A computational approach to detection of conceptual incongruity in text and its applications

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Amogh Mahapatra

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

May, 2013

Acknowledgements

This dissertation is a result of years of thinking, about what leads to the presence of conceptual incongruity in text and why do human subjects find the presence of conceptual incongruity interesting at all! As at the heart of it, this work is about fun, humor, interestingness and engagement, first of all, I would like to thank all my best buddies over the years (Ron, Cheeku, Atul, Ishaan, Subhra, Surya, Abhijit and many others) whose uninhibited sense of humor provided me with the deluge of data points this work derives its inspiration from.

I would like to express my deepest gratitude towards my advisor Dr. Jaideep Srivastava. I had joined Dr. Srivastava's lab at a point of time in my life where I was heavily confused with way too many questions like "what do I really like doing, why do I get so easily bored etc.". All these grinding years in graduate school other than being the intellectual roller-coaster ride that they truly are, also happened to be my first steps towards self-discovery. Thank you Sir, not just for guiding this thesis and ensuring my well-being during my stay at this university but also for implicitly helping me find these necessary life-answers.

Though in no direct involvement with this thesis, I would like to express my deepest gratitude towards Dr. Carl Sturtivant. Thank you Sir, intellectual involvement with you has blessed me with some of my most cherished moments in graduate school. In you, I found both a teacher and an intellectual companion who understood me personally. The mathematical sparks that you have generated in my head should last me a lifetime of creativity.

A special thanks to all my lab-mates at the University Of Minnesota. First, for being fellow sufferers of my dry humor on many occasions. Second, for being energetic collaborators who have taught and inspired me in many ways during the lifespan of

various projects.

Finally, I thank my parents, Lisa, Tara and the rest of my family for being supportive in every possible way over the last few years.

Dedication

To Ila, for being the first one who listened to these ideas

Abstract

Given a text corpus, which particular pieces of text would be most interesting to human subjects? Is it possible to quantify a subjective idea like "interestingness" in the domain of text data and build algorithms to detect it? This thesis provides a computational investigation of the above questions.

The incongruity theory of curiosity postulates that humans deem the optimal presence of conceptual incongruity in their observations as "interesting" . Based on this idea, we propose that, incongruity of a textual topic can be detected by measuring two things, the statistical rarity of the topic in the given corpus and the contextual deviance of the words in the given topic measured from a universal distribution of word co-usage in the society. Based on this concept, we present algorithms to quantify conceptual incongruity and detect different kinds of interestingness (at a sample level) in text data.

We first present an algorithm to detect incongruous topics in large scale text corpora. We could detect incongruous emails from the Enron corpus, deviant paper abstracts and incongruous blog posts. We then extend this algorithm to present a computational model of humor, which was used to detect funny videos from YouTube using a given video's tag-set. We then provide different flavors of this algorithm to detect choice of words considered creative by humans and most popular set of media objects in social networks. We then show the information theoretic motivations behind our proposal and demonstrate that it maps directly to some basic principles. Finally we investigate, if it's the mere presence of incongruity or its eventual resolution which is the real cause of interest stimulation. We present an algorithm to carry out this test and report some interesting results. The generalizability of our results in finding interestingness across these different domains using algorithms derived using intrinsic human motivations, opens up exciting new avenues in the field of knowledge discovery.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Let's say Alice goes out to watch a movie A. Based on the reviews, trailers and her experience with similar movies she expects the movie to be a certain kind of romantic comedy. The movie turns out to be exactly what she had expected , her expectations aren't belied by much and hence, she doesn't find the experience very surprising. Now, Alice watches another movie B. Based on the reviews, trailers and her experience with similar movies she expects the movie to be a certain kind of political comedy. But it turns out to be a suspense thriller. Her expectations are belied heavily in a playful manner, which is why she finds the experience both surprising and "interesting". There exist a class of stimuli which invoke a set of endearing responses like laughter, applause, claps, wows and ahas etc. in human subjects. The root cause behind all such responses are mostly attributed [45] to the presence of an interest instigating stimuli. The incongruity theory, which happens to be one of the most well-accepted explanations provided by the behavioral psychology literature, postulates that interest instigation is the result of an innate human tendency of making sense of the world which is fueled by the failure of subjective expectations of prior beliefs held by the agent. E.g. the prior beliefs of Alice were belied playfully during her second experience and the movie B was deemed more interesting by her. The need for sense-making seems to be aversive at extreme values which implies interest is fueled by the presence of "optimal incongruity" in observations. Stated otherwise, the presence of too much incongruity in an observation leads to the agent perceiving the given observation as noise. E.g. If Alice who happens to be an entertainment movie fan, watches a slow documentary, she is not

very likely to deem the experience interesting. On the contrary, the presence of too little incongruity in an observation doesn't instigate an agent towards sense making hence it eventually leads to boredom. E.g. If Alice who loves romantic comedies in general, watches a movie which is very close to her viewing habits, she is likely to perceive the movie as boring. Can this qualitative idea that need for sense making fueled by the presence of optimal incongruity causes interest stimulation be modeled quantitatively in the domain of text data? This dissertation provides a computational investigation to this question.

Based on above, the presence of conceptual incongruity in a piece of text should be highly indicative of the interestingness of the given piece. This proposal seems both intuitively correct and has been backed by quite a few qualitative studies done in the past. [43] observe that humans consider a dissimilar choice of words pointing towards a central meaning as a creative set. Likewise, researchers in the area of humor, [35], have reported strong correlation between conceptual incongruity and the presence of humor. This spans across most kinds of humorous stimuli, like puns, parody, spoof, sarcasm etc. Hence, the problem reduces to being able to detect the presence of conceptual incongruity in a piece of text. How is conceptual incongruity defined in the domain of text? Given a text corpus, is it the most rarely co-occurring pieces of text which are incongruous? E.g. Words used in emails about a fantasy football league co-occur less with words used in other emails exchanged in a corporate setting. Hence, their statistical rarity makes them interesting. Is it the inherent contextual deviance of given set of words which makes them interesting ? In the set of words "baby, cute, video, blood", the acontextual use of the last word adds some incongruity to the given set. We take both the above factors into account and propose that conceptual incongruity of a textual topic can be detected by measuring the simultaneous presence of two things. The statistical rarity of the topic in the given corpus and the contextual deviance of the words in the given topic measured from a universal distribution of word co-usage in the society. We do provide the necessary mathematical arguments behind our proposal. Based on this concept, we present algorithms to detect interesting emails from Enron corpus, funny videos from YouTube using their tag-set, choice of words considered creative by humans, most popular set of media objects etc. as some of the likely applications of our proposal. The rest of this section provides a road-map of this document and a brief

summary of the rest of the chapters.

Chapter 2 describes an algorithm to detect incongruous topics in large scale text corpora. Traditional, anomaly detection techniques in text data have mostly relied on statistical frequencies of word co-occurrences in a given text corpora, with little reference to their semantic content. This technical constraint restricts the ability of such systems to identify the usage of words in an idiomatic/colloquial/local sense within domain-specific datasets. We propose a novel two-stage algorithm which considers both divergence from the statistical patterns seen in particular data-sets and divergence seen from more general semantic expectations associated with words in the society. Computational experiments on traditional data-sets like Enron emails and Daily-Kos blogs shows that this algorithm does lead to better detection of deviant emails and blogs and reduces false positives significantly. This algorithm is also good tool to characterize conceptual incongruity in text data, and some of its applications are outlined in subsequent chapters.

Chapter 3 provides a general algorithmic framework for the quantification of different kinds of interestingness[1] in the semantic space. First we provided the mathematical arguments behind our proposal and showed that our approach mapped directly to the idea of maximizing information gain in text. Second, we present the different measures used by our algorithm under different generative conditions. Third, empirical results were reported from different kinds of data-sets measuring different kinds of interestingness. We could detect, categories of words considered creative by human subjects, most liked set of media objects on a social network and emails considered interesting in an email corpus. It was found that incongruity is a necessary but not a sufficient condition for characterizing interestingness and also interestingness seems to be monotonic, it increases with increase in the amount of incongruity except at the boundaries. Humans associate too much incongruity with noise and too little incongruity with boredom, otherwise more the incongruity higher the interestingness.

Humor is also a kind of interestingness. Chapter 4 presents a computational model of humor.The incongruity theory of humor postulates that humor results due the failure

[1] It should be noted that our quantification of interestingness based on incongruity theory is measuring what a sample of diverse set of people would most likely find interesting. Currently, we are not measuring a personal cognitive notion of interestingness. For example, there might be a few outlying human subjects who don't like any incongruity in their observations at all.

of subjective expectations of an agent in a playful manner which in turn arises due to the presence of conceptual incongruity. We propose a computational model of humor which is characterized using two factors called (Diversity, Anomalousness) whose simultaneous presence leads to the presence of conceptual incongruity and hence leads to the presence of humor. For this purpose, we developed a methodology for extracting semantic distance from tags associated with YouTube videos manually identified as humorous or not by their existing community of users. We found that a novel quantification of conceptual incongruity, operationalized via this technique, shows strong correlation with humor and proves to be a necessary but not sufficient condition for the existence of humor. It was further observed that conceptual incongruity is also strongly positively correlated with and predictive of video popularity across both humorous and non-humorous genres of videos.

Chapter 5 investigates the debate between between incongruity and incongruity resolution in a data-driven fashion. As mentioned earlier "Incongruity Theory" is currently the most well-accepted explanation as the underlying mechanism behind humor. This theory postulates that humor is caused due to the playful violation of the subjective expectations of an agent; caused by the presence of incongruous stimuli in the agent's observations. But whether humor is caused merely due to the presence of incongruous stimuli or whether it is caused by the resolution of an apparent incongruity by the sudden realization of another competitive explanation still remains an open question. We attempt to address this question in a data-driven fashion. We develop an algorithm inspired from information-theory to investigate this question in the semantic space and report results on three different data-sets. We observe that simpler stimuli like slapstick jokes, puns, one-liners etc. are better explained by the mere presence of incongruity, whereas more complex stimuli; like high-quality jokes, jokes with hidden meanings etc. contain greater amount of conceptual dissonance and hence, are better characterized using the idea of resolution of incongruity. Our methodology opens up exciting new avenues of research in the areas of humor and engagement and our results could find potential applications in areas like marketing and advertising.

The individual contributions are listed at the end of every chapter, and the overall contributions are detailed and discussed in the final chapter. The final chapter also contains a useful discussion on the shortcomings of our approach, the proposed future

work and likely application of our approach in other domains of knowledge discovery.

As a reading guide, readers solely interested in the theoretical contributions of this work should directly skip to chapter 3. Readers solely interested in the algorithmic contributions of this work should read the second sections of each chapter. Readers who solely care for the work done on humor should read chapter 4 and 5. Readers solely interested in getting a quick snapshot of the results, should look at the last section of chapter 3. Care has been taken to keep the chapters self-contained by providing reference to the appropriate chapters/sections as required. Some of the material, like explanation of the properties of semantic distances etc. is repeated in a couple of places to ensure readability of the particular chapter.

Chapter 2

A Contextual Anomaly Detection Algorithm

2.1 Background

The ability to detect anomalies in streams of text data finds broad applicability in a number of web applications. For example, it can be used to detect the occurrence of important events from Twitter streams, the occurrence of fraudulent communication on email networks and even fault descriptions in maintenance logs. An important emerging application of fault detection obtains in the domain of organizational security, where it is critical for the integrity of the organization that sensitive information not be leaked. In all these cases, anomalies are detected through statistical comparisons with typical data and subsequently evaluated by human analysts for relevance and accuracy. The necessity of this latter step arises from the insight that, for the detection of interesting events in media streams, an aggressive approach, leading to significantly many false positives, is often more useful than a conservative approach that promotes false negatives [4, sections:3.1-3.2]. This observation holds the most strongly for security applications, where false negatives, while rare, can carry enormously higher costs as compared to the cost of weeding out irrelevant false positives.

It is therefore desirable, in many application settings, to be conservative in rejecting data samples as anomalies and then post-processing detected samples with human analyst input to identify real anomalies. However, the cognitive limitations of human

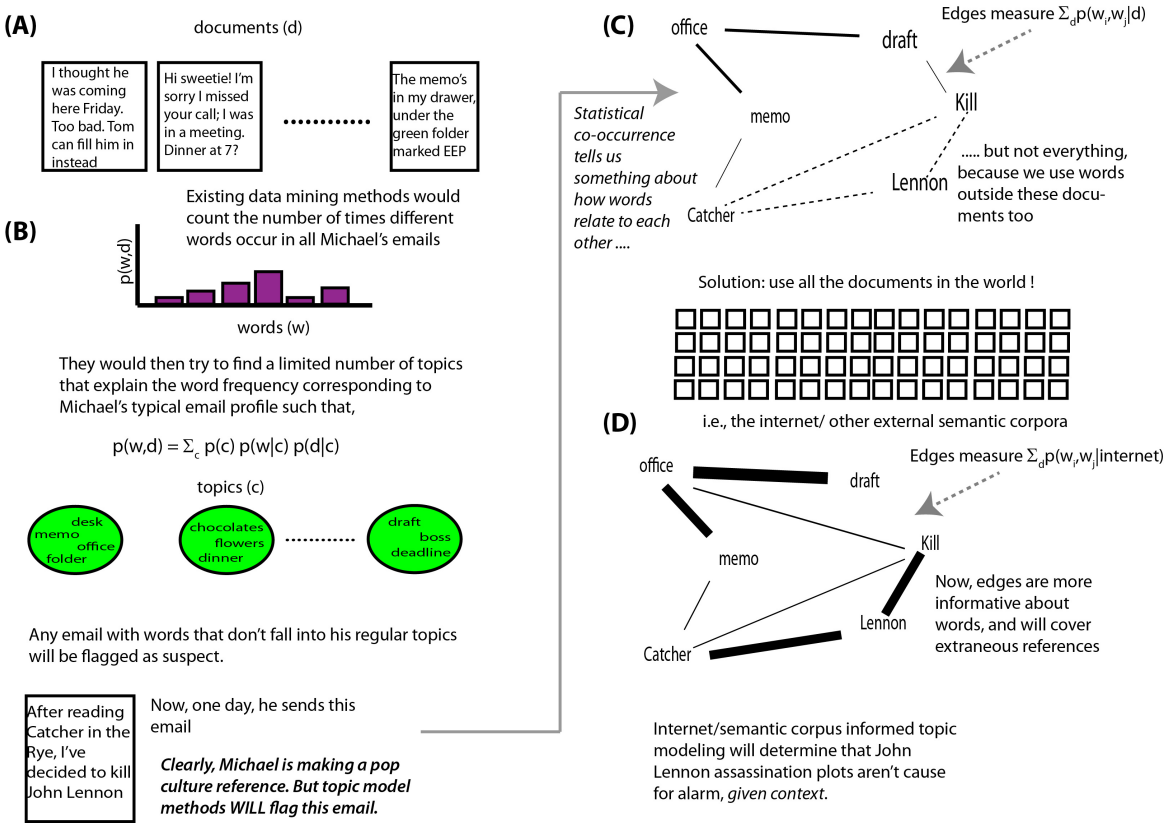

Figure 2.1: This figure shows a notional application scenario of our algorithm highlighting its benefits over the state of the art techniques. A shows a set of emails populating Michael's inbox. B (above) shows how the topics models would extract latent themes, based solely on statistical frequencies of word co-occurrences. B (below) shows that a pop culture reference made by him is flagged anomalous due to the statistical rarity of the used words in the given corpus. C shows the relational-language graph used by topic models where edges only capture the statistical co-occurrence between the words in the corpus. D shows the contribution made by us where the edges now also capture the semantic distance between the words tuned over a very large set of external documents and are hence more contextually informed, which eventually flags Michael's email as harmless.

analysts restrict the scalability of any such proposals. It is, therefore, useful to consider the possibility of augmenting data-driven statistical anomaly detection with a post-processing filtering step that automatically determines the relevance of detected anomalies using alternative criteria. For text data, an intuitive alternative criterion for such an evaluation immediately presents itself. Current statistical text analysis techniques tokenize text data into generic categorical objects. Hence, the semantic content of text data is ignored. Reintroducing semantic information in text data analysis creates the possibility of evaluating the relevance of anomalies detected using background semantic context from a much larger corpus of data.

