Universiteit Leiden

ICT in Business

Humor Recognition Based on Background Knowledge:

A Preliminary Study

Name: **Yan Zhang**

Student-no: **S1551159**

Date: **26/08/2016**

1st supervisor: **Enrique Larios Vargas (LIACS)**

2nd supervisor: **Peter van der Putten (LIACS)**

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Humor is said to be an essential cognitive activity of humans. The amusing effect of humor helps to repel tensions or anxieties and to build common understandings among participants in a conversation. Besides, humor always touches social or moral boundaries, which breaks taboos and clichés, and inspires different perspectives of understanding on existing social conventions. Such contrasts of interpretations, or surprisingly associating incompatible concepts or frames, attract attentions and stimulate memorizing while arousing laughter. In addition, the enthymematic characteristic of humor has powerful persuasive capabilities, via letting hearers agree on the unstated premises hidden in punchlines. These powers of humor will benefit human world a lot, ranging from technology developments such as human-machine interactions and artificial intelligence, to society and economics such as business communications, advertising and commercials, and educations. As the more advanced development of artificial intelligence and deeper involvement of machines in humans' lives, humor is a fantastic tool to make machines more ‘like' humans and to provide human-like services. Thus, understanding humor automatically becomes a critical task for scientists. Despite not receiving enough attention, nor being taken as a serious problem, computational humor has been studied for decades and it has been proven that computational approaches can be successfully applied to humor recognition. However, the existing studies neglected the importance of background knowledge when understanding verbal humor, which can be critical to ‘get' the humor. This background knowledge is often not stated directly in the text because they can be simply inferred by common senses, social conventions or cultural knowledge. But humor cannot be understood automatically if this background is lacking. This thesis proposes a machine learning based method to recognize and detect verbal humor in texts based on background knowledge using three knowledge databases: WordNet, ConceptNet, SentiWordNet, and to automatically extract the words that reveal humorous correlations between incompatible concepts.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Humor is an essential human cognitive activity. It also strongly relates to topics such as attention and memory, motivation and persuasion (Chung, 2003). Humor in conversations eliminates the distance between people, removes anxieties, helps to shape common understandings, and therefore improves interpersonal communications. These interactive characteristics of humor will bring benefits to various fields, such as intelligent systems development, education, advertisement, games and entertainment. If researchers are going to create an intelligent interactive agent, humor should be used as a powerful instrument for more natural human-machine interactions. Through the purposeful use of humor, machines can pretend to be creative and have human-like 'emotions'. Also, some researchers leverage this benefit of humor in educations for kids, to attract continuous attentions and shape a relaxing and pleasurable learning environment. More importantly, humor, as a creative language of humans, reveals surprising and latent correlations among concepts, which always breaks social conventions or common understandings, but at meanwhile lead to the second reflection on those clichés. Hence, there are many benefits of humor in practice, which gives motivations to the study of computational humor.

Computational humor is a subclass of natural language studies, detecting humor is a sub-category of computational humor. Many models or prototypes have been established to detect humor in texts, with two main approaches: stylistic-based and content-based. The stylistic-based approach detects verbal humor by analyzing humor-specific stylistic features. Phonetic style of verbal humor is a typical humor-specific stylistic feature in one-liner, which includes word repetitions or rhymes that produce a comic effect [1]. One example could be "*Q: Why was six afraid of seven? A: Because seven 'ate' nine.*" Word 'eight' sounds similar to 'ate', which lead to different interpretations of the text. Ambiguity, or polysemy of words, is another stylistic feature of jokes, which has been adopted as a key factor to evaluate incongruity of humor in previous studies. Taking the joke "*Two fish in a tank. One turns to another and says: 'Do you know how to drive this?'*" as an example, the multi-meanings of word 'tank' make two contrast situations here. These humor-specific stylistic features are usually formulated merely for a certain type of joke.

However, the understanding of humor does not only rely on these surface features but also requires background knowledge and logic reasoning. *"It is uncontroversial to observe that understanding a particular joke may require not just knowledge of the language used to convey the joke, but also other types of knowledge (factual, cultural, social, etc.)* [2]*"* Sometimes, humor could be hard to understand, due to the lack of cultural or other background information. And in order to understand humor as the same as it is perceived by humans, not only the language features but a representation of knowledge and an ability to reason should be combined in the model [3]. Very few studies have stressed the importance of background knowledge as a complementary dimension to understanding verbal humor. Therefore, this thesis will give a preliminary test on creating a model to detect verbal humor using background knowledge, and to extract humor words with highest comic effects.

1.1 Research Question

The main research question for this thesis is:

How to detect verbal humor based on background knowledge? And what are the significant factors for a text to be categorized as humorous?

Sub-questions used to investigate on this question are:

- How to use background knowledge to better reflect the meanings of humorous texts?
- How to construct background knowledge for humorous texts?
- How to represent text features efficiently?
- How to extract humor words those made most contributions to the comic effects of a text?

1.2 Research Contribution

This thesis proposes a way to detect verbal humor through both analyzing its semantic meanings and the unstated background knowledge related to the texts. To put in another word, this thesis attempts to explore how background knowledge, including related common factual information and cultural/social knowledge, can be leveraged in automatic humor recognition. At

the end, a classification model will be created to distinguish humorous texts from serious or normal texts, and a list of co-occurred concepts will be extracted as the key humor factors for each joke. These co-occurred concepts bear creative interrelations in a certain degree that meet with the incongruity theory of humor, and it will give insights to future studies to discover rules in creative bisociations and then be applied to educations or business applications.

1.3 Thesis Outline

The paper is organized in the following sequence. Some theoretical basis of humor and existing computational humor models are introduced in the theoretical background in Chapter 2, as well as the motivation of introducing background knowledge into humor detection. Chapter 3 will explain how we plan to construct background knowledge for humor or non-humor texts, as well as background introductions of the three databases used to extract such semantic and background knowledge. Chapter 4 contains main parts of the methodology and procedures to build up the model and extract key humor factors. Then details of the experiment are presented in Chapter 5, followed by Chapter 6 giving the results. Chapter 7 discusses challenges and contributions of the model, and possible improvements in the future. Finally, main conclusions will be summarized in Chapter 8.

Chapter 2 Theoretical Background

Many philosophers, psychologists, and linguists have shared their viewpoints on the theory of humor. And many models have been built up to generate or detect humor in recent decades, yet, a few of which has included background knowledge as a key feature of humor. This chapter will give a broad introduction to these theories and models, and to explain the inspirations from these previous works and their relations to our model.

2.1 Humor Theories

Although many researchers have put their views on the essence of humor, it is also pointed out by Attardo that, a general theoretical definition or a general category for humor is difficult to be defined. The study of humor is so complex because humor can be triggered by a great variety of stimulus. It is hard to grasp a single rule from these triggers. Humor is a phenomenon that has relations to various topics as well. Researchers used the method of 'semantic fields' to illustrate words and concepts that relate to this field, as shown in Figure 2.1. In spite of the complication of humor, three classes of theories have raised broadly agreements among the field of humor studies, those are the theory of superiority, relief, and incongruity. This chapter will give brief introductions to each of the three theories. Comparing all of them, the theory of incongruity and 'bisociation' theory will be adopted as the basis to conduct our research in this thesis.

