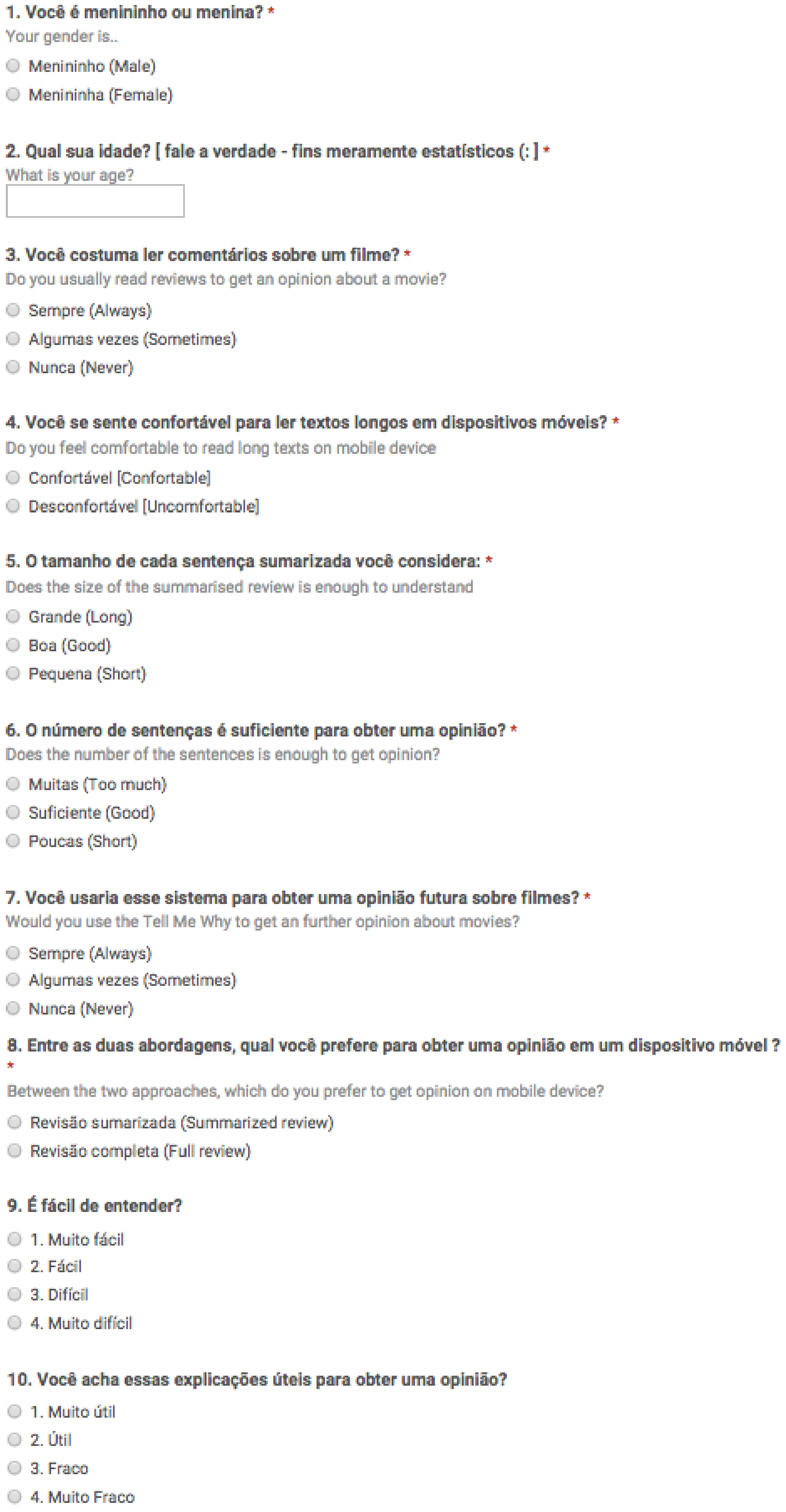
Rating: 4.5

The first time I saw this I was like, 12 years old, and it was on VHS. It looked cool at the time, but my teenage brain couldn't handle anything past the loud voices and constant cussing, but kudos for style points. The second time I saw it (maybe another 12 years after the first) I believe it was supposed to be a DVD rip, but the sound encoding was atrocious, so I only enjoyed it in part. The the third time I saw it (about 3 months ago) was off of Netflix streaming with decent enough audio, but not so great video. Ironically enough, the one time I finally get to see and hear the darn thing with quality A/

Source: Developed by authors

After use the prototype to get a movie opinion, the population were invited to response a survey. Figures 7.6 show the form that the users answered.

Figure 7.6 – TMW2 comparative evaluation form



1. Você é menininho ou menina? *

Your gender is..

○ Menininho (Male)

○ Menininha (Female)

2. Qual sua idade? [fale a verdade - fins meramente estatísticos (:] *

What is your age?

3. Você costuma ler comentários sobre um filme? *

Do you usually read reviews to get an opinion about a movie?

○ Sempre (Always)

○ Algumas vezes (Sometimes)

○ Nunca (Never)

4. Você se sente confortável para ler textos longos em dispositivos móveis? *

Do you feel comfortable to read long texts on mobile device

○ Confortável [Confortable]

○ Desconfortável [Uncomfortable]

5. O tamanho de cada sentença sumarizada você considera: *

Does the size of the summarised review is enough to understand

○ Grande (Long)

○ Boa (Good)

○ Pequena (Short)

6. O número de sentenças é suficiente para obter uma opinião? *

Does the number of the sentences is enough to get opinion?

○ Muitas (Too much)

○ Suficiente (Good)

○ Poucas (Short)

7. Você usaria esse sistema para obter uma opinião futura sobre filmes? *

Would you use the Tell Me Why to get an further opinion about movies?

○ Sempre (Always)

○ Algumas vezes (Sometimes)

○ Nunca (Never)

8. Entre as duas abordagens, qual você prefere para obter uma opinião em um dispositivo móvel ? *

Between the two approaches, which do you prefer to get opinion on mobile device?

○ Revisão sumarizada (Summarized review)

○ Revisão completa (Full review)

9. É fácil de entender?

○ 1. Muito fácil

○ 2. Fácil

○ 3. Difícil

○ 4. Muito difícil

10. Você acha essas explicações úteis para obter uma opinião?

○ 1. Muito útil

○ 2. Útil

○ 3. Fraco

○ 4. Muito Fraco

Source: Developed by the author