Here, we show how information about the semantic content of text data allows us to identify occurrences of words and topics that are statistically infrequent, but contextually plausible for the monitored system, and hence, not anomalous. We do this by bootstrapping a novel context-detection algorithm that operates on an external corpus of general semantic relationships (Wordnet and Internet) to an LDA-based text clustering algorithm for anomaly detection that operates on particular datasets of interest. Clusters of co-occurring words (topics) that are rated highly anomalous by both modules are considered truly anomalous, while those that are flagged by statistical analysis, but considered typical through evaluating semantic relatedness are considered explainable data artifacts.

In order to further display the effectiveness of using contextual similarity measures to solve the problem of anomaly detection in textual data, we conducted a set of experiments on tags generated by users on popular social multimedia network, YouTube. Each video contains about 4-12 tags. We empirically demonstrate that the use of contextual measures, could help us do anomaly detection at word-level as well, i.e. we can detect anomalous topics and go even further and also detect anomalous words in a given topic.

The rest of this chapter is organized as follows: in Section 2.2, we discuss the state-of-the-art in current anomaly detection schemes for text data and in systems for evaluating semantic context. We describe the rationale behind our approach, the technical details of its individual components and its algorithmic implementation in Section 2.3. A description of our empirical evaluation strategy and results follows in Section 2.4, following which we conclude with a short discussion of the implications of our results in Section 2.5.

2.2 Related Work

Anomaly detection techniques in the textual domain aim at uncovering novel and interesting topics and words in a document corpus. The data is usually represented in the format of a document to word co-occurrence matrix which makes it very sparse and high dimensional. Hence, one of the major challenges that most learning techniques have to deal with while working on textual data is being able to deal with the curse of dimensionality. Manevitz et. al. [9] have used neural networks to classify positive documents from negative ones. They use a feed forward neural network which is first trained on a set of positive examples (labeled) and then in the test phase the network filters out the positive documents from the negative ones. In another work which also uses the principle of supervised learning, Manevitz et. al. [10] have used one class SVMs to classify outliers from the normal set of documents. They show that it works better than techniques based on naive bayes, nearest neighbor algorithms and performs just as good as neural network based techniques. The above approach might not be very useful in an unsupervised setting, whereas our approach can work in both supervised and unsupervised settings. This problem has been studied in unsupervised settings as well, Srivastava et. al. [18] have used various clustering techniques like k-means, Sammons mapping, Von Mises Fisher Clustering and Spectral Clustering to cluster and visualize textual data. Sammons mapping gives out the best set of well separated clusters followed by Von Mises Fisher Clustering, Spectral clustering and K-means. Their technique requires manual examination of the textual clusters and doesn't talk about a method of ordering the clusters. Agovic et. al. [1] have used topic models to detect anomalous topics from aviation logs. Guthrie et.al. [6] have done anomaly detection in texts under unsupervised conditions by trying to detect the deviation in author, genre, topic and tone. They define about 200 stylistic features to characterize various kinds of writing and then use statistical measures to find deviations. All the techniques we describe above rely entirely on the content of the dataset being evaluated to make their predictions. To the best of our knowledge, ours is the first attempt at finding anomalies in text logs which makes use of external contextual information. Since the topics detected in statistical content analysis strip lexical sequence information away from text samples, our efforts to reintroduce context information must look to techniques of

automatic meaning or semantic sense determination.

Natural language processing techniques have focused deeply on being able to identify the lexical structure of text samples. However, research into computationally identifying the semantic relationships between words automatically is far sparser, since the problem is much harder. In particular, while lexical structure can be inferred purely statistically given a dictionary of known senses of word meanings in particular sequences, such a task becomes almost quixotically difficult when it comes to trying to identify semantic relations between individual words. However, a significant number of researchers have tried to define measures of similarity for words based on, e.g. information-theoretic [20, 27] and corpus overlap criteria [7, 11]. Cilibrasi and Vitanyi have shown [5], very promisingly, that it is possible to infer word similarities even from an uncurated corpus of semantic data, viz. the Web accessed through Google search queries. This observation has been subsequently developed and refined by [3] and presents possibilities for significant improvements to current corpus-based methods of meaning construction. Semantic similarity measures have been used in the past to accomplish several semantic tasks like ranking of tags within an image [13], approximate word ontology matching [21] etc. and thus, hold the promise of further simplifying cognitively challenging tasks in the future.

2.3 Methods

We would like to detect the occurrence of abnormal/deviant topics and themes in large scale textual logs (like emails, blogs etc.) which could help us in inferring anomalous shifts in behavioral patterns. Our system should then use two inputs - a text log whose content is of immediate interest and an external corpus of text and semantic relationships to derive contextual information. The output would be the set of final topics considered anomalous by our system. It is possible to subsequently evaluate the topics discovered by our data set manually and flag a certain number of topics as anomalous. The efficacy of our approach is measured as the ratio of number of anomalies correctly detected by the system to the number of anomalies flagged through manual inspection.

Consider all words to live on a large semantic graph, where nodes represent word labels and edges are continuous quantities that represent degrees of semantic relatedness

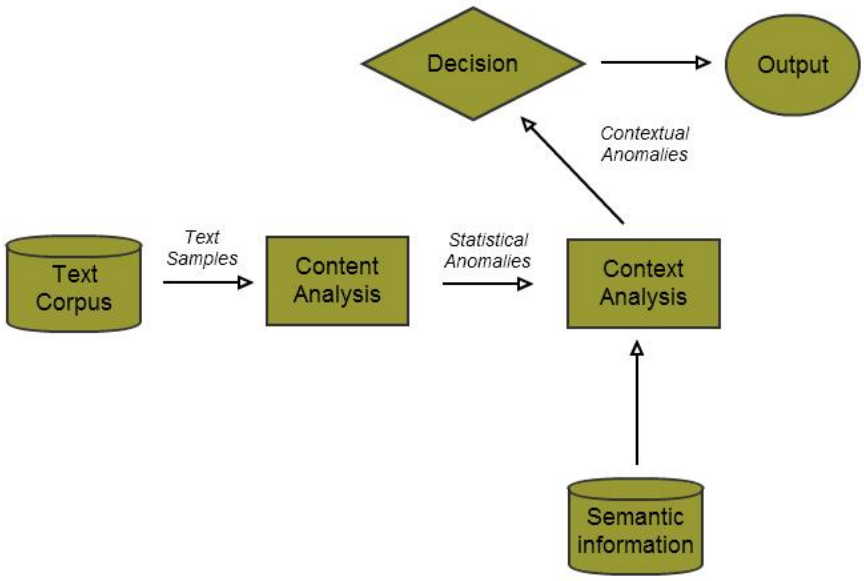

Figure 2.2: Flowchart of the decision engine

to other words. Some subset of these words populates the documents which we evaluate in clustering-based text analysis schemes. A document can then be described by an indicator vector that selects a subset of the word labels to populate an individual document. Traditional document clustering creates clusters of typical word co-occurrences across multiple documents, which are called topics. However, such an approach throws away all information embedded in the semantic network structure. The goal of a semantically sensitive anomaly detection technique is to integrate information potentially available in the edges of the semantic graph to improve predictive performance.

Rather than attempt to include semantic information within a traditional clustering scheme by flattening the relatedness graph into additional components of the data feature vector, we attempt to introduce context as a post-processing filter for regular topic modeling-based anomaly detection techniques. By doing so, we simplify the construction of our algorithm and also ensure that the results from both unfiltered and filtered versions of the algorithm are clearly visible, so that the relative value added by introducing semantic information is clearly visible. Since our approach involves combining two separate modes of analyzing data - one content-driven and one context-driven

- we now describe both these modalities in turn and subsequently, our technique for combining them.

Algorithm 1 The Sematic Anomaly Detection Algorithm

Input: Documents D, number of topics n, partition parameter m, anomaly threshold k

Output: Set of anomalous topics

B=ComputeBagOfWords(D) {Computes the statistical frequencies of word co-occurrences in the corpora}

T=LDA(B,n) {Cluster bag-of-words into n topics}

T_1=Rank(T) {Rank topics based on document co-occurrences}

$Test$=Partition(T_1,m) {Partition into typical and test topic sets}

FindContext($Test$) {Find context from sematic measures}

R=Decision($Test$,k) {Pick k lowest context score topics}

Output R

Figure 2.3: The Semantic Anomaly Detection Algorithm

2.3.1 Statistical Content Analysis

Statistical topic models based on Latent Dirichlet Allocation (LDA) (Blei .et al. 2003 [2]) have been applied to analyze the content of textual logs in order to identify the various topics/concepts that exist in them at different levels of thematic abstraction. LDA represents each topic as a distribution over words and allows for mixed memberships of documents to several topics. LDA is very effective in identifying the various thematically coherent topics that exist in document collections and producing a soft clustering solution of the documents. In addition, LDA's iterative nature allows it to scale to document collections containing millions of documents (Newman et. al. 2007 [14]).

LDAs generative model is illustrated in Figure 2.4. Topics are assumed to be distributions over words. A Dirichlet distribution with parameter α is assumed as the prior over all documents. From Dirichlet(α) each document is drawn as a multinomial distribution θ. The dimensionality of this distribution is determined by the number of topics. For each word within a document we draw a topic from this multinomial distribution.

We subsequently go to the corresponding topic distribution and draw a word from it. Ultimately a document is assumed to be a mixture of topic distributions. In training the model the objective is to learn the topic distributions along with the underlying multinomial distributions θ, representing the documents. In LDA the resulting topics can be represented by the most likely words occurring in them. For example, Figure 2.6 shows some of the most important terms identified from the top four topics from the Enron email dataset, which can be used as a summary of the identified topic.

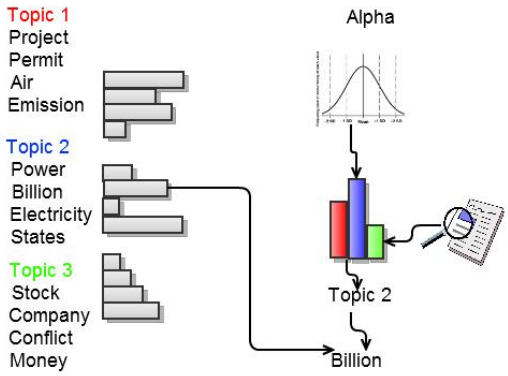

Figure 2.4: LDA's generative model illustrated with 3 topics and four words

The documents in the text logs were converted to the bag of words format after tokenization, noise removal and removal of stop words. Then the following two steps were carried out:

1. Clustering : The text log was divided into k-Clusters using the LDA model. The values of k and other parameters like θ, α etc. were decided based on the size of the data set and our understanding of the nature of the data set. The top 10 most likely words were extracted as a representative summary of each topic.

2. Ranking : The topics were ordered based on their co-usage in documents. LDA assumes every document to have been created by a mixture of topics distributions. We obtain an ordering of topics based on the assumption that topics that appear together are similar to each other should have a low relative difference in their rankings. First, the topic distributions for each topic was calculated. Then, for each pair of topics i.e. (P, Q), the symmetrized KL divergence between topic

distributions P and Q was calculated. Equation 2.1 shows the divergence measure between the probability distributions P and Q. Equation 2.2 shows the symmetrized KL divergence measure (henceforth SD) which has the properties of being symmetric and non-negative (equation 2.3 and equation 2.4). The symmetrized KL divergence was subjected to dimensionality reduction and the first dimension was used to rank the topics. [25] We made use of the "Topic Modeling Toolbox" to conduct out experiments.

$$D(P,Q) = \sum_{i=1}^{n} P(i) ln \frac{P(i)}{Q(i)} \tag{2.1}$$

$$SD(P,Q) = D(P,Q) + D(P,Q) \tag{2.2}$$

$$SD(P,Q) >= 0 \tag{2.3}$$

$$SD(P,Q) = SD(Q,P) \tag{2.4}$$

2.3.2 Computing contexts

Determination of semantic context is, in general, a difficult question. For example, suppose we are analyzing a user's behavior on a social network over a period of time using his activity logs. Just the measurement of deviation of a user's behavior from his past behavioral patterns might not be indicative enough to make accurate predictions about his anomalous behavioral patterns. The context regarding his behavioral changes in this case could be derived from his peers' activity logs, local demographic information etc. which could help in taking a more informed decision. Additionally, the trade-off between adding extra contextual information to improve predictive power and the resultant computational complexity is also important to examine.

As we are dealing with large scale text logs, we decided to derive the contextual information from two well-known external semantic corpora namely Wordnet [23] and the Internet. We describe the technical details of our procedure below:

Computing Semantic Similarity Using Wordnet

Wordnet is a lexical database for English language which has been widely used so far in many natural language processing and text mining applications. Wordnet puts similar English words together into sets called synsets. Wordnet has organized nouns and

verbs into hierarchies of is-a relationships (example dog-is-a-animal). It provides additional non-hierarchical relations like has-part, is-made-of etc. for adverbs and adjectives. Hence, two words could be considered similar if they are derived from a common set of ancestors or share similar horizontal relationships. As a result, various kinds of similarity and relatedness measures like path, wup, lch, gloss, vector etc. [15] have been defined which quantify the semantic similarity/relatedness of two words present in the database.

Pederson et al. [15] have enumerated 9 measures to compute the similarity/relatedness between two words in Wordnet. These measures can be broadly classified into two categories. The first ones are those which quantify the similarity/relatedness between concepts based on different kinds of network distances between the concepts in the network. The second ones are those which use occurrence overlap between glossary texts associated with the concepts to measure their similarity. We have used one of each of these measures, to account for both classes of distances.

1. Path Measure : Path is a network distance measure. It is simply the inverse of the number of nodes that come along the shortest path between the synsets containing the two words. It is a measure between$0-1$.The path length is 1 if the two words belong to the same synset.

2. Gloss Vectors : Gloss is a relatedness measure that uses statistical co-occurrence to compute similarity. Every set of words in Wordnet is accompanied by some glossary text. This measure uses the gloss text to find similarity between words. Relatedness between concepts is measured by finding the cosine between a pair of gloss vectors. It is also a measure between $0-1$. If two concept are exactly the same then the measure is 1 (angle between them is 0° and $Cos(0)$ is 1).

We take the average of those two measures for any given word pair, reflecting the fact that we have accounted for both semantic relatedness (from the first measure) and real-world statistical co-occurrence of two words (from the second measure) in our decision making process. We have used Wordnet 2.05 in our experiments.

Computing Semantic Similarity Using the Internet

In spite of the apparent subjective accuracy of the Wordnet corpus, the static nature of this corpus lacks the rich semantic connectivity that characterizes natural language interactions in social settings. Words, phrases and concepts change meanings during the course of their everyday usage and neologisms (e.g. lol, rotfl) and colloquialisms are added everyday to the human lexicon. Therefore, for its ability to offer higher conceptual coverage than any known semantic network or word ontology, we base our semantic meaning extraction on the entire World Wide Web by using the Normalized Google Distance for this purpose, as we describe below.