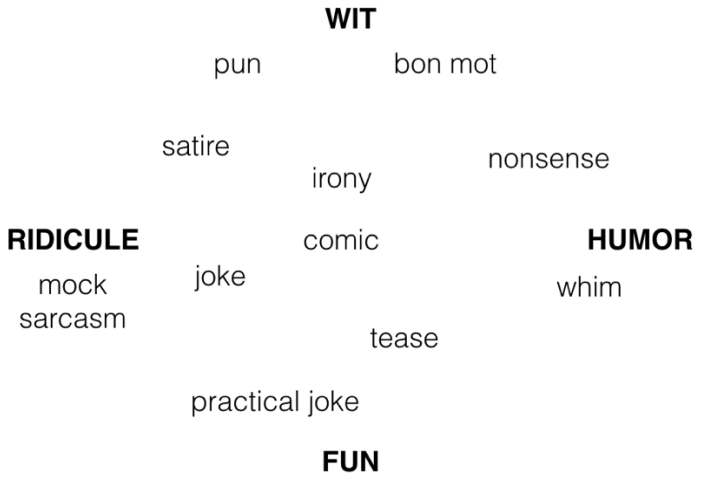

Figure 2.1 The Semantic Field of "humor"

2.1.1 Theory of Superiority

According to superiority theory, there is a kind of high-and-low relationship between the addresser and the target people, group or idea, specifically the addresser always looks down to what he laughs at [4]. Thomas Hobbes is the first researcher proposed this theory. He saw "laughter as a kind of sudden glory". This glory usually caused by aggressive satisfactions based on political, ethnic or gender ground. It is true that people laugh at situations such as a clown's retarded and exaggerated body actions, or unusual accents and dialects, or a politician's inappropriate speech. However, the theory of superiority is not sufficient to explain many cases. For example, people also laugh at themselves, which is not able to have a 'high-and-low' relation within oneself. Also, since there are no obvious benchmarks to measure the superiority character of humor, this theory will not be taken as the main theoretical basis for the thesis. One inspiration from this theory to our model, though, is that some kind of contrast must exist in humor.

2.1.2 Relief Theory

Another theory giving more psychoanalytical explanations to humor is the relief theory. Sigmund Freud in the book *Jokes and Their Relation to the Unconscious* [5] said that conventional morality and social rules refrain us from natural impulses. And humor is a release of those suppressed urges, aggressively or sexually, in a socially acceptable manner. Spencer also pointed out in his book *The Physiology of Laughter* that laughter releases nervous energy that might lead to larger practical actions [6]. More specifically, the setup of a pun usually leads people to a certain understanding or expectation of a scenario inferred by simple logic or life experiences, but the expectation is forced into a totally different interpretation by the punchline at the end. For example, people will find an insult funny if it, at first sight, appears as a compliment. The words of compliment in this example might lead people to feel self-satisfied but which later in a while being proved to be superfluous by the words of degradation. In Spencer's words, this self-satisfaction is a kind of nervous energy, which will accumulate until being released at the end and finally caused laughter. This theory seems to be more generalized than the superiority theory to include self-mockery, but still, it is hard to monitor such nervous energy with computers. Moreover, each person might feel different levels of nervous energy

under the same situation. So this theory will not be adopted as the main theory in the thesis either, but only gives inspirations to our model that we should include sentiment as a feature of humor.

2.1.3 Incongruity Theory

One widely adopted theory of humor, the theory of incongruity, or contradiction, 'bisociation' is more from a cognitive point of view. This class of theory assumes that humorous texts or acts involve two different planes of content, or called frames or scripts. These two frames are incompatible or contradict but have some kind of overlap which makes frame-shifting from one to another possible. Kant posits *"Laughter is an affection arising from the sudden transformation of a strained expectation into nothing...* [7]*".* A joke or a one-liner always consists of two parts: a setup and a punch(line). The setup will intentionally lead spectators to a certain expectation about a topic or a situation, based on one common frame of reference like life experiences, but the punchline at the end will present a surprising and different frame of explanation about the situation. The two frames of references are usually incompatible or incongruous with each other, but still contains some linkages that force the shifting from one reference to another. The former frame of reference always follows one's mental patterns established from experiences or educations, but the latter frame reveals a logically correct and creative way of interpretation. This steep twist between the mental pattern/expectation to the actual sense perception makes people laugh. Salvatore Attardo also agreed with this theory in his book *Linguistic theories of humor* [8]. This theory has aligned with Arthur Koestler's theory in a certain degree, though Koestler called it as 'bisociation' rather than 'incongruity'.

2.1.4 Humor and Creativity

Besides these three classical theories of humor, a unique and fascinating feature of humor is its close relation to creativity. In the book *The Act of Creation* [9], Arthur Koestler explained how creativity works in art, science, and humor. In his point, there are three types of creativity broadly: humor, discovery, art. He posits that these three types of creativity have different "emotion mood": scientific discovery is more of neutral creativity, art is sympathetic and tragic creativity, and humor is aggressive creativity. All of them are founded by 'bisociation'. As Figure 2.2 shows, M_1 and M_2 are two different planes of content, also called frames. They are

incompatible with each other but be linked by creativity, letting shift from one into the other possible. He concludes "*recipients feel the humor when emotions fall behind bisociation of two habitually incompatible matrices, namely associative contexts and frames of reference*". Since this theory has common features with the theory of incongruity, and there are many successful attempts to build a model based on it, we will adopt Koestler's theory combined with incongruity theory to build our model.

He also believes humor is motivated either by self-defense or assaulting impulse. "*Laughter is a luxury reflex which could arise only in a creature whose reason has gained a degree of autonomy from the urges of emotion, and enables him to perceive his own emotions as redundant to realize that he has been fooled.*" [9] This quote reveals a very philosophical relation between humor and critical thinking. Compared to art and science, humor could be the most pervasively understandable creative human activities. Everybody can understand humor or be a producer of humor, no matter age or culture. Understanding humor is a way trying to think out of the box, or let people re-think about some axioms or logic reflex. This could bring much joy to the gradually settled and fixed adult world.

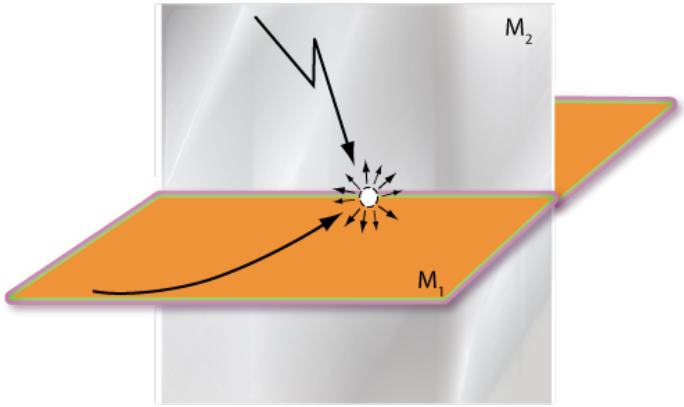

Figure 2.2 The pattern underlying creative acts

2.2 Computational Humor

As a relatively new research area, the first international computational humor conference was held in Twente Workshops on Language Technology, Enschede, University of Twente, Netherlands, 1996. The main task of computational humor is to understand and produce humor

automatically using artificial intelligence technologies. Although many pieces of researches have been conducted, the contribution made to build a model to analyse and generate humor automatically is still limited. And more efforts have been put into joke generation rather than humor understanding since the nuanced nature of humor is difficult to be generally defined [10]. Many existing computational model focus on the analysis of humor through its linguistic styles, lexical or semantic meanings of words. *HAHAcronym* is a computational humor system developed to generate humorous interpretations for existing acronyms, or to build humorous acronyms with the start words provided by users. As one example shown in the article [11] "*it turned IJCAI—International Joint Conference on Artificial Intelligence— into Irrational Joint Conference on Antenuptial Intemperance."* The method used in the application is based on the incongruity theory which was realized by coupling concepts from different domains, such as 'religion' versus 'technology' or 'sex' versus 'religion'. It is a good attempt to imitate incongruity in a rough way, but it is still focused only on the lexical or stylistic style of texts.

Chapter 3 Knowledge Bases

The most prevailing theory about humor is the incongruity theory, and many researchers have pointed out that understanding humor requires background knowledge. The main humor theory we choose here is the theory of incongruity combined with Koestler's 'bisociation' theory. They have one similarity that: verbal humor consists of some kind of frame-shifting. Setup of a joke will lead the listener or receiver to a most salient frame of understanding about a topic or a situation, but the punchline will be an obstacle for people to go further under the former frame or script. "*The next and the most critical step – a leap from the failed script to a suitable alternative – remains totally outside the capacities of combinatorial rules and the receiver will be able to achieve it only through intuitive trial and error, using his/her encyclopedic knowledge, or WORLD INFORMATION, as Raskin calls it.* [12]" In order to include this 'world knowledge' in detecting verbal humor, this thesis proposed a content-based method for humor recognition, leveraging both semantic meanings and background knowledge. The background knowledge includes factual, cultural and social knowledge, which are retrieved from mainly two knowledge databases, namely WordNet and ConceptNet. And the third one, SentiWordNet is used to formulate the sentiment fluctuation for each text. In this chapter, it will give introductions about background and functionalities of each knowledge databases, as well as their possible contributions to our model.