Given any two concepts, (a, b), the Normalized Google Distance (NGD, henceforth) between them is given by equation 5.1. $f(a)$ and $f(b)$ denote the number of pages containing a and b as returned by Google. $f(a, b)$ is the number of pages containing both a and b and N is the total number of pages indexed by Google.

$$NGD(a, b) = \frac{max(\log f(a), \log f(b)) - \log(f(a, b))}{\log N - min(\log f(a), \log f(b))} \tag{2.5}$$

The range of this measure is (0,∞). The value 0 would indicate that the two concepts are exactly the same and ∞ would indicate they are completely unrelated. Normalized Google Distance is a non-metric and hence triangle inequality (equation 5.2) doesn't always hold. If a, b, c are any three random words, then:

$$NGD(a, c) \not<= NGD(a, b) + NGD(b, c) \tag{2.6}$$

Normalized google distance is symmetric.

$$NGD(a, b) = NGD(b, a) \tag{2.7}$$

Equation 5.1 implicitly assumes that the ratio of the total number of pages returned by Google for a given term divided by the total number of pages indexed by Google is equal to the probability of that search term as actually used in the society (at any given point of time). Normalized Google Distance which is based on the idea of normalized information distance and Kolmogorov Complexity [5], exploits this contextual information hidden in the billions of web-pages indexed by Google to generate a sense of semantic distance between any two concepts. NGD based natural language processing experiments have shown up to 87% agreement level with Wordnet [5].

2.3.3 The Decision Engine

The content based anomaly detection engine clusters and ranks the topics as described earlier. The decision engine, based on a certain threshold value set by the operator (in our case $m = 2/3$, reason discussed later) divides the topics into two fractions, the first fraction consisting of two-thirds of the highest ranked topics which we assume to be normal and the second fraction consisting of the rest of the topics which are initially assumed to be potentially anomalous. Each topic is represented by a tuple consisting of its top ten words. Each anomalous topic is compared with each of the normal topic in the following manner:

1. We find the semantic distance between the first word in an anomalous topic and the first word in one of the normal topics. We aggregate it over all words in the given normal topic and aggregate that over all the words in the given anomalous topic. Let $S^m(i,j)$ denote the similarity between the i^{th} word in a test topic and j^{th} word in the m^{th} typical topic. I and J stand for the total number of words in a test topic and a typical topic respectively (10 in our case). Then relatedness of the test topic with one typical topic is measured as follows. The semantic distances used by us were the ones based on the Wordnet and Normalized Google Distance as described above.
$$r = \sum_{j=1}^{J} \sum_{i=1}^{I} S(i,j) \tag{2.8}$$

2. We then aggregate that over all the normal topics which finally gives us a measure of similarity of the given anomalous topic with all the normal topics. M stands for the total number of typical topics. Hence, relatedness between the given p^{th} test topic and all the typical topics is measured as follows:
$$r_{total} = \sum_{m=1}^{M} \sum_{j=1}^{J} \sum_{i=1}^{I} S^m(i,j) \tag{2.9}$$

3. We repeat the two steps above for all potentially anomalous topics.

Based on the newly available contextual information, the anomalous topics are sorted in ascending order of the context score obtained. Topics with the k lowest scores are considered anomalous. In our current setup, m is determined via user input, since we

are modeling anomaly detection in unsupervised settings. This technique can easily be adapted to supervised settings, where a white list of allowable or typical topics informs our judgment of the m threshold in the decision engine. It is also possible to adaptively use analyst input to change these parameters to refine the accuracy of our system during operation. For example, based on past statistics or domain knowledge, if the operator decides that a certain kind of topic shouldn't be considered anomalous then he can incorporate it in the list of normal topics, so that topics with scores close to this one are no longer considered anomalous. Similarly, if the operator decides that a certain topic should always be considered anomalous, the threshold m can be revised upwards.

2.3.4 Anomaly Detection In Tag Space

Text corpora consisting of a collection of structured textual units like blogs, emails, abstracts etc., have a closer resemblance to the natural language interactions prevalent between human subjects in the society. Hence, a question worth investigating is if the idea of using contextual information from semantic networks to accomplish the task of anomaly detection could also work at the word level and not just at the topic level. We investigated this question by conducting the anomaly detection experiments on tags collected from YouTube videos.

Table 2.1: Example of Tag-Sets(Visually determined anomalous tag In Bold)

Set 1	Education, Tutorial, Teacher, Class, **Somersault**
Set 2	Yoga, Health, Exercise, **Shocking**
Set 3	Lonely, Island, Holiday, Adventure, **Rap**
Set 4	Cute, Babies, Funny, Laughter, **Blood**

We use the "Normalized Google Distance" as our semantic measure in this case due to the noisy/colloquial nature of user supplied tags. Our dataset consisted of 50 sets of tags with one anomalous word each. Table 2.1 shows five such sets with the anomalous word in bold.[1] . The effectiveness of our technique is judged based on the number of times the most anomalous tag adjudged by the human reviewers is matched by the one flagged by our technique.

[1] The anomalous word is shown in bold. The tag labeled anomalous was the one majority of our human reviewers (N=10) agreed upon to be anomalous.

Suppose a video V is accompanied by a set of n tags $I = i_1, i_2, i_3, ..., i_n$. In order to find the most atypical tag in this set, we used a modified version of the K-farthest neighbor algorithm [4], as described below. We calculated the distance of a tag from every other tag in the set and the sum of these distances would result in a score which quantifies the overall divergence of a tag from the rest of the set. Based on value of the parameter k (1 in our case), the top k most divergent tags are flagged as anomalous. We believe that an approach similar to this could be applied in ontologies, taxonomies, concept hierarchies etc., to detect the presence of anomalous content.

Algorithm 2 The Sematic Anomaly Detection Algorithm For Tags

Input: Tag-set T, number of anomalies k

Output: Set of anomalous tags

Initialize the array *scores* to zero {This array will contain the divergence of each tag from the rest of the set} {Find the score for each tag}

While i=1 to n **do**

    While j=1 to n-1 **do**

    D=NGD(i,j)

    Score(i) + = D {Find the score for one tag}

    End while

End while

Sort *score* in descending order

Output Top k tags

Figure 2.5: The Semantic Anomaly Detection Algorithm for Tags

2.4 Results

We conducted experiments on three standard text datasets - Enron emails, DailyKos blogs and NIPS abstracts, which we detail below. The underlying methodology was as follows: we clustered each dataset into 50 topics. Then the topics were ranked by the content engine. The bottom one-third (in our case 16) of the topics were flagged as

potential anomalies. Thereafter, each of the topics was compared with the top 34 topics by the context engine and a score was generated. Finally, based on a new threshold, some of the lowest ranked topics were declared to be anomalous.

The heuristic assumption of considering the top two-thirds of the topics normal is based on two important reasons. Firstly, anomalies are rare and are hence topics that are statistically rarer and hence ranked lower in our list are more likely to be semantically atypical. Secondly, during our preliminary empirical investigations we experimented with different values of the threshold and found the above assumption being found most effective in terms of removing anomalies. We would like to emphasize that this threshold parameter could be set using a domain expert's insights rather than relying completely on empirical tuning like in our case.

Each of the next three subsections has the following format. We start with a brief description of the dataset and then show a diagram which consists of 5 tables. The table in the center shows the top 4 most representative topics of the data set which reflects the general tenor of the corpus. The table in the upper right corner shows the top 4 most anomalous topics in the dataset based on content based ranking which basically shows the most statistically "rare" topics in the given corpus. The table in the upper left corner shows the top 4 topics identified as interesting or anomalous by human evaluators. For human evaluation, we have used a majority voting scheme using 5 human subjects who were asked to rate the anomalousness of bottom 1/3rd of the topics generated by our technique on a scale of 1-10 (10 being the most anomalous). The human subjects were graduate students in the department of Computer Science at the University Of Minnesota who were informed about the general tenor of these three corpora before conducting the poll. The inter-rater agreement was 0.54 which is considered to be in the range of good agreement. The tables in the bottom left and right corners show the top 4 most anomalous topics in the dataset after context matching using Wordnet and Normalised Google Distance respectively. We then conclude each subsection with a discussion on the quality of our obtained results.

Table 2.2, shows all the quality metrics obtained by us on these three data-sets. We report precision, recall, F-score, sensitivity and specificity scores of the classification of the anomalous class. The range of all these values is between $(0, 1)$ (1 being optimal). Our evaluation measure was F-score as it captures both the accuracy and sensitivity of

a model. A brief description of these measures is provided below. Let tp, fp, fn, tn denote the number of true-positives, false-positives, false-negatives and true-negatives respectively, obtained from a classification task. Then, $Precision = tp/(tp + fp)$, $Recall = tp/(tp + fn)$ and $F = Harmonic\_mean(Precision, Recall)$. We also report sensitivity and specificity scores as they are quite popular in the anomaly detection literature. $Sensitivity = tp/(tp + fn)$. $Specificity = tn/(tn + fp)$.

Table 2.2: Results on all the data-sets with and without the addition of contextual information. Notice the increase in specificity (False Positive Rate=1-Specificity) and hence the decrease in false positive rate upon augmenting contextual information. The F score increases with the addition of context. Wordnet performs slightly better than NGD

Data-Set	Precision	Recall	F Measure	Sensitivity	Specificity
Enron without context	0.62	1	0.77	1	0
Enron with Word-net	0.889	0.8	**0.84**	0.8	0.83
Enron with NGD	0.77	0.7	0.73	0.7	0.67
NIPS without context	0.25	1	0.4	1	0
NIPS with Word-net	1	0.8	**0.88**	1	0.9167
NIPS with NGD	0.5	0.5	0.5	0.5	0.833
Kos without context	0.43	1	0.60	1	0
Kos with Word-net	0.8	0.57	**0.67**	0.57	0.89
Kos with NGD	0.75	0.42	0.54	0.4286	0.88

2.4.1 Enron Email Data Set

The corpus consists of a set of email messages. This data set after de-identification of private fields was made public during a legal investigation. The original dataset contained 619,446 email messages from 158 users. A cleaned version of this corpus contains 200,399 messages from 158 users with an average of 757 messages per user [26]. We have used a bag of words version of this cleaned data set which contains 28,102 unique words and approximately 6,400,000 total words in the entire collection. Please note that a lot words are filtered in the process of tokenization, stop-word removal etc.

Expert evaluation flagged 10 topics as deviant/interesting/noisy from the set of 50 topics in this dataset. These were topics like discussion about fantasy football, travel discussions, football teams, legal agreements, spam messages, system maintenance emails, discussion among new MBA graduates etc.

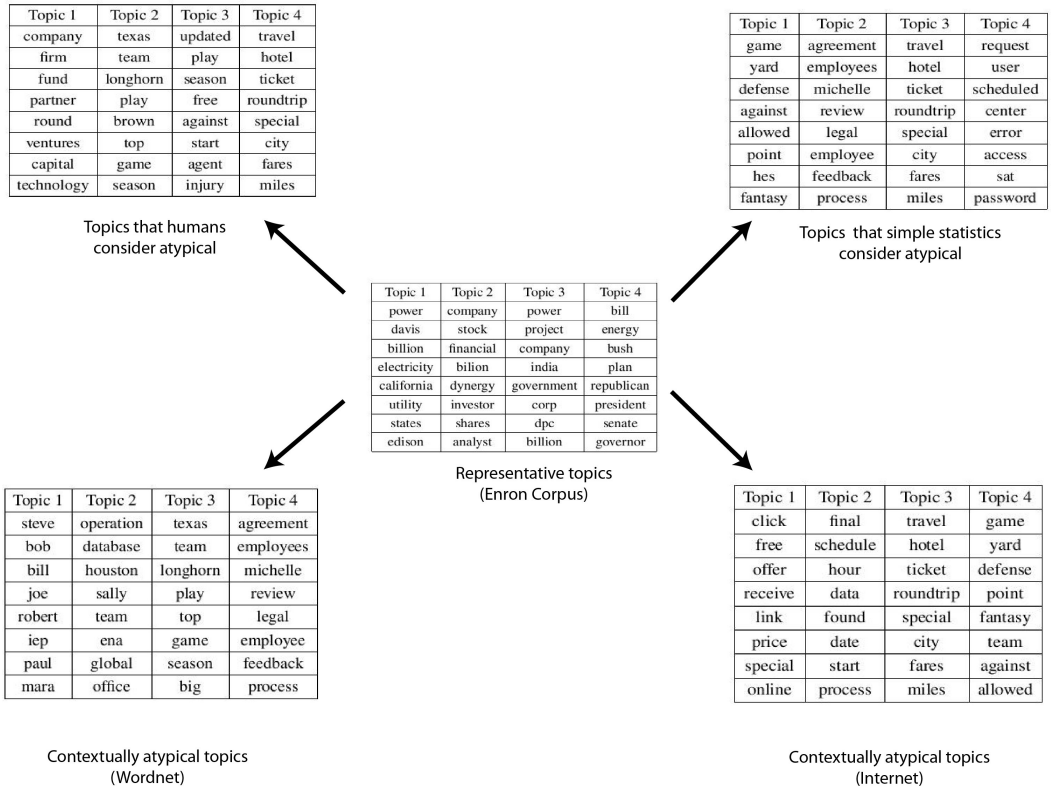

Figure 2.6: Results from Enron dataset: Table in the center shows the most representative topics, table in the upper left corner shows the human adjudged anomalies, table in the upper right corner shows the statistical anomalies and tables in the bottom show the final contextual anomalies based on Wordnet (left) and Internet (right) respectively.

Our system based on the Wordnet similarity measures flagged 9 topics as anomalous and 8 of those matched with the ones picked out via human selection. The two topics it couldn't predict correctly were as follows, one was a topic with spam keywords and one was a discussion thread among new MBA graduates. However, it was able to identify most noisy clusters really well. It had just one false positive (which happened to be a topic about marketing). It was able to flag a topic about legal agreements correctly as anomaly which was ranked 35 by content ranking. It was able to filter out topics (like topic 1 in the upper right table) which are contextually uninteresting, though statistically interesting.

Our system based on Normalized Google Distance flagged 9 topics as anomalous and 7 of those matched with the ones picked out via human selection. The three topics it couldn't predict correctly were as follows,first one was a topic about fax and phone communication issues, the second one was a topic about legal agreements and the third one was a topic about discussions among new MBA graduates. Again it was able to identify most of the noisy noisy clusters really well. It had two false positives, first was a topic about football and second was a noisy cluster. It was able to flag the topic with spam keywords as anomalous unlike Wordnet based measures which were unable to do so. It should be noted that the topic about discussion among new MBA graduates was the only topic which was considered anomalous by human evaluators but not flagged by either Wordnet based measures or Normalised Google Distance. As summarized in Table 2.2, F-score increases with the augmentation of context. Another thing worth noting is the increase in the specificity of the model.

2.4.2 DailyKos Blogs Data Set

Dailykos.com is an American political blog that publishes political news and opinions, typically adopting a liberal stance. This data set consists of a set of blogs taken from this website.It has 3430 documents, 6906 unique words and approximately 467714 total words in it, in the bag of words format.