3.1 WordNet

In order to understand humor texts, lexical and semantic meanings of words should first be analyzed. WordNet [13] [14] is a large lexical network formed by synsets and their interrelations. It contains English words of *nouns, verbs, adjectives* and *adverbs*. Senses of words are grouped into synsets (synonyms), and each synset represents a distinct sense of a word. Synsets are interlinked by lexical relations. WordNet has 117,000 synsets and each synset contains a brief definition ("gloss") and an example sentence. The lexical and semantic relations revealed by WordNet consist of synonymy, antonymy, hyponymy, meronymy, troponymy and entailment. Examples and meanings of these relations are summarized in Table 3.1.

Semantic relation	Syntactic Category	Examples
Synonymy (Similar)	None Verb Adj Adv	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (Opposite)	Adj Adv Noun Verb	wet, dry rapidly, slowly top, bottom rise, fall
Hyponymy (Subordinate)	Noun	sugar maple, maple maple, tree tree, plant
Meronymy (Part)	Noun	brim, hat gin, martini ship, fleet
Troponymy (Manner)	Verb	march, walk whisper, speak
Entailment	Verb	drive, ride divorce, marry

Table 3.1 Types of semantic relations exist in WordNet [15]

"WordNet was designed to model lexical memory rather than represent lexical knowledge, so it excludes much of a speaker's knowledge about both semantic and syntactic properties of verbs. [16]" Hence there is a necessity to introduce ConceptNet as a complementary common-sense knowledge source.

3.2 ConceptNet

ConceptNet [17] is a semantic network created by MIT Media Lab, aiming to give computers access to common-sense knowledge, the kind of information that ordinary people know but usually leave unstated. This semantic network is stored in a graph structure, in which words or short phrase are presented as nodes, also called terms or concepts, and links between nodes represent lexical relations between terms. An assertion is a directive edge linking different nodes, revealing the common-sense relations between concepts, which could be daily basic knowledge, cultural or scientific knowledge. ConceptNet has 321,993 English concepts and fully 1.6 million assertions over 27 types of relations. These different types of relations have fallen into mainly 8 categories, categories and a number of assertions as shown in Table 3.2.

Category	Assertions
K-Lines	**1,035,035**
ConceptuallyRelatedTo	816,737
SuperThematicKLines	160,181
ThematicKLines	58,117
Functional	**103,556**
CapableOfReceivingAction	57,600
UsedFor	45,956
Agents	**89,313**
CapableOf	89,313
Things	**46,828**
IsAs	16,720
PartOf	12,934
PropertyOf	9,135
DefinedAs	6,520
MadeOf	1,519
Events	**35,317**
SubEventOf	22,764
FirstSubEventOf	4,453
PrerequisiteEventOf	4,092
LastSubEventOf	4,008
Affective	**31,196**
MotivationOf	24,483
DesireOf	6,713
Spatial	**28,805**
LocationOf	28,805
Causal	**15,303**
EffectOf	9,057
DesirousEffectOf	6,246

Table 3.2 Breakdown of concepts over categories in ConceptNet [18]

On the one hand, ConceptNet provides factual knowledge to be the presuppositions for various kinds of text understandings. Taking concept 'table' as an example, some typical concepts related to it and relations between 'table' and these nodes are:

- food_AtLocation_table: Something you find on a table is food.
- table_AtLocation_kitchen: You are likely to find a table in the kitchen.
- leg_PartOf_table: A leg is part of a table.

- table MadeOf wood: a table can be made of wood.
- table UsedFor eat meal: a table is for eating meals.

On the other hand, ConceptNet provides more complex social and cultural background knowledge as the basis of simple inferences. Another example that gives a graphic explanation to this 'simple inference' functionality of ConceptNet is shown in Figure 3.1. It is a semantic graph excerpted from ConceptNet to describe compound concepts that composed by a verb (e.g. 'drink') with a noun ('coffee') phrase or a prepositional phrase ('in morning') [17]. It reveals the ability of ConceptNet to connect related concepts of a given node to make a semantic network concerning this core concept. Such a semantic network

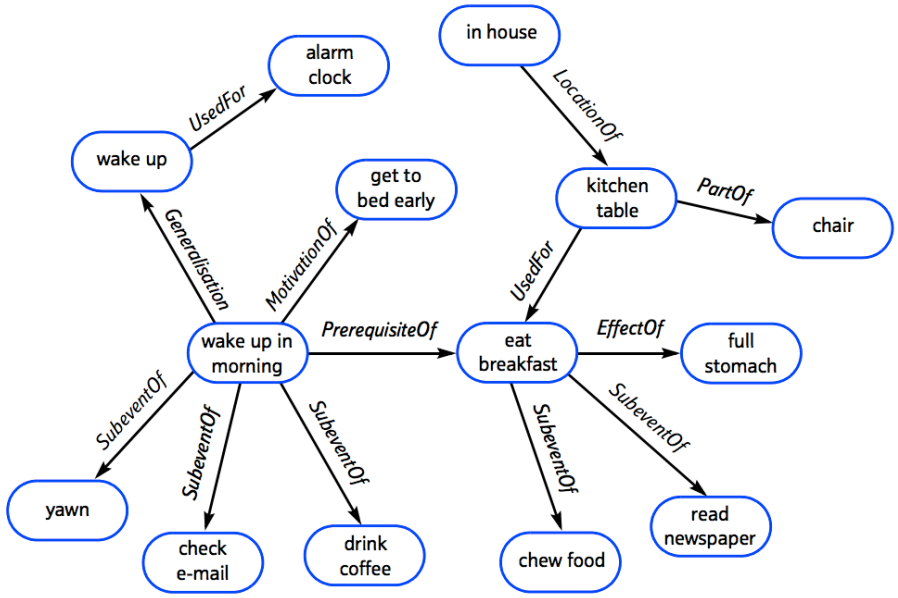

Figure 3.1 An excerpt from ConceptNet' semantic network of commonsense knowledge.

This tremendous coverage of daily life concepts and common-sense relations can be used to understand text meanings, calculating concept similarity, words meaning disambiguation, topic clustering, sentiment analysis, document classification and context-oriented inferences.

3.3 SentiWordNet

One important feature of humor is sentiment. The theory of relief shows that humor contains some kind of nervous energy or emotional energy. And the theory of

superiority points out the aggressiveness nature of humor. All these characteristics are related to human emotions and subjectiveness. Also, humor has been an outlet for sentiments in the modern society [19]. Hence, sentiment analysis is applied here to be an additional feature for text classification. SentiWordNet is a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications [20]. It defines opinion related properties for each synset in WordNet. Three scores are given to different senses, which illustrate how ‘positive', ‘negative’ and ‘neutral’ the synset is [21].

Chapter 4 Methodology

In order to create a better model to distinguish humorous and normal texts in general, this thesis attempts to leverage knowledge-based inferences to rebuild text samples and also adopts different statistical text processing techniques to reconstruct and represent text data for the machine learning model. The process to build my model contains five main steps: Data Preparation, Concepts Expansion, Text Representation, Automatic Classification, Humor Words Extraction, as shown in Figure 4.1 below.

1) Data Preparation: Filtering valid samples, clean the data, and tokenize each sample into a list of words.
2) Concept Expansion: Three well-defined knowledge bases (WordNet, ConceptNet, SentiWordNet) are introduced to expand concepts in original samples.
3) Feature Extraction/ Text Representation: Statistical text processing techniques will be used to reconstruct the texts.
4) Automatic Classification: Train a binary classifier with the training data, then use this model to predict on testing data.
5) Humor Words Extraction: Use the trained classifier to find out the words that have the most comic effect for each joke.

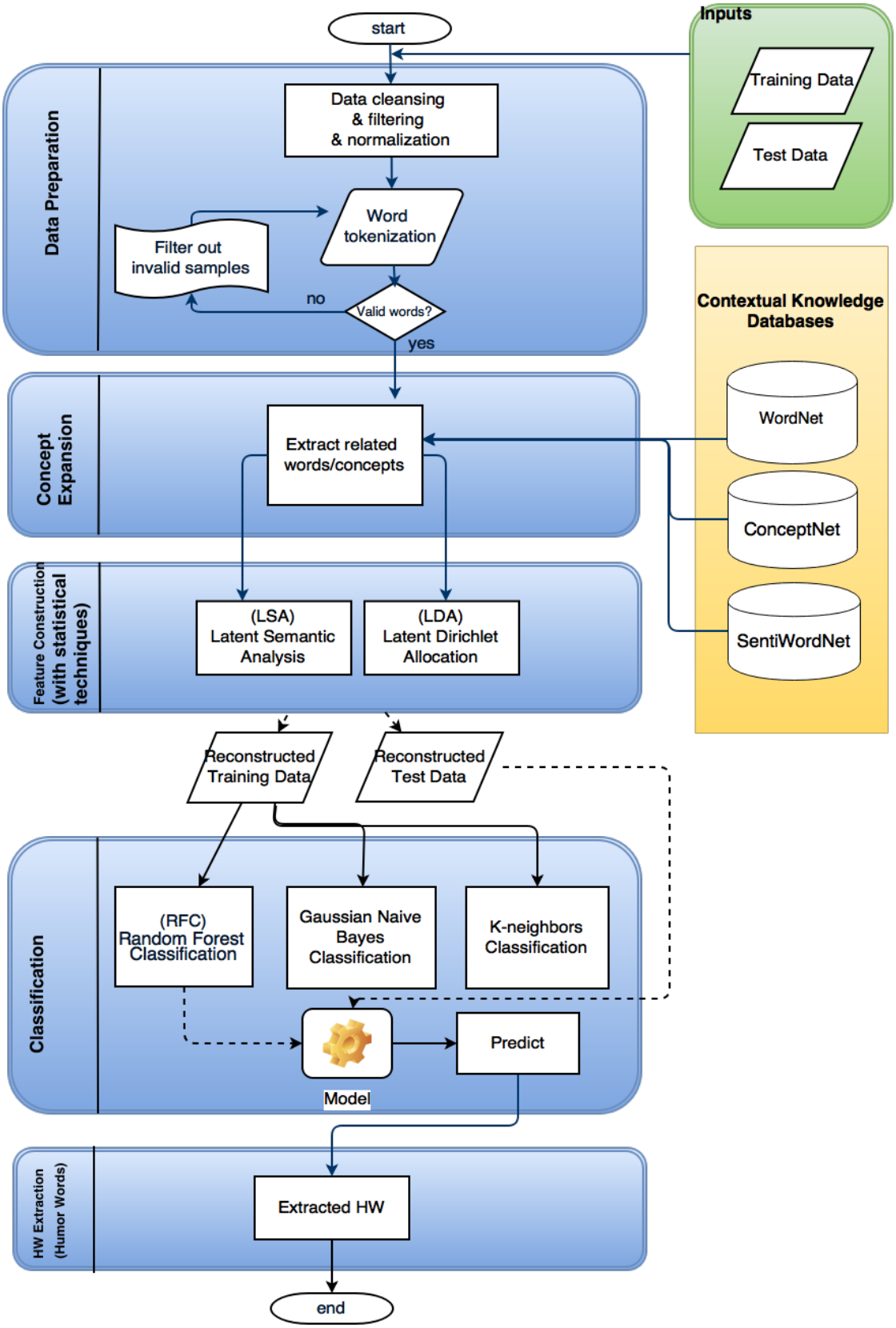

Figure 4.1 The flowchart of the methodology in the model

4.1 Data

To train the model, positive and negative data are needed. To simplify the problem, and reduce the noise in texts, humorous texts are restricted to one-liner and few sentences jokes here. The negative data is composed of data from three sources, sentences from news articles, Wikipedia articles, and proverbs. This mix of different sources and types of serious texts will help the model recognize humor texts based on essential differences between two datasets. To avoid the model classifying regarding obvious texts styles and domain difference, serious data are chopped into similar length with the humorous texts, and only those serious texts using semantically similar words with the positive dataset are chosen as negative data.

4.2 Expansion with Background Knowledge

In order to combine lexical and semantic meanings of words, unstated background knowledge in the humor detection, the method proposed in this thesis will expand the concepts in a text sample with all semantically related words, and correlated concepts. Through this concepts expansion, the original text will contain all possible elements to explain the meaning of itself and the presuppositions needed to make simple inferences. Among these related concepts and their relations linking each other, there should be a sub-map that can best explain the meaning of the original text. Three knowledge databases will be used to obtain such information. The lexical and semantic meanings of each word can be extracted from WordNet. Background information includes common senses, social and cultural background knowledge can be retrieved from ConceptNet.

After using such correlated words and concepts to expand original text, the sentiment fluctuation will be used as additional features for each sample. The sentiment fluctuation is calculated as the standard deviation of the subjectiveness and polarity of each word. These sentiment scores of each word are retrieved from SentiWordNet.

4.3 Text Representation Algorithms

After we got the expansions of words or concepts, the model needs to find a way to extract features to represent such text samples. The text representation should be efficient for

computing, consuming acceptable computing memory, but at the same time can not lose much information of the original data. Hence, three different text representation techniques are introduced. TF-IDF is the most basic and popular one but it loses a lot more information than the other two methods. Latent semantic analysis is mainly based on singular value decomposition, and is usually used to reduce the dimensions of input data. The best bag-of-words model could be latent Dirichlet allocation. It represents texts with a rather smaller vector of topics. It largely reduces dimensions of input data and also keeps the original meaning of texts through mapping words into a dictionary of topics. We combined LSA and LDA respectively with TF-IDF to extract the feature matrix.

4.3.1 TF-IDF

One basic method used to represent words in documents is TF-IDF scheme, short for term frequency–inverse document frequency. TF-IDF [22] is used to represent documents with words occurrence. It counts the occurrence of each word in a document. These words occurrence counts are normalized on a log scale to represent their occurrence frequency in the whole corpus. The output result is a document-term matrix, each row of which contains words occurrence frequency for a document. Though TF-IDF is widely adopted in IR (information retrieving), the benefits of this method is limited to lexical features since it does not take different possible formats of a single word, or various semantic meanings into consideration.

4.3.2 Latent Semantic Analysis (LSA)

Another technique that is able to consider synonymy and polysemy is Latent Semantic Analysis (LSA). It provides a method to extract and represent the contextual-usage meaning of each word in a document through applying statistical computations to the whole corpus. The main idea of this approach is to take advantage of implicit higher-order structure in the association of terms with documents ("semantic structure") to improve the detection of relevant documents on the basis of terms found in queries [23]. LSA uses single-value decomposition technics to get the low-rank approximation of a $M*N$ matrix C, given as:

$$C_k = UE_kV^T \tag{3}$$

C is a document-term matrix representation of all training data, with *M* document and *N* terms. *U, V* are orthonormal matrices, and E is a diagonal matrix containing singular values. Documents can be represented by a document-to-term matrix, then be transformed into a semantic structure *k* dominant components. These *k* components are considered as *the k* most important topics for the space of documents, and each document is represented by a distributed term vector over these *k* topics. In this way, LSA can identify a hidden semantic structure for each document. Thus, the power of LSA is that it can recognize similar items even if they share no common words. [24]

4.3.3 Latent Dirichlet Allocation (LDA)

The third type of text representation technique adopted in this thesis is Latent Dirichlet Allocation (LDA). LDA is a statistical model that allows documents being transformed into a matrix over a certain number of topics [25]. The output results can be seen as a "*probabilistic factorization of the matrix of word counts n (where* n_{dw} *is the number of times word w appears in document d) into a matrix of topic weights θ and a dictionary of topics β.* [26]*"*

4.4 Classification Algorithms

In order to get a better result, different classification methods (Gaussian Naive Bayes Classifier, K-neighbors Classifier, Random Forest Classifier) will be tested on the same data input. Combined with characteristics and limits of each method, the performances of each model will be compared at the end of next chapter.