Our results on this dataset are not as clear-cut, but illuminating nonetheless. Human evaluation did not find any of the topics generated in our content analysis to be particularly anomalous (low anomaly scores assigned by humans in the scale of 1-10, 10 being the most anomalous). Our system flagged 5 topics as anomalous using Wordnet

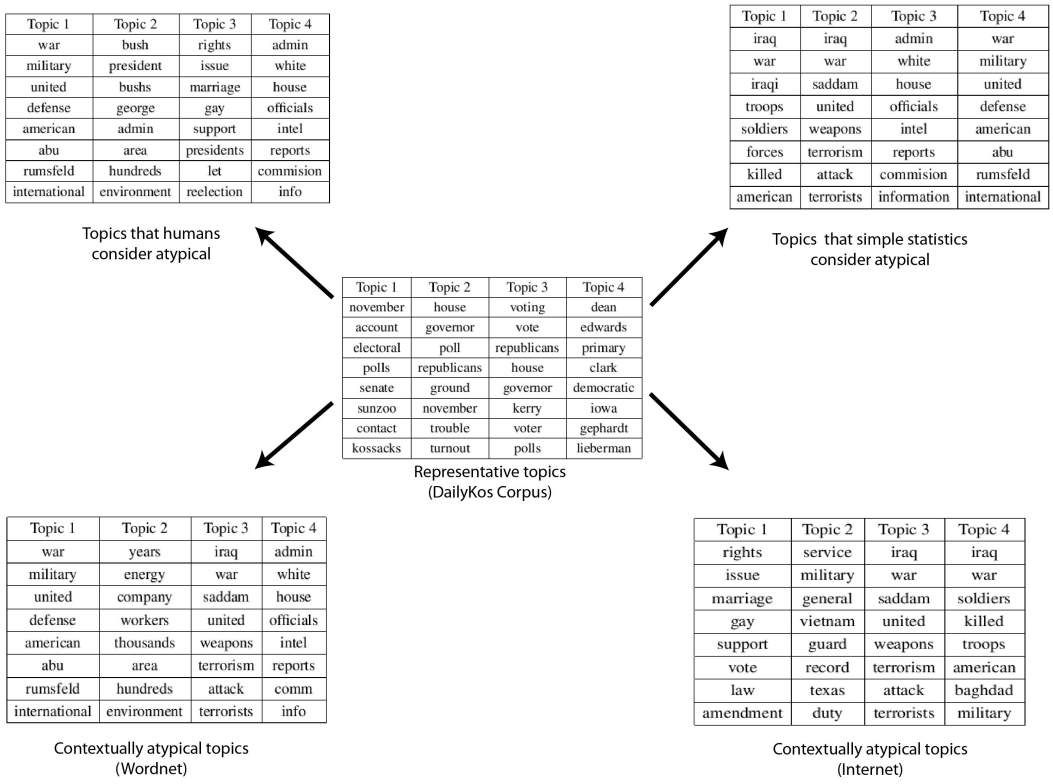

Figure 2.7: Results from the DailyKos dataset: Table in the center shows the most representative topics, table in the upper left corner shows the human adjudged anomalies, table in the upper right corner shows the statistical anomalies and tables in the bottom show the final contextual anomalies based on Wordnet (left) and Internet (right) respectively

similarity measures and 4 topics as anomalous while using Normalized Google Distance, none of which, however, are intuitively out of place in the Daily Kos setting. The anomalous topics as shown in the bottom two tables are about iraq war, gay rights, budget cuts, Abu-Ghraib prison incident etc. Some, such as discussions of the Iraq War etc. are quite representative of the tenor of this website. As in the previous case, our algorithm was able to eliminate several topics in the second stage of contextual analysis which might have otherwise resulted in false positives. However, unlike in the case of Enron, where clear explanations that determine whether a particular statistical anomaly or not are available, in this case, no such clarity is forthcoming. In particular, it appears that none of the topics determined to be anomalous in a statistical sense are not, in fact, real anomalies in the political context in which Daily Kos operates. We believe this is a limitation of the external corpus that we resort to obtain contextual information.

Recall that both Wordnet and Internet are general semantic relation corpora, whereas Daily Kos text is immersed in a far more sophisticated context of political discourse. As a consequence, the context inferred fails to capture the semantic associations between topics such as gay rights, Iraq war etc. with typical content on Daily Kos. A more dynamic approach to context detection, potentially leveraging internet resources, would be expected to perform significantly better in this case. We believe these negative results are interesting, because they accentuate the fact that our findings in the Enron dataset are not statistical artifacts, but consequent to value added by generic semantic context.

2.4.3 NIPS Papers Data Set

NIPS which stands for Neural Information Processing Systems, is a conference on computational neuro-science. This data set consists of a set of full papers taken from collection of papers published in this conference. It has 1500 documents,12,419 unique words and approximately 1.9 million total words in it, in the bag of words format.

In comparison to the previous two data sets this was the most difficult data set to deal with. This was because there was very little divergence between the various topics and there was also very little noise in the data set. Expert evaluation flagged 4 topics as deviant/interesting/noisy from the set of 50 topics. These topics were slightly less frequent than the top 34 topics and were related to acoustics, principal component

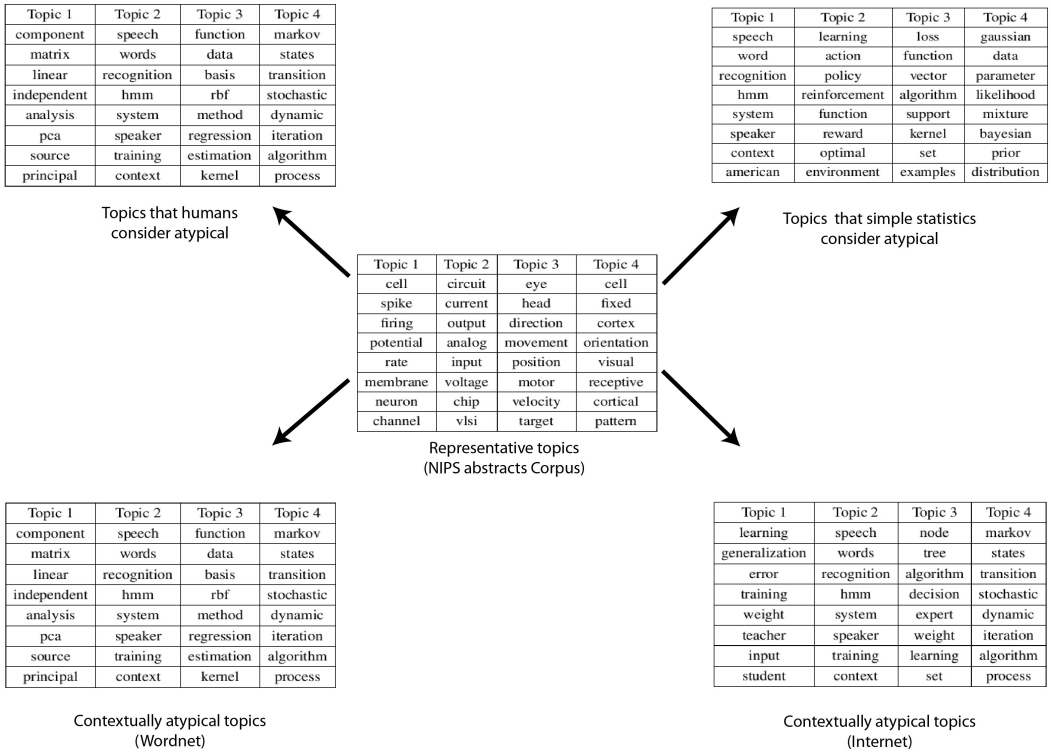

Figure 2.8: Results from the NIPS abstracts dataset: Table in the center shows the most representative topics, table in the upper left corner shows the human adjudged anomalies, table in the upper right corner shows the statistical anomalies and tables in the bottom show the final contextual anomalies based on Wordnet (left) and Internet (right) respectively

analysis, regression etc. Our system based on Wordnet measures flagged 5 topics as anomalous and 4 of those matched with the ones flagged via human analysis. We had just one false positive (it was a topic related to support vector machines). On the contrary our algorithm based on "Normalised Google Distance" didn't perform just as well. It flagged 4 topics as anomalous and only 2 two of them matched with human judgment. The amount of noise in the NIPS dataset was quite low. As a result, Wordnet which is more accurate semantic network turned out to be a better match as compared to the Internet. We believe this observation could guide future decisions regarding choice of semantic measures to perform anomaly detection. The choice between Wordnet and Normalized Google Distance, is the choice between greater accuracy versus greater coverage.

As this corpus consisted of a set of papers from a technical conference, it is somewhat natural to find that the text samples were heavily focused towards the central theme of the conference which is neuroscience, artificial intelligence and machine learning. The biggest victory of our technique with this dataset was the fact that only content based analysis could have resulted in a very high number of false positives which is something our algorithm could avoid. The number of false positives was reduced heavily as the specificity increased from 0 to 0.9167. (Table 2.2)

2.4.4 YouTube video tags dataset

YouTube is a well-known community contributed video repository. Users upload videos on various themes and topics from all over the world on this website. During the process of upload, users assign certain tags which characterize the content of the given video. For example, the tag set "*smartphone*", "*apps*", "*demo*" would indicate that the given video is likely to be a demonstration of smart-phone applications. Hence, tags can be assumed to contain a concise summary of the given video's content. Due to various reasons like presence of inherent unexpected content, user errors, use of technical jargon, cross lingual colloquials etc., some videos end up having some anomalous/noisy/atypical tags. We would like to use our methodology to be able to detect those anomalous tags.

The average number of tags per video is between 4-12[2] . The number is too low for the effective application of statistical text clustering algorithms which rely heavily

[2] This observation is made on a dataset of 42000 videos.

on statistical frequencies of word co-occurrences to extract meaning. Hence, we directly jump to the second step of our algorithm, which is to test for contextual atypicality. Finally, we evaluated the judgment of contextual atypicality obtained from our algorithm against the judgment of human evaluators.

Out of the 50 sets containing one-anomaly each, our algorithm was able to match the human judgment 41 times correctly.(Accuracy=0.82) Our algorithm failed to detect anomalies in tag-sets containing multiple conflicting tags or the sets where the anomalous word seemed too subtle. A few examples of the failed cases are shown in Table 2.3. For example, in set 3 in Table 2.3 the human reviewers marked the word "orange" as anomalous possibly because the words mouse, computer and monitor have strong correlation with the computer company "Apple" so the word orange seems anomalous to them, whereas the word that stands out as semantically anomalous is the word "computer", possibly because the rest of the words have some connection with biological world except this one. This experiment supports our conjecture (with reasonable amount of confidence) that contextual information could be used to accomplish anomaly detection even at word-level.

Table 2.3: Example of Tag Sets where the judgment of our algorithm proved inadequate.(Human adjudged anomalous tags are shown in bold.The ones detected by our algorithm are shown in italic)

Set 1	fish, *net*, hook, island, **U2**
Set 2	*motivational*, speaker, speech, inspiration, **sony**
Set 3	mouse, *computer*, monitor, apple, **orange**
Set 4	*angry*, birds, game, mobile, **playstation**

2.5 Discussion Of Results

We make three contributions to the state of existing anomaly detection techniques. First, we show how the use of external context information can reduce the false positive rate in existing systems by explaining away spurious statistical deviations. Previous research has settled the theory of anomaly detection for categorical and numeric data quite decidedly. Using similarity metrics on test data samples to estimate statistical deviance from typical system behavior proves to be a robust anomaly detection strategy

across domains and datasets [4]. Because of the acontextual nature of such data objects, the use of external data to inform and contextualize anomalies detected has typically not been considered, although [19] have recently described how contextual and functional constraints system constraints can be used to bias anomaly detection techniques towards finding systemically meaningful anomalies. However, their approach is primarily concerned with using contextual constraints as a pre-processing step to reduce the dimensionality of the feature space to be searched for anomalies, whereas our technique uses such information as a post-processing filter.

Introducing contextual information from a different data corpus helps us find semantic explanations for anomalies and in filtering out false positives. For example, in our analysis with the Enron email data set, the most anomalous topic after content analysis was a theme related to venture capital investment in a technology firm. This was filtered out by the decision engine after contextual analysis as a false positive, a conclusion in concordance with the evaluation of human subjects. By augmenting contextual information, we can also flag previously suspect data points as anomalies with greater confidence.

Second, contextual information can be used to detect previously undetected anomalies, if we allow for adaptive threshold manipulation based on operator input. If we can augment these anomalies in our contextual stream appropriately then it could help in detecting anomalies similar to the one flagged by our operator more easily in the future. For example: during our analysis with the Enron email data set, a topic about legal agreements was ranked higher (less anomalous) by the content analysis engine, but context analysis was able to flag it out as an anomaly. Such disparity in content-context rankings can inform operators' judgment to tune the threshold better so that such topics aren't missed out in the future, or to construct better similarity metrics for the content analysis task itself. Additionally, by augmenting a dummy topic in our contextual database, in the spirit of 'must-link' proposals [22] in earlier constrained clustering approaches, containing words related closely to legal agreements, we can promote subsequent detection of related topics going forward.

Third, contextual information can be used to perform anomaly detection in datasets made up of user generated exemplars (YouTube tags in our case). This shows that this idea could be implemented at various levels of abstraction as we have operationalised it

at word level and topic level. The overall complexity of our technique is as follows:

1. The first step which clusters the text corpus into topics has complexity: $O(((NT)^{\tau}(N+\tau)^3))$ [17], where N is the number of words in the corpus, T is the number of topics in the corpus and τ is the number of topics in a document. This has polynomial run-time if τ is a constant.

2. The second step which performs context incorporation has complexity: $O(m*n*k*k)$, where m is the number of training topics, n is the number of test topics and k is the number of words in a topic. (k=10 in our case)

3. The overall complexity is thus: $O(((NT)^{\tau}(N+\tau)^3))+O(m*n*k*k)$. The second step doesn't affect the overall complexity of the algorithm asymptotically.

The two major computational tasks performed by our algorithm are performing topic modeling to extract topics and computing semantic distances. We mentioned earlier that the iterative nature of LDA can scale it up to millions of documents. Calculation of semantic distances using Wordnet can be performed in polynomial time. Calculation of semantic distance using the web is restrictive in practice, as search engines throttle automated queries (e.g., Google permits only up to 1000 queries per day from an IP address). In order to tackle this challenge, we suggest the use of a dictionary which enlists the number of search queries returned for a list of common terms by a search engine. Hence, we believe that with the use of map-reduce framework to perform clustering and with the use of a pre-computed Internet semantic dictionary, it is possible to apply our technique on much larger datasets.

To conclude, in this chapter, we have proposed that by augmenting topic modeling techniques with contextual information derived from semantic networks we can improve the detection of deviant topics in large scale text logs. We were able to validate this empirically and build a system which could accommodate the human judgment of anomalies as well. Our results show both reductions of false positives and detection of previously undetected anomalies in existing datasets.

Chapter 3

Characterising Interestingness In Semantic Space

3.1 Motivation and Related Work

The desire to solve a puzzle, the uncontrolled urge to look through your friend's inbox and many other such curiosity driven information seeking activities exhibited by humans seems to share the common need for sense making in a given situation. Various kinds of information seeking responses are triggered in humans based on how surprising, deviant or salient a particular set of observations are. A stimulus that instigates curiosity and elicits information seeking behavior in humans is often called an "interesting" stimulus. For example: a humorous joke is often surprisingly deviant, an interesting metaphor is usually generated by a contextually dissimilar choice of words etc.

Investigations conducted in the field of psychology, in order to find the underlying causes of curiosity have led to many popular theoretical accounts (like Competence Theory, Drive Theory etc. [45]), of which one of the most accepted accounts is the "Incongruity Theory". This theory postulates that curiosity is the result of an innate human tendency of making sense of the world which is fueled by the failure of subjective expectations of prior beliefs held by the agent. The need for sense-making seems to be aversive at extreme values which implies curiosity is caused by the presence of "optimal incongruity" in observations. Stated otherwise, the presence of too much incongruity in an observation leads to the agent perceiving the given observation as noise. On the

contrary, the presence of too little incongruity in an observation doesn't instigate an agent towards sense making hence it eventually leads to boredom. Can this qualitative idea that need for sense making fueled by the presence of optimal incongruity which causes curiosity and hence interest stimulation be modeled quantitatively?