4.5 Humor Words Extraction

After we trained a classifier to distinguish humor texts and serious texts, we will extract the humor words in each sample. In this stage, we get inspired by a previous study by Yang, D. [27]. We agree on the importance of humor anchors in humor understanding. And we use the same method to measure the humor effect of each word in a joke, but we do not define humor words as same as in this model.

Humor words are the words in a text sample that make the text rather a humorous text than a non-serious text. While deleting one word each time from the original sample text, the

classifier predicts a new probability for the sample to be humorous, namely the humor score of the text. The new prediction should be different compared to the original sample, which normally has a slight decrement than the original humor score. Then each word will be assigned respectively a decrement rate, also called the humor score of the word in this sample text. The larger decrement rate a word generates, the more contribution to the comic effect it has. Therefore, we get a list of words with their humor effects for a sample text.

From the incongruity and superiority theory of humor, a relation of contrast or contradiction exists in a joke, which means a joke should contain at least one pair of incompatible concepts. A concept can be a single word, or constitutes with multiple words, that is a compounded phrase from 2 to 3 words. Then the pair of incompatible concepts should contain at least 2 words and 6 words at most. Considering that one-liner jokes are short texts, we limit the length of humor words to a list of 2 to 5 words. Therefore, for each sample, we choose top 5 words with highest decrement rate as the humor words. If less than 5 words has positive decrement rate, we ignore those words with negative decrement rate.

Chapter 5 Experiment

The experiment mainly consists of two parts, the first of which is to build up the text classifier, using three auto-classification techniques respectively, based on the two statistical text representation techniques. The second part is to extract a list of humor words (HW) for each humor text, with the use of humor scores generated by the classifier.

5.1 Data

We use the online corpus, short text corpus for humor detection, as our training data. In this corpus, humorous jokes and one-liners are crawled from Twitter.com and other joke websites. And we choose one-liners as the positive data in our model. Serious texts from Reuters News articles, Wikipedia articles and proverbs are treated as negative data. We merge texts from all the 3 sources into one negative dataset. Originally it contains 10,076 one-liners, 10,142 Reuters headlines, 1,019 English proverbs, and 10,076 Wikipedia sentences.

Firstly, the simple cleansing should be done upfront to avoid empty strings or special characters. Secondly, for each sample in the positive dataset, we select 20 most similar negative texts through latent semantic analysis and combine them as a new negative dataset. These filtering processes will reduce the domain distance between the positive and negative datasets. Finally, we get 5,300 positive and 6,100 negative data catered for it.

Then we tokenize each sample text into a list of words. Take a joke as an example:

"I just gave my girlfriend a ring and proposed… that we break up.
Then I put the phone down on her." (1)

→ Word tokens:

gave, girlfriend, ring, proposed, break, phone. (2)

5.2 Concept Expansion

From the three knowledge databases, any words or concepts that have relations to the term appears in the sample data are extracted. Different types of relations covered by these databases can be summarized as below:

- **WordNet**: synonyms, antonyms, hypernyms, hyponyms, and meronyms;
- **ConceptNet**:
    - **K-Lines**: ConceptuallyRelatedTo, SuperThematicKLines, ThematicKLines
    - **Functional**: CapableOfReceivingAction, UsedFor
    - **Agents**: CapabelOf
    - **Things**: IsA, PartOf, PropertyOf, DefinedAs, MadeOf
    - **Events**: SubEventOf, FirstSubEventOf, PrerequisiteEventOf, LastSubEventOf
    - **Affective**: MotivationOf, DesireOf
    - **Spatial**: LocationOf
    - **Causal**: EffectOf, DesirousEffectOf
- **SentiWordNet**: polarity and subjectivity rate of words

→ Related words extracted from WordNet for the example in (1) is shown in Table 5.1.

	Gave	**Girlfriend**	**Ring**	**Propose**	**Break**	**Phone**
Synonym	feed, springiness, render, give, spring, contribute, move over, kick in, consecrate, have	girlfriend, girl, lady friend	ringing, hoop, **telephone**, echo, pack, knell, resound, ring, skirt, border	aim, purport, pop the question, propose, offer, nominate, project, **purpose**, **declare oneself**, **suggest** ...	recess, founder, soften, conk out, go, gaolbreak, pause, violate, ruin, bust ...	sound, earphone, telephone set, telephone, headphone, phone, speech sound, call, ring, earpiece ...
Antonym	starve, take	--	open chain	--	repair, make, promote, conform to, keep	--

Hypernym	show, consent, distribute, come about, evince, gauge, utter, take place, toast, parcel out ...	woman, lover, friend, adult female	sound, **jewelry**, chain, platform, telecommunicat e, touch, toroid, chemical chain, jewelry, band ...	intend, request, declare, pick out, choose, take, plan, think, select, mean ...	shot, founder, invalidate, dance, fall, terminate, come about, trespass, commute, cease ...	telecommunicat e, language unit, linguistic unit, electro-acoustic transducer, electronic equipment ...
Hyponym	distribute, spit up, pony up, bequeath, wet-nurse, slop, open up, parcel out, supply, combine	--	ding, carabiner, hem in, prize ring, heterocyclic ring, mourning ring, signet ring, rim, tintinnabulation , knell ...	offer, advance, advocate, urge, nominate, make a motion, move, submit, state, proposition ...	cut-in, splintering, capillary fracture, rest, dissociate, trespass, abatement, blab out, disunite, come out ...	handset, French telephone, speakerphone, phoneme, vowel sound, telephone receiver, telephone extension, call in, glide, desk phone ...
Meronym	--	--	canvass, canvas	--	fault line	mouthpiece, telephone receiver, receiver

Table 5.1 Lexical and semantic information retrieved from tokens in (2)

As shown in Table 5.1, word ‘ring’ can mean a subclass of ‘jewelry’, but also is related to ‘telephone’. And word ‘propose’ can mean ‘propose’ for marriage, and also it can be interpreted as ‘declare oneself’ or ‘suggest’. The setup in this joke is ‘*I just gave my girlfriend a ring and proposed*’. The words ‘girlfriend’, ‘ring’ and ‘propose’ mislead listeners understand it under a pleasurable frame that the man is planned to propose marriage to his girlfriend. Suddenly, with the word ‘that’, this joke makes it clear that the man is just proposing an idea, and the ring refers to a phone call, and just make it a sorrowful situation that their relationship is going to end. And

the distinguished contrast between these two frames of understanding is enabled by purposefully relating ‘a jewelry ring’ to ‘a phone ring’, ‘propose a marriage’ to ‘a proposal of an idea’. These hidden correlations are obvious to humans but remain hidden for computers if not including background knowledge.

→ From ConceptNet, we also retrieve all concepts that are related to each word in the sample:

Gave	Girlfriend	Ring	Propose	Break	Phone
give, contribution, donation, provide, donate, give back, given, distribute, receive, bid	girlfriend, find woman, take clothe, undress, find partner, remove clothe, take off clothe, take clothe off, find mate	ring, **wed ring**, choker, jingle, necklace, earring, bracelet, peal, clang, dowry	propose, peal, offer, **ring**, **suggest**, ask, provide, bid, beg, cope	break, bend over, raze, alcoholism, break glass, crush, drunkeness, lack money, cartilage, ruin	phone, telephone, use telephone, dial phone number, cellphone, cell phone, mobile phone, look up phone number

Table 5.2 Related concepts retrieved from ConceptNet for tokens in (2)

Table 5.2 also reveals correlations between a ‘wedding ring’ and a kind of sound. Types of relations between original words and extracted concepts are not listed here, but it is not difficult to notice their relations. From Table 5.1 and 5.2, we can have a clear picture of different semantic meanings each word may have, and the core semantics of each concept can be illustrated by a list of concepts. These correlations are hidden under the original texts but are dominant factors to cause frame shifts.