There have been many notable attempts of operationalizing the incongruity theory of curiosity. [45] proposed an information theoretic measure called information gap which is the divergence between what the agent already knows about the environment and what he seeks to know further. [48] have provided an exhaustive literature survey of computational methods used in robotics which make use of entropy based measures quantifying incongruity. Methods like uncertainty motivation, information gain motivation and distributional surprise motivation etc. all try to capture the basic idea behind incongruity theory of curiosity. Previously, in 2, we operationalized the incongruity theory of humor algorithmically to detect funny videos from YouTube, using a given video's tag-space.

This chapter presents a domain independent theoretical and algorithmic framework which characterizes incongruity and hence interestingness[1] in the semantic space. We detect, categories of words considered creative by human subjects, most liked set of media objects on a social network and emails considered interestingly deviant from the central theme of Enron email corpus. Our operationalization results in three important findings. First, incongruity is strongly correlated with data-points considered interesting across different domains and hence leads to their easier detection. Second, incongruity is a necessary but not a sufficient condition for characterizing interestingness and third, there exists a monotonically increasing relationship between the amount the incongruity and the amount of interest (till the upper bound) i.e. higher amount of incongruity leads to greater interest stimulation.

The rest of the chapter is structured as follows. Firstly, we present the information theoretic arguments behind our instantiation and then show how that leads us to derive the two measures (Diversity and Anomalousness) whose simultaneous presence characterizes incongruity in a given semantic space. We then describe a computational

[1] It should be noted that our quantification of interestingness based on incongruity theory is measuring what a sample of diverse set of people would most likely find interesting. Currently, we are not measuring a personal cognitive notion of interestingness. For example, there might be a few outlying human subjects who don't like any incongruity in their observations at all.

methodology for measuring conceptual incongruity for a set of words under varying degrees of noise and then present empirical results on various kinds of datasets. This work presents the first known operationalization and extensive empirical validation of the incongruity theory of curiosity in the semantic space.

3.2 Theoretical Motivations

The idea behind incongruous stimuli being able to instigate curiosity in living organisms which in turn triggers information seeking behavior in them can be formally grounded in an account of the "value of information" to an information foraging species. Humans constantly weigh in and weigh out the value of the different options presented to them during the process of decision making, preferring options with greater value. Treating information like a resource makes it evident that incongruous or surprising data points (i.e. data points with high divergence from the current belief distribution) contain a greater amount of information that the agent can use to update his model about the world. Data points which contain redundant information are deemed uninteresting and ones which are divergent from the current belief distribution are deemed interesting.

In order to characterize a given data-set as interesting or not, we have to compute the divergence of the content of this dataset from the expected belief distribution regarding the content of the dataset. Theoretically, the divergence between any two probability distributions can be computed using the measure KL-divergence. The divergence between expected prior distribution and obtained posterior distribution can then be used to find the most surprising/incongruous regions in the sampling space.

We will now consider a notional scenario to illustrate the above point. Let us say an agent has prior beliefs on k hypotheses about all possible "themes" of a movie, before actually watching the movie and his prior beliefs are built upon his previous experiences with similar movies' trailers. Let a_1, a_2,..,a_k be k possible hypothesis and let $P(a_k/H)$ denote the prior probability of the agent on one of the hypothesis and let H denote the history of the agent. Hence, the probability of watching a certain kind of movie given an observation is given by the posterior probability. (Equation 3.1).

$$P(a/b, H) = \frac{P(b/a, H) * P(a/H)}{\sum_{a_1}^{a_k} P(b/a_k, H) * P(a_k/H)} \tag{3.1}$$

If the observation b is a surprising one, then the posterior probability should diverge from the assumed prior (e.g. expected comedy movie turns out to be thriller) which can be calculated using the "Information Gain" of this sample.

$$InformationGain(b) = \sum_{a} P(a/b, H) * log \frac{P(a/b, H)}{P(a/H)} \tag{3.2}$$

Figure 3.1 shows two notional scenarios. In the both the cases the amount of surprise can be quantified using equation 3.2. And finally, the expected value of the information gain of the entire data-set (assuming there are n points in the data-set) quantifies the overall incongruity of the data-set. $Incongruity(dataset) = \sum_{1}^{n} E(InformationGain(d_i))$.

The idea of information gain (as explained above using our notional scenario) has been used by many others in the past to characterize different aspects intelligence. We will now explain how the above formulation can be used in the domain of text-data to quantify incongruity and expected informativeness. Given a set of words S, if we can compute the KL-divergence between the probability of co-usage of the given set of words (the posterior in our case) from a universal prior belief distribution which captures the probability of co-usage of all the words in English language, then we can compute the inherent conceptual incongruity in the given set. This in turn can lead us to find the interestingness of the dataset as described above. Words which have high statistical co-occurrence frequency (put simply words which come together more often) are likely to have shared lots of common contexts and their future co-occurrence shouldn't be deemed surprising. Hence, conceptual incongruity could only result with the use of dissimilar terms together. We use Normalized Google Distance (NGD henceforth) [5] as our semantic measure, which calculates the information distance between any two words based on their co-usage in the entire World Wide Web. We assume that the Internet captures the context of general co-usage of all the words in the society and can hence better capture the idea of similarity between words based on co-usage as compared to any word ontology like Word-net or semantic similarity measure like LSA. Additionally, NGD can also find the distance between noisy misspelled words, internet neologisms (e.g. lol, rotfl) etc.. Let the prior belief distribution for the co-usage of all the words in language extracted using NGD be denoted by X which is same as distribution of $P(a/H)$ in equation 3.2 and the posterior distribution for the co-usage

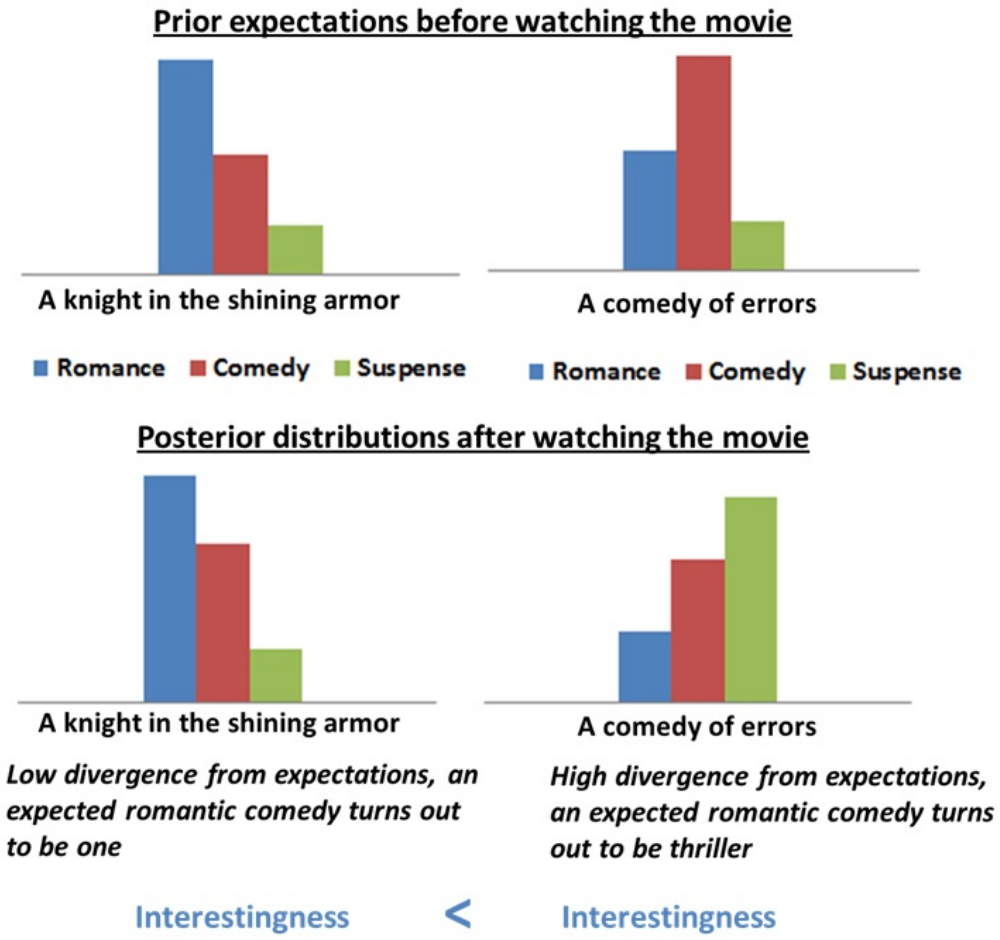

Figure 3.1: This figure shows a notional scenario involving two movies. The movie on the LHS is expected to be a romantic comedy (as shown by priors) and turns out to be one (as shown by posteriors) and hence the information gain is lower. The movie on the RHS is expected to be a comedy with some romantic elements in it, but it turns out be a thriller. The information gain is higher and so is the likelihood of it being deemed more interesting of the two movies.

of all the words in a given piece of text be denoted by Y which is same as distribution of $P(a/b, H)$. Let the $InformationGain$ of two set of words ($S1$, $S2$) be denoted as $IG1$ and $IG2$ respectively. If $IG1 - IG2 > \epsilon$, where ϵ is a parametric threshold, then we can say the $S1$ has greater divergence from the prior beliefs as compared to $S2$ and is hence likely to be more interesting. Let the entropy of a distribution X be denoted by $H(X)$ ($H(X) = \sum_i^n -p * logp$). Using information theoretic identities, we can restate equation 3.2 as equation 3.3 where distributions of $P(a/H)$ and $P(a/b, H)$ are denoted as X and Y respectively.

$$IG(Y; X) = H(Y) - H(Y/X) \tag{3.3}$$

As argued, earlier ($IG1 - IG2 > \epsilon$), an algebraic difference of two information gains can help us compare the interestingness of two sets. Information Gain consists of two factors, $H(Y)$ denotes the entropy of a given set of words $S1$ and $H(Y/X)$ denotes the entropy of $S1$ given the normal co-usage of words in the universe. The first factor $H(Y)$ (we call it *Diversity*), is simply the amount of dispersion in a given set of words and the second factor $H(Y/X)$ (we call it *Anomalousness*) is simply the fraction of number of anomalous words with respect to normal co-usage in a given set. Figure 3.2 considers a few examples to further clarify our point. The first set of words (apple, banana, mango) has no apparent ambiguity and clearly refers to a set of fruits. The second set of words (apple, mango, banana, monkey, orangutan) has two different sense clusters namely; fruits and primates. The presence of multiple contexts is causing a minor conceptual dissonance and is hence adding some incongruity. Hence, *Diversity* of a given set, is simply a measure of semantic dispersion ($H(Y)$) of a given set. The third set, (apple, banana, mango, treadmill) has one statistical anomaly (treadmill) which doesn't fall in line with the main theme of the set. In the third example, the incongruity is being introduced by a statistical anomaly which cannot be accommodated in the main theme of the set. Hence, *Anomalousness*, is a measure of number of anomalous words ($H(Y/X)$) in a given set; given the normal usage of words. The fourth set is a semantic deconstruction of the funny statement with some inherent conceptual ambiguity ("Health conscious monkeys and orangutans, prefer fruits for lunch after working out on the treadmill"). It shows that the simultaneous quantification of *Diversity* and *Anomalousness* can characterize the innate conceptual

incongruity in a given semantic space.

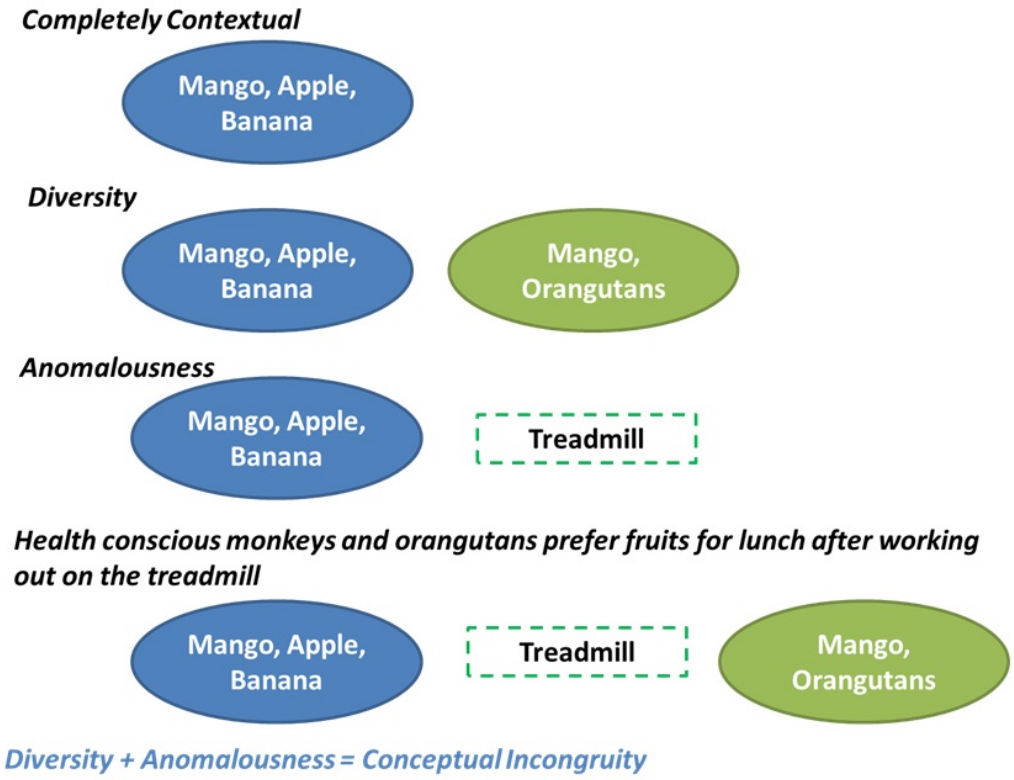

Figure 3.2: This figure represents a notional semantic space and shows the intuition behind the factors used by us to quantify conceptual incongruity. A detailed explanation is given in section 1.

3.3 Models and Methods

As mentioned previously, we expect conceptual incongruity to be directly correlated with *Diversity* and *Anomalousness*. First, we create a concept space (in the form of a similarity matrix) for a given set of words using NGD. Then we compute the Diversity and Anomalousness scores of that matrix. Finally we combine these measures suitably based on the problem objective.

3.3.1 Calculating Normalized Google Distance

Given two words x and y the NGD between them is given by equation 5.1. $f(x)$ and $f(y)$ denotes the number of pages containing x and y respectively and $f(x,y)$ denotes the number of pages containing both (x,y) as returned by Google and N denotes the total number of pages indexed by Google.

$$NGD(x,y) = \frac{max(logf(x), logf(y)) - log(f(x,y))}{logN - min(logf(x), logf(y))} \tag{3.4}$$

This measure assumes that ratio of the frequency of total number pages containing a term divided by the total number of pages indexed by Google, approximates the probability of that term's presence in the World Wide Web. NGD is our tool of choice because of reasons described in detail in Chapter 4. The range of this measure is (0,∞) where 0 indicates complete similarity and ∞ indicates complete dissimilarity.

3.3.2 Defining The Similarity Matrix

In order to simplify our analysis of the non-metric concept space of a given set of words we first create a similarity matrix. Let I denote a set of concepts/words $I = \{i_1, i_2, .., i_n\}$ where i_k is the k_{th} word. The cardinality of the set I is assumed to be n. We then construct a n*n matrix where each entry of the matrix is the NGD between any two words (i_1, i_2). This matrix is symmetric in nature because NGD is a symmetric measure.

3.3.3 Defining Diversity

The first thing we need to calculate is presence of statistical *Diversity*, indicating the presence of conceptual dissonance in the given similarity matrix. Firstly, we sum the similarity matrix across the rows to produce the sum of deviations of each word from the rest of the words. Assuming, the cardinality of the set is n, we now obtain n such sum of deviations of each word from the rest of the set and we now need to compute the amount of dispersion in this set. We use two well-known measures of dispersion for doing so, as detailed below.

Least absolute deviation or L1 norm measures dispersion as the least absolute cumulative deviation within a set. $L1$ norm is known to be robust to the presence of noise and outliers and hence it becomes a suitable choice while dealing with crowd-sourced

data like tags taken from videos. Let us say a certain data-set consists of a set of points (x_i, y_i) with $i = 1, 2, ..n$. Our goal is to find a function f(x) of the form $y = mx + c$ which approximates the values of y. In this case, we estimate the values of m and c to minimize the cumulative sum of absolute deviations.

$$S = min(\sum_{i=1}^{n} |y_i - f(x_i)|) \tag{3.5}$$

S denotes the least absolute deviation norm of the set. Finally the number of words varies from set to set; hence we normalize the above calculated measure with the cardinality of the set. NLAD stands for normalized least absolute deviation. $NLAD = S/n$.

Median absolute deviation (MAD henceforth) measures dispersion as the median of the absolute deviations from the data's median. It gives a robust estimate of the dispersion in a univariate data-set.

$$MAD = median(|(x_i - median(X)|) \tag{3.6}$$

x_i denotes a data-point and X denotes the entire data-set. In our initial empirical analysis, we realized that if the words in the set were taken from dictionaries and word-ontologies MAD provided a better estimate of the dispersion in the set. In noisy data-sets like the ones taken from social networks which consisted of neologisms and words with latent contextual meanings, use of $NLAD$ resulted in a better estimation. This probably has to do with the fact that MAD is more suitable for univariate data and we can safely assume that crowd-sourced data has been affected by more contextual variables in general.

3.3.4 Defining Anomalousness

The second thing we need to quantify is the presence of semantic anomaly/anomalies in the concept space. We do that in the following two ways depending on the amount of noise in the data-set. After creating the similarity matrix, we first compute the sum of deviations of each word from all the other words in a given set. Let the similarity matrix be denoted by S where each entry $S(i, j)$ denotes the distance between $word(i)$ and $word(j)$. The cumulative divergence score d_i for the i_{th} word is given by equation 4.6. After obtaining the divergence scores; d_i for all the words, the problem reduces to

finding the most deviant points from this set. We use two well-known techniques from the anomaly detection literature [4].

First one is called inter-quartile measure and is summarized in equations 4.7 and 4.8. $Q1$ denotes the first quartile, $Q3$ denotes the third quartile and IQR denotes the interquartile range. Points lying outside the ranges specified by equation 4.7 and 4.8 are counted as anomalies. UB and LB denote the upper and lower bounds respectively. This measure is robust to minor amount of noise and does effective anomaly detection in highly noisy data-sets.

The second one assumes that the divergences follow a Gaussian distribution and then finds the mean and variance of this distribution. It then finds the number of points above and below the $k * \sigma$ threshold (Equations 4.9, 3.11), where k is a multiplicative constant. μ denotes the mean and σ denotes the variance. This measure can be tuned using different values of k for varying degrees of noise in the data. Let $n1$ and $n2$ be the number of points above and below the upper and lower bounds respectively. As the number of words per set is likely to vary, we normalize the anomaly score of a set by its cardinality. (Equation 3.12)

$$d_i = \sum_{j=1}^{n} S(i,j) \tag{3.7}$$

$$LB = Q1 - 1.5 * IQR \tag{3.8}$$

$$UB = Q3 + 1.5 * IQR \tag{3.9}$$

$$LB = \mu - k * \sigma \tag{3.10}$$

$$UB = \mu + k * \sigma \tag{3.11}$$

$$Score = (n1 + n2)/n \tag{3.12}$$

3.3.5 Perception Of Incongruity

As described in the earlier anecdote in Figure 3.2 and then later mathematically in equation 3.3 we quantify, "conceptual incongruity" as the simultaneous presence of diversity and anomalousness. In order to quantify the simultaneous presence of the above two factors (Diversity, Anomalousness) in the concept space we combine their individual scores using simple linear regression or a simple positional rank aggregation

technique called Borda [30], depending on whether we seek a parametric or a non-parametric model. Chapter 4 contains a brief description of Borda's technique.

3.4 Results

3.4.1 Creative Categorization

Our first investigation was conducted on a data-set introduced by [43]. This data-set consists of 105 categories constructed by human subjects where each category consists of a certain set of words surrounding a central theme (E.g. Bulb, Humans, super natural powers etc.). These categories are then rated by other human subjects, judging the amount of creativity required to come up with these categories. Another set of participants were then asked to rate the amount of similarity in the choice of words in the least and most creative categories respectively. Based on these experiments the authors postulate that creative categories are formed by grouping dissimilar words together. This follows the intuition that people use their imagination to put together dissimilar concepts which together carries some holistic meaning.

Does conceptual incongruity characterized by semantic diversity and anomalousness correlate with human creativity? To investigate this question, we applied our technique to each of these categories. As data-set consists of words taken from dictionary only, we used the Wikipedia pages returned by Google search engine as our estimate of $f(x)$ (equation 5.1). This being a completely noise free data (in a semantic sense) we used median absolute deviation as our measure of diversity (equation 3.6) and $k*\sigma$ deviation (k=1.5) from mean as our estimate of anomalousness (Equations 4.9, 3.11).

We then divided the data-set into 7 clusters, the first cluster contained the 3 most and 3 least creative categories (total 6), the second cluster contained 5 most and 5 least creative categories (total 10) and so on and the final cluster contained all the 105 categories. Figure 3.3 shows the Spearman correlation of *Diversity*, *Anomalousness* and the combined *Incongruity_Score* (generated using simple linear regression) with the human generated creativity ratings of these categories. The overall correlation of *Incongruity* with creativity ratings for 105 categories was about 0.38. But the key observation here is that correlation between creativity ratings and incongruity keeps increasing as the number of categories is decreased, which clearly indicates that the

most creative categories have high semantic incongruity and the least creative categories have low semantic incongruity. The correlation values for the cluster containing 6 and 10 categories are as high as 0.8873 and 0.7701, which clearly indicates that a creative choice of words which instigates interest in human beings can be characterized using semantic diversity and statistical anomalousness. (Note:All p-values are less than 0.01)

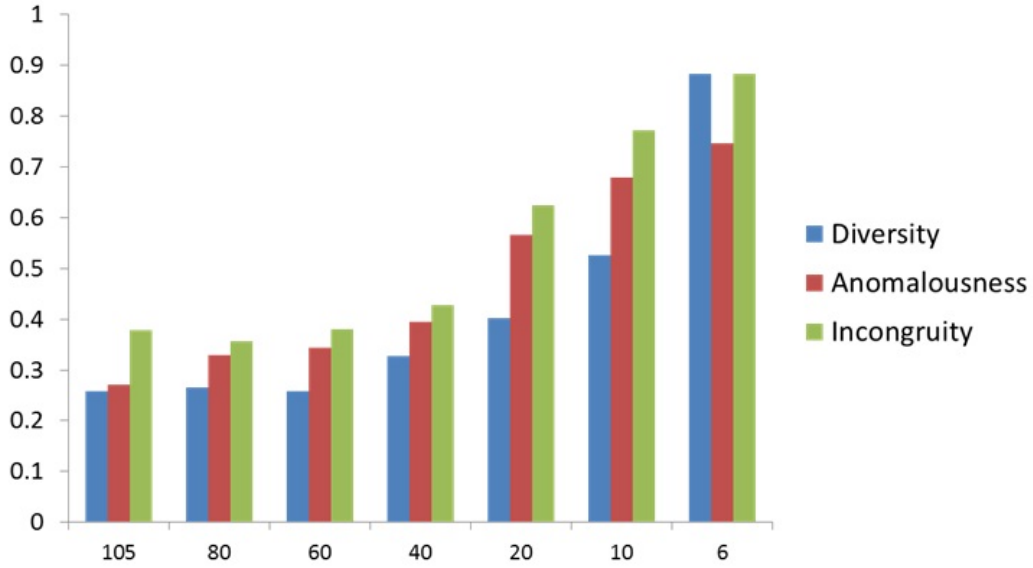

Figure 3.3: This figure shows the correlation between the categories and the creativity ratings. X-axis denotes the number of categories and Y-axis denotes the values of spearman correlation rho.

3.4.2 Perception Of Humor

Humor is also a kind of interestingness as in, a humor inducing stimuli like a joke or a funny video instigates interest seeking response in a subject. Here we briefly reinterpret some of the results discussed in great depth in Chapter 4 to show the generalizability of our technique and the similarity in monotonic trends and sufficiency conditions, also observed with the other data-sets. The mean anomalousness, diversity and Incongruity scores (calculated using equations 4.7, 4.8, 3.5) of 14 YouTube video categories are shown in Table 3.5. The table clearly shows that these measures capture the strong

Table 3.1: Mean diversity, anomalousness and incongruity scores for all 14 video categories ranked in descending order of incongruity

Category	Diversity	Anomalousness	Incongruity
Comedy	0.3229	0.0302	**23**
Games	0.3106	0.0258	19
Music	**0.3719**	0.0174	16
Tech	0.2723	0.0293	16
Sports	0.2848	0.0269	16
Entertainment	0.2922	0.0237	15
People	0.2969	0.0194	13
Autos	0.3100	0.0176	13
Films	0.3289	0.0153	13
Movies	0.2472	**0.0304**	13
Nonprofit	0.2683	0.0231	10
News	0.2626	0.0220	7
Animals	0.2801	0.0246	6
Education	0.2637	0.0116	2

correlation between humor and conceptual incongruity. Additionally, the ordinal list of categories in Table 3.5 also gives a listing of categories which is in close harmony with the general sense of which categories people find most interesting. Comedy, Games are usually considered more interesting than Education and Non-Profit.

Two binary classification experiments were conducted to in order to classify comedy videos from general videos. The feature space of each video consisted of two features; Diversity and Anomalousness. Two classes, namely, "Comedy" and "Others" were created, where the "Others" class contained videos from all other categories except *Comedy*.

Experiment 1, (Table 3.2), consisted of 50 videos from each class the best set of results was obtained using the Naive Bayes classifier. The fact that precision is moderate and recall is very high means that number of false positives is high, which means that most Comedy videos are incongruous but not all incongruous videos are funny. This clearly suggests that incongruity is a necessary but not a sufficient condition for humor. [33] proposed the same idea (that incongruity is a necessary but not a sufficient condition for humor) in a literary criticism of the incongruity theory of humor.

The anomalousness score was found to be zero for a majority of videos in this

dataset. Hence, in Experiment 2, 25 videos from each class "Comedy" and "Others" with non-zero anomalousness scores were considered. The impressive F-score (using Naive Bayes) indicates that videos with high incongruity identify more with Comedy. This result is analogous to the result we obtained in previous section where the most creative categories had the highest incongruity and the least creative categories had the least incongruity.

Table 3.2: Predicting Humor

Exp. Num	Class	Precision	Recall	F Measure
1	Comedy	0.69	0.94	0.79
1	Others	0.59	0.67	0.62
1	Overall	0.64	0.80	0.71
2	Comedy	0.87	0.93	0.90
2	Others	0.83	0.78	0.80
2	Overall	0.85	0.85	0.85

3.4.3 Predicting Popularity Ratings

Are videos which are incongruous in nature also more "liked" on an average than normal videos? We conducted four binary classification experiments to investigate this line of thought. Our data-set consisted of YouTube videos and we assume that the context of a video can be derived from its set of associated tags. We calculated the "Diversity" and "Anomalousness" score of each video's tag space. The ratings provided by users served as our ground truth of popularity. A video on YouTube can take any value in the range $(0, 5)$ as its average rating. We created two classes, one where the average rating was lower than 3.5 (Low class) and one where the average rating was above it (High class).

Experiment 1 (results shown in Table 3.3) was conducted on 184 videos equally sampled from low and high classes. The feature space consisted of merely two features, diversity and anomalousness. The best set of results was obtained using the Naive Bayes algorithm. Notice the high recall value for the *High* class. Incidentally, a high recall value was observed for the high class using most of the learning techniques. Again, the fact that precision is low and recall is very high means that number of false positives is high, which means that most highly rated videos have high diversity and anomalousness values but not all lowly rated videos are low on these two values. This and an analogous

observation made in Experiment 1 of previous section strongly suggest that incongruity is a necessary but not a sufficient condition for the presence of interestingness.

The anomalousness score was found to be zero for a majority of videos in this dataset. Hence, to remove this bias, in Experiment 2, we considered a subset of 50 videos each with non-zero anomalousness scores for both classes. Our best set of results eventuated from the use of a k-nearest neighbor classifier (k=3) and the impressive numbers demonstrate once again that high incongruity correlates strongly with high interestingness (in this case high average rating).

In Experiment 3, we investigated the performance of state of the art text classification technique on this dataset (n=184). The classes and labels were the same as described in Experiment 1. The feature set used by us is the well-known bag-of-words model. We created a $Video * Tags$ matrix which captures the co-occurrence of tags in videos. The data set had 1522 unique tokens, hence the feature set size was 1522. We then used maximum entropy based classifier to obtain the results as shown in Table 3.3. It is worth noting that numbers (for the High Class) shown in Experiment 1 and 3 are quite comparable, but Experiment 1 merely uses two features (diversity, anomalousness) proposed by us, whereas Experiment 3 uses 1522 features. This shows that with that the use of a theory driven feature set like ours could give comparable results to a very rigorous data dependent technique, because it captures the intrinsic phenomenon behind interestingness.

Finally, in Experiment 4, we investigated if the incorporation of our feature set (diversity, anomalousness) to the bag-of-words classifier improves the classification accuracy further (n=184). We simply appended our feature space to the $Video * Tags$ matrix described above and conducted binary classification. As we can see, the overall F-measure has improved and it indicates that a data-driven approach could benefit with the augmentation of a theory driven approach like ours if the classification task aims at capturing an intrinsic phenomena like interestingness. The fact that a data-intensive technique (as used in Experiment 3) which would implicitly pick up all kinds of implicit signals, could still benefit with the appendage of theory driven features, has two implications. First, the theory driven features showed up among the top 5 features using most feature selection techniques. Hence, the use of a reduced feature space with the top k features (including the two theory driven features) would be more preferable over using

Table 3.3: Predicting Popularity Ratings

Exp. Num.	Class	Precision	Recall	F Measure
1	Low	0.643	0.283	0.391
1	High	0.438	0.837	0.655
1	Overall	0.586	0.56	0.52
2	Low	0.70	0.76	0.72
2	High	0.74	0.79	0.77
2	Overall	0.72	0.77	0.75
3	Low	0.743	0.913	0.82
3	High	0.887	0.685	0.773
3	Overall	0.815	0.799	0.79
4	Low	0.757	0.913	0.8277
4	High	0.89	0.79	0.8370
4	Overall	0.8235	0.8515	0.83

n features (assuming $k¡¡n$) as it would be computationally less expensive. Second, the use of theory driven features makes the decision space more interpretable. For example: The use of a reduced feature space containing features which carry semantic meaning could result in a smaller decision tree or a simpler set of rules generated by a rule based classifier which could be more easily interpreted by an analyst. Hence, overall this lead to better model building.