5.3 Feature Extraction

Statistical techniques are used to reconstruct the text inputs, namely, a sparse document-term matrix is generated (due to the formatting issue, the result of the matrix will be handed in

separately) to represent training data. Two different representing techniques (LSA, LDA) are used upon the same training data set, the one with better performance will be chosen. For both LSA and LDA, the model set the number of features as 400, which means features for representing texts will be reduced to dominant 400 features. The sparse matrix will be used as the input data to train the classification model.

5.4 Classification

Machine learning method is applied to distinguish humor and non-humor texts. And three different classifier techniques, Gaussian NB, K-Neighbors, RFC (Random Forest Classifier), are used to give comparisons of the results. For each method, a 6-fold cross-validation is used to evaluate the classifier. Classifiers are trained by the same training dataset, but non-CE (without concept expansion) will be taken as a baseline to evaluate whether if the new model has improvement. The ROC (receiver operating characteristic) curve are used to show the results of cross-validation. And Area Under the ROC Curve (AUC) is used to measure the predictiveness of each classifier.

5.5 Humor Words (HW) Extraction

The last step of the model is to extract humor words for each humorous texts. The humorous words are a list of words that have highest contributions to the humor effect of the sample. In this model, the classifier predicts a probability of humor of the whole text. As explained in Chapter 4.5, humor words are select by the decrement of the humor score if remove those words. The list of humor words for each text sample will be limited to 2-5 words. According to the incongruity theory, humor should contain a kind of incongruity, namely, a pair of contrast concepts or situation, each of them might have multi-words. And considering that we use short one-liners as positive data, the length of humor words should not be too long. So at least, there should be two concepts with at least one word each and at most 2 or 3 words.

The quality of these extracted humor words is compared with a manual baseline. A database of 300 humor texts is tagged manually. Those words regarded as essential for comic effect of the sample will be tagged as HW. Also, the HW extracted by the method with CE (concept expansion) and without CE will be compared in Chapter 5 Experiment.

Chapter 6 Results

From the corpus of 11,400 training data which consists of half positive and half negative data, 17,843 distinct words have been tokenized in the first step. The average word length of jokes and normal texts is 80.38. After expansions with related words or concepts, the average length of training data is 106.58. The amount of words extracted to expand the original text is 26 in average. After expansion, the total number of unique words increased from 17,843 to 28,061, as shown in Table 6.1.

	Before CE	**After CE**
Number of words	17,843	28,061
Average sample length (number of words)	80.38	106.58

Table 6.1 Average number of words in sample data before and after CE (concepts expansions)

Although the two databases are well developed, many words still can not find related words or concept. For the total 17,843 unique words tokenized in the training dataset, we found 80.9% of them exist in WordNet and SentiWordNet, and 44.7% words were defined in ConceptNet, as shown in Table 6.2 below.

	WordNet	**ConceptNet**
Number of words exists in knowledge bases	14,441 (80.9%)	7,974 (44.7%)
Number of related words/concepts extracted	23.606400269	4.46897943171

Table 6.2 Number of related words or concepts extracted from WordNet and ConceptNet

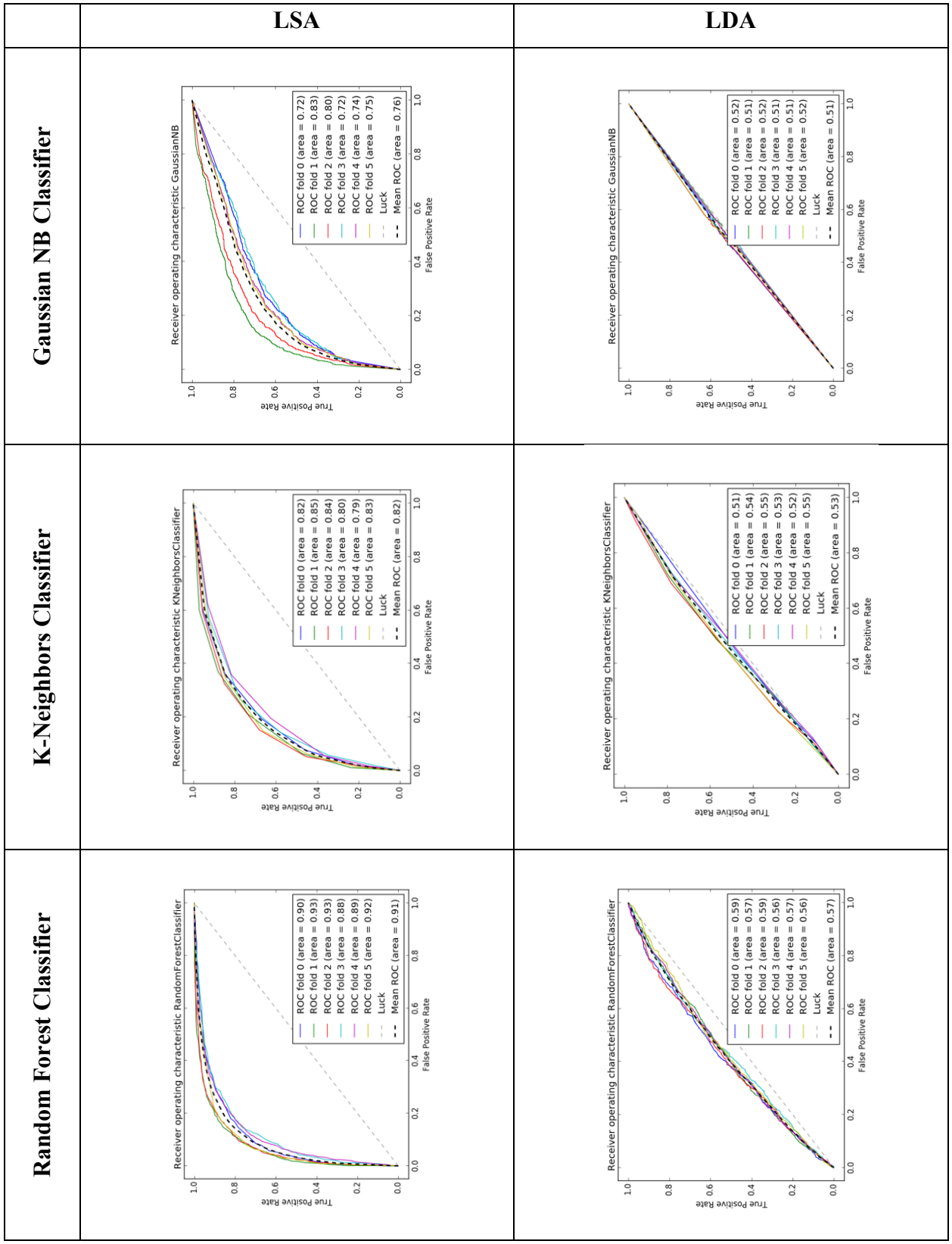

Figure 6.1 The ROC curves for cross validations on LSA and LDA WITHOUT Concept Expansion, based on three machine learning classification methods