3.4.4 Identifying incongruous topics in untagged text corpora

The tag space of YouTube videos, is particularly well-suited for discovering relationships between concepts since the cognitive process of tagging, essentially compresses information about multimedia objects into semantic objects, which become easy to study with our technique. Could our method be extended into settings where there are no pre-existing tags? To this end, we extended our analysis towards finding interesting topics in a well-known text corpora, Enron emails [26]. We extracted 50 topics from this corpora, using techniques detailed in chapter 2. Each topic was represented by a set comprising of the 10 most probable words in its distribution. We apply the same set of techniques described above, to find the most diverse, anomalous and incongruous topics. Representative topics derived from the Enron Email corpus were about discussions on work, gas pipelines, agreements, international deals, spam, fantasy football leagues,

Table 3.4: Most Diverse Topics in Enron Corpus

Topic1	Topic2	Topic3	Topic4	Topic5
request	scott	internet	customer	lynn
user	sell	website	direct	wine
scheduled	fax	online	access	confirmed
error	free	service	contract	meter
center	phone	network	cost	description

Table 3.5: Most Anomalous Topics in Enron Corpus

Topic1	Topic2	Topic3	Topic4	Topic5
gas	power	california	refund	click
capacity	project	power	committee	free
pipeline	company	market	document	offer
social	india	summer	energy	receive
naturalgas	government	prices	prices	link

travel discussions, scheduled outage mails etc. The ones marked most diverse were discussions regarding system maintenance, communication centers, online services, utility bill payments, and post-game celebrations. The topics marked most anomalous were issues related to pipeline extensions, project dealings with Indian government, power prices during summer in California, customer issues and spam messages. The topics marked as most incongruous/interesting (combined using rank aggregation and shown in Table 3.6) were issues related to pipeline extensions, project dealings with Indian government, power prices during summer in California, customer issues and labor troubles. We conducted a small survey among the graduate students (number of students=25) in the Department of Computer Science, where we asked them to rate the interestingness of all the topics on a scale of 1-10. The Spearman correlation of the user based ratings with the incongruity scores generated by our technique was 0.64 and the inter-rater agreement among the reviewers was 0.78. These results demonstrate that our methods are technically capable of knowledge discovery even in the absence of user-generated semantic exemplars (tags), which makes them accessible to a wider range of possible text-mining applications.

Table 3.6: Most Incongruous/Interesting Topics in Enron Corpus

Topic1	Topic2	Topic3	Topic4	Topic5
gas	power	california	refund	page
capacity	project	power	committee	court
pipeline	company	market	document	law
social	india	summer	energy	labor
naturalgas	government	prices	prices	worker

3.5 Discussion

We made the following contributions in this chapter.

1. First, we provided information theoretic motivations behind incongruity eventuating to interestingness in the semantic space and also showed how the two measures designed by us for doing so were the equivalent of computing entropy and conditional entropy respectively in any given semantic space.

2. Second, we described methodologies to compute $Diversity, Anomalousness$ in a semantic space under varying degrees of noise.

3. Third, we provided empirical validation across different data-sets with differing notions of interestingness in them and also find that incongruity correlates strongly with subjective human judgment of interestingness.

4. Fourth, we empirically find that high and low incongruity classes correlate strongly with high and low interestingness.

5. Fifth, we also provide the first known empirical validation (in semantic space) of the fact, that incongruity is a necessary but not sufficient condition for interestingness.

We would like to add that the overall generalizability of our results is constrained in the granularity of information present in a given piece of text. On many occasions, lots of contextual factors like implicit social understanding, cultural norms, locations, professional terms etc. add implicit meaning to a piece of a text which is something our techniques cannot completely capture. A possible way to deal with this limitation could be by making use of available meta-data like country, location etc. along with the

text at hand. The overall big picture contributions and possible future applications of this framework have been discussed in detail in the Conclusion section.

Chapter 4

A Computational Model Of Humor

4.1 Background and Related Work

The inscription on the metal bands used by the US Department of the Interior to tag migratory birds was recently abruptly changed. The small rings, typically attached to one of the birds' legs, used to bear the address of the Washington Biological Survey, which was abbreviated "Wash. Biol. Surv." This practice continued until the agency received a letter from an irate Arkansas farmer stating: "Dear Sirs, I shot one of your pet crows, and followed the instructions you had wrote on it. I washed it, an' I biled it, an' I surved it. It was turrible! You shouldn't be making fools of people like that."
The bands now read Fish and Wildlife Service [39]. There exists a class of anecdotal stimuli that results in a uniquely and endearingly human response, humorous laughter. Yet a quantitative characterization of this class of stimuli has remained elusive. What is humor? Traditional inquiries into the nature of humor have taken the form of philosophical speculation (see e.g. [35] for a review), whereof three broad classes of theories have emerged. The first type of theory, known as superiority-based, proposes that humor results from a sudden perception of ones own eminence/superiority/competence with reference to the subject of the stimulus' referent subject. In the case of our example, this theory would suggest that our sense of unlikelihood of making the same mistake as the

poor farmer provides the source of humor in the episode. Yet, it is well-understood that a sense of superiority is not sufficient, on its own, to generate a humorous response. Further, examples abound of amusing stimuli that do not easily admit a superiority-based account (see e.g., Figure 4.1).

Figure 4.1: Do I amuse you? Why?

The second type of theory, known as relief-based, draws upon Freudian principles to argue that humor is a form of release for psychological tension. Again, it is possible to argue that extreme agita aroused by psychological tension is hardly conducive to appreciation of humor. Likewise, it is not clear how suddenly noticing a nihilistic kitten (see Figure 4.1 again) would provoke the release of heretofore unnoticed psychological tension. Furthermore, if humor were presumed to stem entirely from drawing upon a store of stimulus-independent psychological resources, their expenditure should lead to

diminishing humor response. Thus, relief-based theories predict response saturation for humor-provoking stimuli across *different* stimuli. This is unlikely to be true, since while it is clear that encountering the same stimulus repeatedly leads to humor response saturation[1] , this is certainly not the case with sequential presentations of different humor-inducing stimuli[2] .

The third type of theory, known as incongruity-based [37], is currently the dominant theoretical account of humor. Such theories consider the violation of subjective expectations to be the principal source of humor and laughter. In Figure 1, for instance, such a theory would explain humor as arising out of the incongruity between the cuteness of the kitten and the nihilism of its *weltanschauung*. In the case of the Arkansas farmer, such theories would implicate the incongruity between the implicit purpose of the expression Wash. Biol. Surv. and its misinterpreted meaning as the source of humor in the anecdote. The ability to generate such passable accounts for a very broad range of humorous anecdotes leads to the current popularity of incongruity theory as an explanation of humor.

It is not, however, without its own share of criticisms. It is considered likely, for instance, that its explanatory success is partly because the notion of incongruity is itself hard to characterize, leading to a somewhat vacuous conceptual encompassment of all kinds of humor-inducing anecdotes [40]. This conceptual ambiguity is rendered all the more problematic by the inability of incongruity-based accounts to predict humorous response. That is, while incongruity (in some operationalization or the other) can be found in most funny episodes, not all incongruous episodes are funny. Thus, while there is undoubtedly some correlation between incongruity and humor, causality is yet to be ascertained.

Partly due to the absence of compelling theoretical hypotheses, partly because of the difficulty of operationalizing such hypotheses in the real world, and partly, no doubt, through disregard of the concept as an object of 'serious' study, empirical studies on the causes of humor are vanishingly rare, with literary criticism [36] and ethnographic accounts [38] dominating the sparse set of extant studies. And yet, with the development of the Social Web, an unprecedented opportunity has presented itself. Users of social

[1] For instance, having to listen to the same joke over and over quickly reduces appreciation of its humor, though aunts seem strangely reluctant to embrace this fact.

[2] For instance, watching a funny movie, or a professional comedian's set.

media, in numbers dwarfing by orders of magnitude the sample sizes of traditional ethnographic studies, incessantly congregate at media aggregation websites to tell other people what they find funny. Given this plethora of Web 2.0 information, it is likely that a new way of studying the sources of humor may now become feasible. What situational characteristics do episodes generally acknowledged as funny share? Is it possible to characterize a class of stimuli that people are likely to find funny? In this chapter, we report promising results from the first Internet-scale empirical investigation of these questions.

4.2 Finding humor on the social web

To instantiate any theory of humor in online media, it is essential to extract episodic context from media objects. Existing multimedia literature doesn't point to any promising algorithm which could be used to do episodic content analysis of videos. This technical limitation rules out the possibility of directly analyzing video data. Fortuitously, though, many online videos have an associated set of user-generated metadata (tags, ratings, comments etc), which we believe may be extremely useful in inferring episodic context. Therefore, as an alternative to direct video analysis, in this work, we use tag clouds to indirectly derive episodic contexts in associated videos, with extremely promising results.

4.2.1 YouTube's Ecosystem and Our Dataset

YouTube.com is a well-known social multimedia network. It is one of the largest repositories of community contributed multimedia content. People contribute videos on it from all over the world on various themes and topics. Besides posting and viewing videos, users are also allowed to respond to the content posted by other users by liking, disliking, commenting, linking, embedding a given video to different social networks, posting video responses etc. Videos on YouTube.com belong to one of the pre-defined categories. These category names have been decided upon by the YouTube administrators. A user assigns his content to one of the categories during the process of uploading. We used the APIs provided by Google to crawl this website across the 14 categories shown in Table 4.1.

Table 4.1: YouTube Video Categories in our Dataset

Animals and Pets	Autos and Vehicles	Comedy
Education	Entertainment	Films and Animation
Games	Movies	Music
News	Nonprofit and Activism	People and Blogs
Sports	Science and Technology	

Table 4.2: Attributes Of A Video

Title	Description	Uploaded By
Tags	Comments	Average Rating
View Count	Thumbnails	Category Name

YouTube APIs allow sorting the videos in each category using four different criteria, namely, Published Date, Number Of Views, Relevance and Rating. We crawled 1000 videos from each of the above 14 categories based on the first three sorting criteria which amounts to 42000 videos in all. Every retrieved video had the set of attributes shown in Table 4.2.

Post retrieval, the videos with metadata containing non-English or non-ASCII characters were removed from the dataset, as dealing with them is outside the scope of this work. We noticed that the videos retrieved by sorting on published date were extremely recent with blank fields returned for views, ratings and tags for most of the videos. We also noticed that the result set which contained the most relevant videos only contained very highly rated videos. The result set which contained the most viewed videos had no such apparent selection biases and was hence selected as the final data set.

Our final dataset has 38,618 unique tags across 12,088 unique videos. The number of tags per video ranges between 0 to 85, presents the appearance of a log-normal distribution, as shown in Figure 4.3. The frequency with which unique tags occur across videos in the entire dataset presents the appearance of a power-law distribution, entirely in line with similar results universally observed in other word frequency data. This distribution's plot is shown in Figure 4.2.

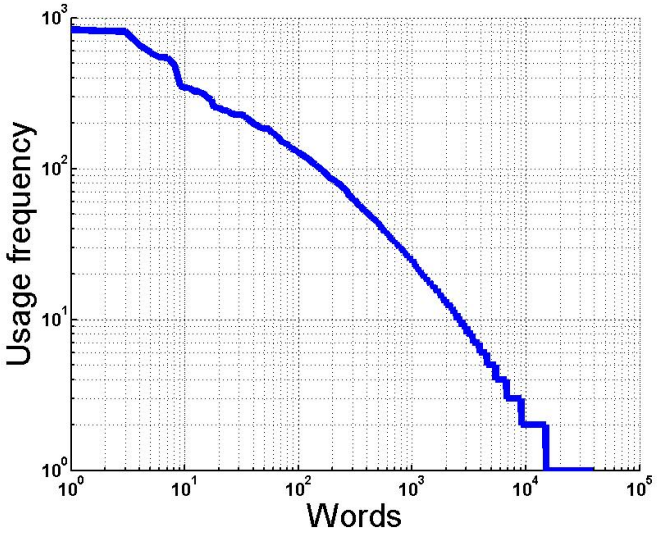

Figure 4.2: This figure shows the frequency distribution of tags in our dataset; it is well-approximated by a power law, justifying a Zipfian natural language intuition for the use of tags. The average frequency of a tag is 4; very few tags show high frequencies. Note that the x-axis logarithmically plots unique tag (word) IDs ranked in decreasing order of occurrence frequency

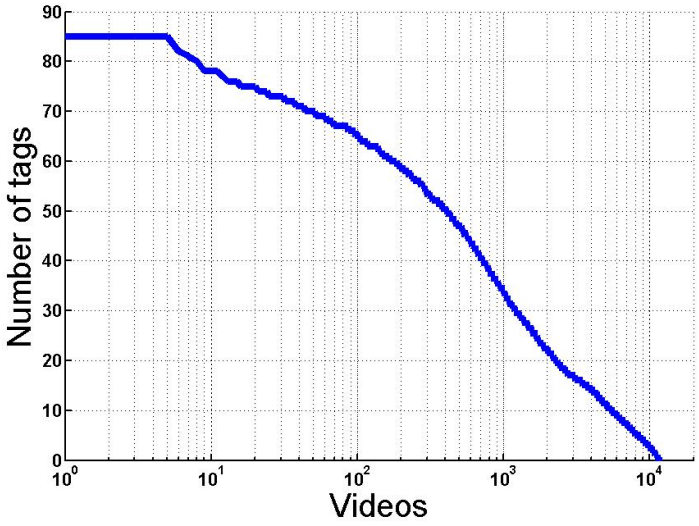

Figure 4.3: This figure shows the distribution of cardinality of tag-set in videos; it appears to be well-described by a log-normal distribution, which follows from the intuition that the use of individual tags for a video is mutually independent. Most videos have 3-5 tags; a few videos have a large number of tags. Note that x-axis logarithmically plots Video IDs ranked in decreasing order of tag-set cardinality, and that the y axis linear.

4.2.2 Meaning extraction

We assume that the context of a video V can be derived from the set of tags I accompanying it. For example, the set of tags I = "*horse*", "*breeding*", "*tutorial*", "*short*", suggests that the related video is likely to be a short tutorial about horse-breeding. Hence, to extract episodic knowledge from the tag space I about video V, we need a measure to quantify the semantic distance between any two concepts.

The idea of using a semantic measure to detect meaning has been used by researchers in the past to accomplish many information retrieval and natural language processing based tasks. [27] and [11] have defined measures of similarity for words based on e.g. information theoretic and corpus overlap criteria. While earlier efforts at extracting semantic meaning relied on the use of specific curated linguistic corpora (e.g. Wordnet), more recently, researchers have realized the value of attempting to harness the semantic connectivity of the World Wide Web itself to identify the semantic similarity of different concepts. The pioneering effort in this direction, Normalized Google Distance [5] has already been used to accomplish several semantic tasks like resolving tag ambiguities [31], mining wikipedia articles etc. [12]. Bollegala et. al. [3] have used the web to extract a different semantic relational measure which has been shown to solve cognitively sophisticated problems like analogy generation promisingly. Mahapatra et. al. [28] have recently used semantic side information to improve anomaly detection in text data by factoring in the context derived from a semantic network. An anomaly in text data means a topic which diverges from the central theme of a document, as statistically measured using topic models. This research showed the value of using semantic side information in detecting anomalousness in text data.

Meaning extraction techniques that depend on static semantic corpora lack the rich semantic connectivity that characterizes natural language interactions in cyberspace. For instance, a Wordnet-based similarity measure is unlikely to predict that 'hot' and 'cool' are semantically similar as adjectives expressing admiration. Furthermore, as is seen in Figure 4.3 and Figure 4.2, our own dataset is highly sparse, which contraindicates using statistical frequencies within itself to define similarities, as is often done in standard text-mining techniques like LDA, pLSA. Additionally, we noticed that several tags used by people were heterodox colloquialisms and net icons (e.g. lol, :) etc.) unlikely to be enumerated in standard semantic networks. Therefore, for its ability to

extract richer semantic connectivity and offer superior conceptual coverage, we base our semantic meaning extraction on the entire World Wide Web, and use the Normalized Google Distance [5] for this purpose, as we describe below.