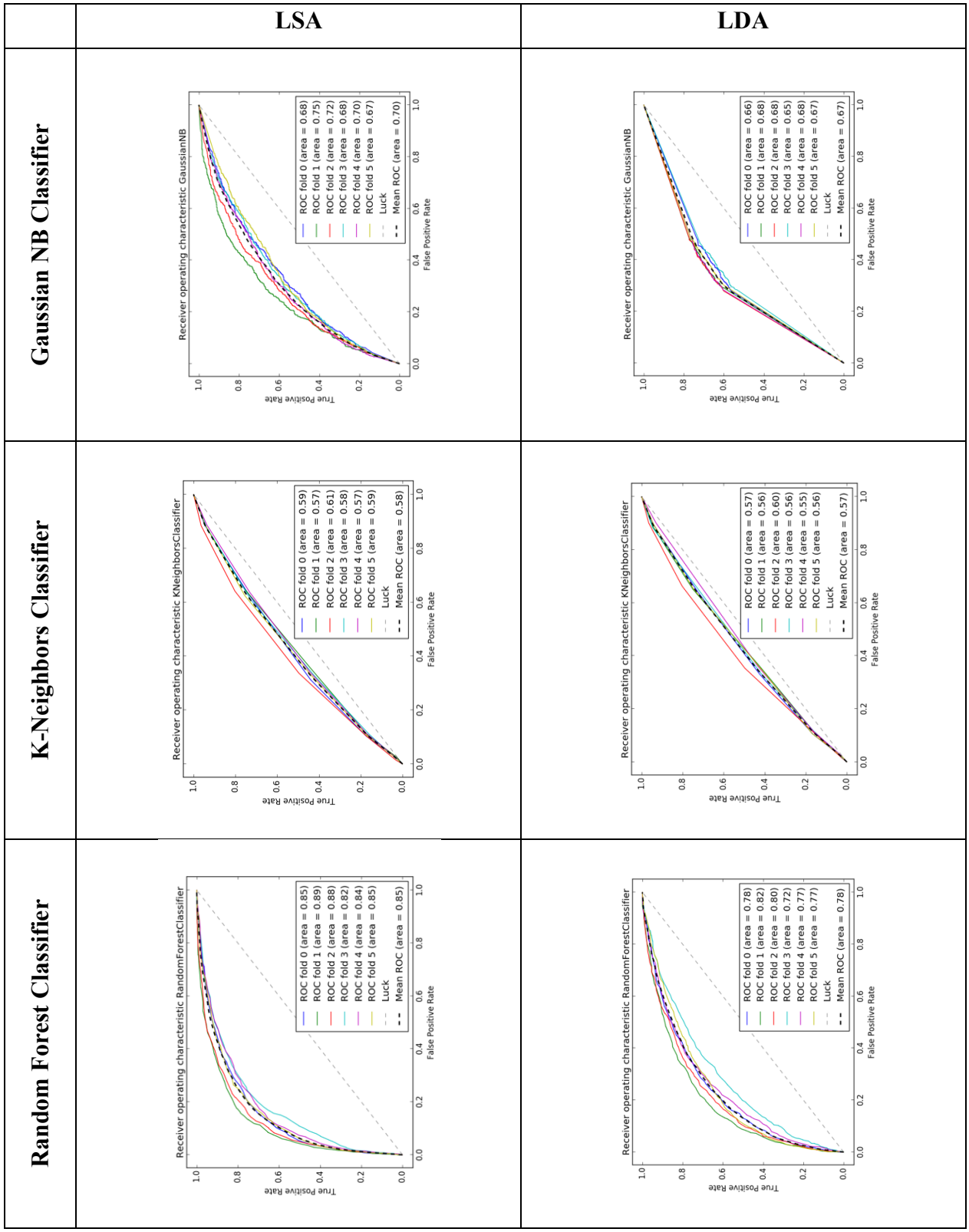

Figure 6.2 The ROC curves for cross validations on LSA and LDA WITH Concept Expansion, based on three machine learning classification methods

The ROC curves of 6-fold cross-validations for 2 text representation methods combined with 3 classification methods are presented in Table 6.1 and 6.2. Table 6.1 gives the validation of tests on the method without concept expansion, while Table 6.2 presents the results of the method with concept expansion. AUC is the area under the curve, which is used to evaluate the performance of a binary classifier. The performance is better if the area is more close to 1. And AUCs of all these models are also summed up in Table 6.3 below.

	Without CE		With CE	
	LSA	LDA	LSA	LDA
Gaussian NB Classifier	76%	51%	70%	67%
K-Neighbors Classifier	82%	53%	58%	57%
Random Forest Classifier	91%	57%	85%	78%

Table 6.3 The summary of AUC of different methods.

In Table 6.3 above, each row represents different statistical text representation methods, which are LSA and LDA. Each column represents a type of machine learning method of classification, hence there are 12 different combinations in total. For LSA technique, 'with CE' method has less AUC in all three classifier models compared to the method 'without CE', which AUC drops from 76% to 70% in Gaussian NB classifier, from 82% to 58% in K-Neighbors classifier, and from 91% to 85% in the case of RFC. However, LDA technique gains better performances, which has 16%, 11% increase respectively in Gaussian NB and RFC, and slightly increased in K-Neighbors classifier of 4%. The best AUC appears in RFC in all cases.

After training the model with different classification methods, the last step is to find out which words are of great contribution to the comic effect for each sample. Since RFC performs better than the other two classifiers, and LSA gets better and stable performance than LDA, here we choose RFC+LSA method to calculate the humor score for each sample. We use the method explained in Chapter 4.5 to locate the top 2-5 humor words for each sample. These words caused the highest decrement on humor score when removing them. In order to compare the method of 'with CE' ad 'without CE', we apply RFC+LSA on both of them.

To evaluate the quality of the humor words extracted, we choose 300 humorous texts randomly from the tainting data, and manually tag the humor words for each one. We tag 2 to 5

humor words for each text as the correct or true humor words that contribute most to the humorous effects. This manual work is taken as the correct benchmark. Then for each humorous text, a list of words with their decrements on the humor score are calculated by the model, from which the top 2-5 words with the highest decrement are tagged as humor words (HW). Additionally, a list of 2-5 words is chosen randomly for each humorous text, which is used as the baseline here to be compared with.

The results of the extracted humor words are listed in Table 6.4:

	Recall	**Precision**	**F1**
A: Random HW	18.94%	26.70%	21.01%
B: Without CE	53.49%	53.85%	51.59%
C: With CE	54.39%	53.24%	51.56%

Table 6.4 Quantitative comparisons of HWs extracted via 3 different methods

In Table 6.4, methods with and without CE both have obvious improvements on recall, precision, and f1 score than randomly chosen humor words. However, the method with CE has no improvement over the method without CE, but they just have similar results.

	Joke samples with HWs tagged manually	**HWs extracted by method without CE**	**HWs extracted by method with CE**
CE method gets better results	Do you ever get half way through **eating a horse** and you **realize** you weren't as **hungry** as you thought?	half way, thought, eating a horse	hungry, realize, thought, eating a horse,
	Thieves have broken into my house and **stolen** everything except my **soap**, **shower gel**, **towels**, and **deodorant**. The **dirty bastards**!	(Null)	dirty bastards, towels, shower gel,
	I haven't **slept** for **three days**. Because that would be **too long**.	haven't, long, days	slept, long, days
Non-CE method gets better results	I just got the morning post and one envelope read '**Photographs** Do Not **Bend'**. **Liars**! I folded it in half really **easily**!	half, folded, photographs, liars, bend	photographs, post, got, bend

	People always say that **tattoos** are a great way of preserving precious **memories**. In case you forget that anchor... or your **Mother's name**.	great way, mothers name, tattoos	preserving, forget, mothers name, precious memories
Similar	I don't know why anyone would ever want a **pocket calculator**... Its really easy to **count** how many **pockets** I have got.	want, know, pocket calculator, many pockets	pocket calculator, count, know, got

Table 6.5 Humor words (HW) extracted via 3 methods (manually, without CE, with CE)

Table 6.5 gives 6 random examples with their humor words tagged by those three methods. The jokes are presented with manually tagged HWs in boldface. The second and third column lists the HWs extracted by without CE and with CE method. We can see overlaps between those three results. The first example shows the CE method is better than non-CE method. The word 'hungry' here is obvious a keyword that makes the text humorous, but it is not identified by non-CE method. The third example shows the opposite, that the non-CE gets better results than CE method. The word 'tattoo' is of great importance to the whole text, without which would make the sentence invalid, but it is identified as a humor word in CE method.

Chapter 7 Discussions and Future Work

The results obtained by the classifier experiment using different methods prove that computational approaches with concept expansion present a viable solution to humor recognition. Though, it does not get improvement compared to the method without concept expansion. The results of humor words extraction show much better performance than the random baseline, but still no significant improvement between methods with or without concept expansion.

Despite our first intuition that background knowledge should benefit the understanding of humor, it is hard to integrate this information into feature representations. In this thesis, we use semantically related words and correlated concepts extracted from ConceptNet to expand original text samples, which is a simple and direct way to include background knowledge. And we expect that the statistical methods like LSA and LDA can handle the latent semantic structures within the expanded texts. Admittedly, the CE (concept expansion) method achieves high performance, with the AUC of 91% in the best case of LSA, and 78% in the case of LDA. Comparing to other existing models which achieved 75% precision rate in average, our classifier managed to improve the performance. However, the slight decrease in AUC in CE method compared to the non-CE method indicates that the concept expansion is reluctant to be the best way to leverage background knowledge in humor recognition. There are two main insights that we can get from our preliminary attempt to include background knowledge:

i) The concept expansion method includes more noise than the correct ones. Currently, all related words or concepts extracted from WordNet and ConceptNet are used to expand the original text, but only a very small part of them has truly contributed to the script-shifting of humor. There must be much non-relevant information introduced to the classifiers. One solution is to fine-tune these expansions, filtering out those irrelevant words or concepts. However, it is difficult for computers to decide which related background knowledge contributes to the humor effect. A researcher has developed a method to reduce dimensions of common sense knowledge [28], which can be leveraged in the future work.

ii) Text representation method can be more 'smart'. Although LSA and LDA method are a very sophisticated method to represent text inputs with statistical calculations, they

still focus the occurrence of words rather than the semantic meaning of a word, nor its correlations to other concepts. The semantic meaning of a word can be represented by a vector, which is the main idea of word embedding. The most popular word embedding technique is the *Word2vec* [29] initiated by Google. Another embedding technique is *GloVe* [30] which is designed to capture as much as possible the meaning specified by the juxtaposition of two words, such as 'woman' and 'man', 'strong' and 'stronger'. One best solution could be used in humor recognition in the future studies is *Conceptnet Numberbatch* [31], which represents words with semantic vectors supported by ConceptNet.

After we build up the classifier, we extract humor words that have the most contribution to the comic effect for each humorous text. As we explained in the theoretical background, the theory of incongruity and Koestler's theory of 'bisociation' both support that some hidden inferences are the key factors to trigger the frame-shifting in humor. To find out such creative correlations, or called analogy matching is the most interesting part of understanding humor. Our model proves humor extractions has a significant improvement compared to the baseline of random selections, but the creative correlations between them should be left to the future studies due to research limitations.

Chapter 8 Conclusions

Since the study of computational humor is still in an explorative level, we cannot expect to solve computational humor problem perfectly before the NLP (natural language processing) and AI (artificial intelligence) techniques are well developed. This thesis provides a preliminary attempt to understand verbal humor with both semantic meanings and background knowledge. Firstly, we prove the necessity to include background knowledge into computational humor. We combine semantic network retrieved from WordNet, with the factual, cultural, social knowledge extracted from ConceptNet, aiming to improve the humor recognition and discover humor words for each humor text. The results of our classifier achieved better performance than existing models, but reasons can be attributed to the statistical text representation technique LSA. When compared with the 'non-CE+LSA' method, 'CE+LSA' method does not improve the performance of the classifier. So the challenge is still how to better represent background knowledge in the humor detection, with reduced dimensions or noise. Therefore, we suggest smarter word embedding method which can better represent word meaning specified in WordNet and ConceptNet: *Conceptnet Numberbatch*. Additionally, the humor words extracted based on our classifier get much better results than the baseline of random selections. We expect further discoveries on the hidden correlations between those humor words in the future studies, which can reveal the mysterious essence of creativity in humor. These creative analogies and correlations can be used for training and education in the future.

Chapter 9 Bibliography

[1] R. Mihalcea and C. Strapparava, "Making computers laugh: Investigations in automatic humor recognition.," in *Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

[2] G. Ritchie, The Linguistic Analysis of Jokes., Routledge, 2004.

[3] J. M. Taylor, "Computational detection of humor: A dream or a nightmare? the ontological semantics approach.," in *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volu*, 2009.

[4] L. La Fave, J. Haddad and W. A. Maesen, "Superiority, enhanced self-esteem, and perceived incongruity humour theory.," 1976.

[5] S. Freud, Jokes and their relation to the unconscious, vol. No. 145, WW Norton & Company, 1989.

[6] H. Spencer, The physiology of laughter, Macmillan, 1860, pp. 1868-1874.

[7] I. Kant, The Critique of Judgment, Hackett Publishing, 1987.

[8] S. Attardo, Linguistic theories of humor, vol. 1, Walter de Gruyter, 1994.

[9] A. Koestler, The act of creation, 1964.

[10] R. Mihalcea and C. Strapparava, "Learning to laugh automatically: Computational models for humor recognition," in *Computational Intelligence*, 2006.

[11] O. Stock and C. Strapparava, "Automatic production of humorous expressions for catching the attention and remembering," 2006.

[12] A. Krikmann, "Contemporary linguistic theories of humour.," *Folklore: Electronic Journal of Folklore,* vol. 33, pp. 27-58, 2006.

[13] G. A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM,* vol. 38, no. 11, pp. 39-41, 1995.

[14] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998, ed..

[15] C. Leacock, G. A. Miller and M. Chodorow, "Using corpus statistics and WordNet relations for sense identification," *Computational Linguistics,* vol. 24.1, pp. 147-165, 1998.

[16] G. A. e. a. Miller, "Introduction to WordNet: An on-line lexical database," *International journal of lexicography,* vol. 3.4, pp. 235-244, 1990.

[17] H. Liu and P. Singh, "ConceptNet: A Practical Commonsense Reasoning Toolkit," *BT Technology Journal,* vol. 22, 2004.

[18] D. I. Diochnos, "Commonsense reasoning and large network analysis: A computational study of conceptnet 4," vol. 1304.5863, 2013.

[19] E. Oring, "Humor and the suppression of sentiment," *Humor: International Journal of Humor Research,* 1994.

[20] B. Pang and L. Lee, "Opinion mining and sentiment analysis.," *Foundations and trends in information retrieval,* Vols. 2.1-2, pp. 1-135, 2008.

[21] S. Baccianella, A. Esuli and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," *LREC,* vol. 10, 2010.

[22] G. Salton and J. Michael, "Introduction to modern information retrieval," pp. 24-51, 1983.

[23] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological review,* vol. 104.2, p. 211, 1997.

[24] S. e. a. Deerwester, "Indexing by latent semantic analysis.," *Journal of the American society for information science,* vol. 41.6, p. 391, 1990.

[25] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation.," *Journal of machine Learning research,* vol. 3, pp. 993-1022, Jan 2003.

[26] M. Hoffman, F. R. Bach and D. M. Blei, "Advances in neural information processing systems," vol. 23, pp. 856-864, 2010.

[27] D. Yang, A. Lavie, C. Dyer and E. Hovy, "Humor recognition and humor anchor extraction," in *2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[28] R. Speer, C. Havasi and H. Lieberman, "AnalogySpace: Reducing the Dimensionality of Common Sense Knowledge.," *AAAI,* vol. 8, pp. 548-553, July 2008.

[29] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space.," *CoRR,* 2013.

[30] J. Pennington, R. Socher and C. D. Manning, "Glove: Global Vectors for Word

Representation," *EMNLP,* vol. 14, pp. 1532-43, October 2014.

[31] R. Speer and J. Chin, "An Ensemble Method to Produce High-Quality Word Embeddings.," 2016.

[32] T. &. B. H. C. Flamson, "The encryption theory of humor: A knowledge-based mechanism of honest signaling.," *Journal of Evolutionary Psychology,* vol. 6(4), pp. 261-281, 2008.

[33] K. F. T. V. Geert Brône, Cognitive Linguistics and Humor Research, Walter de Gruyter GmbH & Co KG, p. 154.

[34] J. C. Meyer, "Humor as a Double-Edged Sword: Four Functions of Humor in Communication," in *Communication Theory, doi:10.1111/j.1468-2885.2000.tb00194.x*, 2000.

[35] T. Fawcett, "An introduction to ROC analysis.," vol. 27, no. 8, pp. 861-874, 2006.

[36] H. Chung and X. Zhao, "Humour effect on memory and attitude: moderating role of product involvement.," *International Journal of Advertising, 22(1),* pp. 117-144, 2003.