4.2.3 Normalized Google Distance:

Given any two concepts, (x, y), the Normalized Google Distance (NGD, henceforth) between them is given by equation 5.1. $f(x)$ and $f(y)$ denote the number of pages containing x and y as returned by Google. $f(x, y)$ is the number of pages containing both x and y and N is the total number of pages indexed by Google.

$$NGD(x, y) = \frac{max(logf(x), logf(y)) - log(f(x, y))}{logN - min(logf(x), logf(y))} \tag{4.1}$$

The idea behind using Google queries is that the probability of a Google search term which is the frequency of the search term divided by the number of pages indexed by Google is approximately equal to the probability of that search terms as actually used in the society. Words, phrases and concepts acquire new meanings based on their usage patterns in the society. Also, new concepts keep getting added to the "human-communication" database(e.g. rotfl) regularly. Hence, NGD which is based on the idea of normalized information distance and Kolmogorov Complexity [5], exploits this contextual information hidden in the billions of web-pages indexed by Google to generate a sense of semantic distance between any two concepts. NGD based natural language processing experiments have shown up to 87% agreement level with Wordnet [5], which is a widely accepted standard semantic network.

4.3 Operationalizing a theory of humor in tag space

The use of semantic distance measures in tag space affords us the possibility of quantitatively visualizing patterns in semantic connectivity that typify humor-inducing stimuli. For instance, Figure 5.3 visualizes one possible deconstruction of the sad tale of the inadvertent crow gourmand using a notional set of tags and an intuitive distance measure. The intuition of incongruity is strongly borne out by the particular pattern that we observe in this visualization. However, it is equally evident that simply measuring

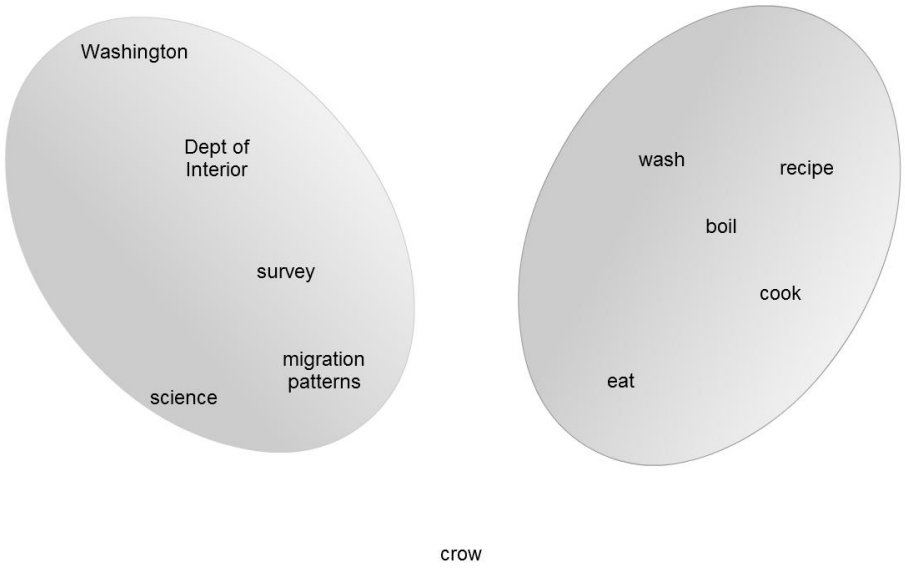

Figure 4.4: This figure shows a semantic deconstruction of our original anecdote in a notional semantic space, with notionally generated tags. Notice the presence of two disparate contexts as well as an anomaly, in accordance with predictions of incongruity-based theories of humor.

statistical anomalousness of tags is insufficient to operationalize the sense of conceptual incongruity that seems to characterize this episode. One possible interpretation of the pattern we observe, along the lines of Raskin's influential proposal [34], is that the semantic discord between disparate sets of tag clusters appears to create a conceptual dissonance, which is subsequently resolved in an unexpected manner via an anomalous bridge term (crow). Thus, in a statistical sense, we expect episodes inducing humor to contain both statistical **diversity**, allowing for the generation of anticipatory expectation and statistical **anomalousness**, allowing for unexpected resolution of the expectation.

Recently, Mihalcea et. al. [29] have attempted to operationalize a computational model for incongruity detection in humorous episodes. They formulated the problem by having a set-up line which had four possible follow ups with one of them being funny and the task was to be able to detect the funny line. They used a dataset of 150 such expert curated setup-follow up combinations. Their model used corpus related measures like LSA, semantic similarity measures based on Wordnet and other domain based heuristics like "alliteration detection" to quantify the divergence of the follow-up lines from the setup lines to detect incongruity. Conceptually, our work complements their effort, since, instead of attempting to find incongruity in humor-producing stimuli, we test whether incongruity predicts the existence of humor in episodes.

4.3.1 Calculating Normalized Google Distance

We use the NGD as given by equation 5.1 to first build a semantic space for each video V given its set of tags I. Calculating the NGD between two given concepts (a, b) requires three Google queries (queries to calculate $f(x)$, $f(y)$, $f(x, y)$). Google APIs allow maximum 1000 queries per day. This puts a slight computational restriction on us in being able to try out our model on a very large number of videos. Hence, we restricted ourselves to the top 50 most viewed videos from each of the 14 categories (700 videos in all). Our general assumption, anecdotally validated, is that the most viewed videos in any category can be assumed to be the true labels of that category, which means e.g. a very highly viewed education video must be educational and more saliently for our purposes, that the most highly viewed comedy videos must be funny.

Let a and b be two concepts and $NGD(a, b)$ be the distance between them. The

range of this measure is (0,∞). The value 0 would indicate that the two concepts are exactly the same and ∞ would indicate they are completely unrelated. Normalized Google Distance is a non-metric and hence triangle inequality doesn't hold for it. If a, b, c are any three random concepts/words/tags, then:

$$NGD(a,c) \not<= NGD(a,b) + NGD(b,c) \tag{4.2}$$

Normalized google distance is symmetric.

$$NGD(a,b) = NGD(b,a) \tag{4.3}$$

The fact that NGD is a non-metric presents a slight technical challenge as most well-known statistical techniques are defined to work only in a metric space. However, we deal with this limitation in ways we describe below.

4.3.2 Similarity in tag space

We assume that the contextual information about any given video can be derived from its set of associated tags. We analyze the non-metric concept space of tags where every mutual distance is the NGD between the two given concepts. To simplify our analysis, we first calculate a similarity matrix. Let V denote a given video, and the set $I = \{i_1, i_2, .., i_n\}$ denote the set of tags which characterize V. The cardinality of the set I is assumed to be n. Then, a $n * n$ similarity matrix was constructed. Each row of this matrix contains the NGD between any given tag and the whole tag-set I. This matrix is symmetric in nature which follows naturally from the fact that NGD is symmetric in nature (equation [4.3]). This cuts down the required number of Google queries by half, as only half of the entries in any similarity matrix need to calculated.

4.3.3 Defining Diversity

As explained earlier, the first thing we need to detect is the presence of statistical diversity, indicating the presence of conceptual dissonance in the video episode. In statistical terms, this will be indicated by greater variation in inter-tag semantic distance. We quantify this intuition in the form of a measure we call "Diversity". To operationalize this measure, we first calculate the similarity matrix for the tag cloud of each video

as described above. Each row of the similarity matrix denotes the semantic divergence of each tag from the entire set I. To handle the non-metric nature of this similarity measure, we use the $L1$ norm to measure variation. Diversity in this sense can be measured in the form of least cumulative absolute deviation within a tag set. Least absolute deviation is also known to be robust to the presence of noise and outliers in data and can hence deal with noisy tag clouds, and is therefore adequate for our purpose. Let's say the data consists of a set of points (x_i, y_i) with i = 1, 2, ..., n. We assume that there is linear function $f(x)$ of the form: $y = mx + c$, which approximates the values y. We now seek out the values of the unknown parameters m and c which would minimize the sum of absolute deviations. The value of n, which is cardinality of the tag set, has a very high range $(1, 85)$ in our data set, which is likely to result in scaling effects in the least absolute deviation measure. Hence, we normalize the value of least absolute deviation by the number of tags n to obtain our final measure of diversity in tag space.

$$S = min(\sum_{i=1}^{n} |y_i - f(x_i)|) \tag{4.4}$$

$$NLAD = S/n \tag{4.5}$$

NLAD stands for normalized least absolute deviation.

4.3.4 Defining anomalousness

The second aspect we need to quantify is the presence of a semantic anomaly in tag space. We detect the presence of anomalous tags in a set of tags as follows:

1. First we calculate, the sum of deviations of each tag from all the other tags in a given set of tags I. Let the similarity matrix be denoted by Sim where each entry $Sim(i, j)$ denotes the distance between $tag(i)$ and $tag(j)$. The deviation score d_i for the i_{th} tag is equal to:
$$d_i = \sum_{j=1}^{n} Sim(i, j) \tag{4.6}$$

2. After obtaining the deviation scores d_i for all the tags, the problem reduces to finding the most deviant points from this set. Again, in order to handle non-metric distances, we use the inter-quartile measure to accomplish this, which is

well-known in the anomaly detection literature [4]. $Q1$ denotes the first quartile, $Q3$ denotes the third quartile and IQR denotes the interquartile range. Any points lying outside the ranges specified by equation 4.7 and 4.8 are counted as anomalies. UB and LB denote the upper and lower bounds respectively.

$$LB = Q1 - 1.5 * IQR \tag{4.7}$$

$$UB = Q3 + 1.5 * IQR \tag{4.8}$$

Let $n1$ and $n2$ denote the number of points outside the LB and UB respectively.

3. As the number of tags per video is highly variable, we normalize the anomaly score of a video by the number of tags.

$$Score = (n1 + n2)/n \tag{4.9}$$

4.3.5 Perception of Humor

As we describe in the deconstruction of our original anecdote, we intuitively expect semantic incongruity to be directly correlated with both diversity and anomalousness. To quantify this intuition, we used a rank-aggregation technique to combine both our empirical measures. The use of rank aggregation instead of algebraic combination is preferred since it allows us to avoid concerns about parameterization and scaling affecting our ultimately rank-based conclusions. We can safely assume that diversity and anomalousness have been used to the rank the categories independently. This allows us to use Borda's Technique [30], a simple positional rank-aggregation technique. Given a set of ranked lists, $T = \{t_1, t_2, .., t_n\}$, it assigns a score $S_i(k)$ to each candidate in t_i which is simply the number of candidates ranked below it in that list. The net score of every element is the sum of all the scores generated across the entire set of ranked lists. The scores are then sorted in decreasing order to give the final list of highest ranking categories. Thus, for instance, a category that is the 3^{rd} most diverse and the 6^{th} most anomalous out of our 14 categories would have a Borda incongruity score of $11 + 8 = 19$.

4.4 Results

First, we calculated the NLAD score for each video across each of the 14 categories using equations 4.4 and 4.5. Then, we conducted a large sample one way analysis of

variance test under the null hypothesis that all samples (diversity scores) in data set D are drawn from populations with the same mean. We obtain extreme statistical significance $p = 0.000026$, decisively rejecting the null hypothesis. A few interesting observations based on the above statistical test are as follows:

1. *Music*, *Film*, *Comedy* and *Games* present as the most diverse categories.
2. *Education*, *Movies*, *News* and *Nonprofit* are the least diverse categories.
3. Categories like *Science & Technology* and *Autos & Vehicles* show the least variance in their deviations, a likely consequence of substantial overlap in the tag clouds in these categories.

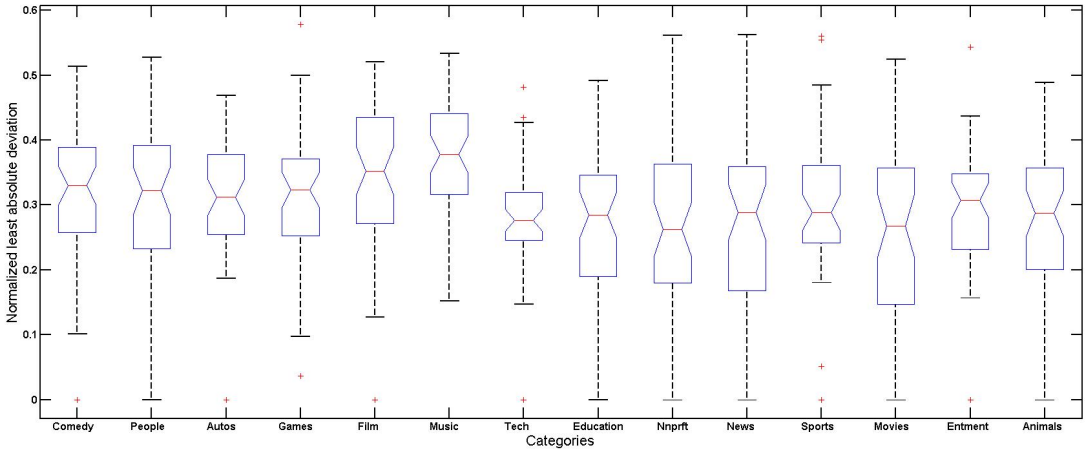

Figure 4.5: Anova Plot: x-axis denotes the categories, y-axis shows the least absolute deviation variable width boxplot.

The observations from this experiment complement a natural intuition of assigning greater creative possibilities to categories that contain a more diverse set of concepts. A recent field study from Srinivasan et. al. [43] arrives at the same conclusion, supporting the basic validity of our statistical measure of diversity. We then calculated anomalousness scores for each video in each category using equations 4.7, 4.8, and 4.9. Since the use of this particular technique for detecting anomalies in non-metric spaces is

Table 4.3: Mean diversity, anomalousness and incongruity scores for all 14 video categories ranked in descending order of incongruity

Category	Diversity	Anomalousness	Incongruity
Comedy	0.3229	0.0302	**23**
Games	0.3106	0.0258	19
Music	**0.3719**	0.0174	16
Tech	0.2723	0.0293	16
Sports	0.2848	0.0269	16
Entertainment	0.2922	0.0237	15
People	0.2969	0.0194	13
Autos	0.3100	0.0176	13
Films	0.3289	0.0153	13
Movies	0.2472	**0.0304**	13
Nonprofit	0.2683	0.0231	10
News	0.2626	0.0220	7
Animals	0.2801	0.0246	6
Education	0.2637	0.0116	2

well-supported by extensive anomaly detection literature [4], no further substantiation of the calculation is felt necessary. The mean anomalousness and diversity scores of the 14 categories are shown in Table 4.3. We notice that intuitively 'exciting' categories like *Comedy* and *Movies* are high in anomalousness while 'mundane' categories like Education and Autos score low.

Finally, we compute Borda scores combining diversity with anomalousness for each category to obtain a final list of categories ranked by measured conceptual incongruity, as shown in Table 4.3. We see that our empirical results are in perfect concordance with the incongruity-based theories of humor, as the category "Comedy" ranks the highest in the combined list. The two dimensional feature space which maps these categories is shown in Figure 4.6. As we can clearly see, *Comedy* is the only category which scores high on both the measures.

Further, the ordinal list in Table 4.3 suggests an increasingly lower possibility of finding humor in games than comedy, music than games, and least of all in videos about animals, news stories and educational videos. Anecdotally, such a listing appears to be in relatively close harmony with the general sense that people have of the relative funniness of these categories of media. Unfortunately, more rigorous validation of this

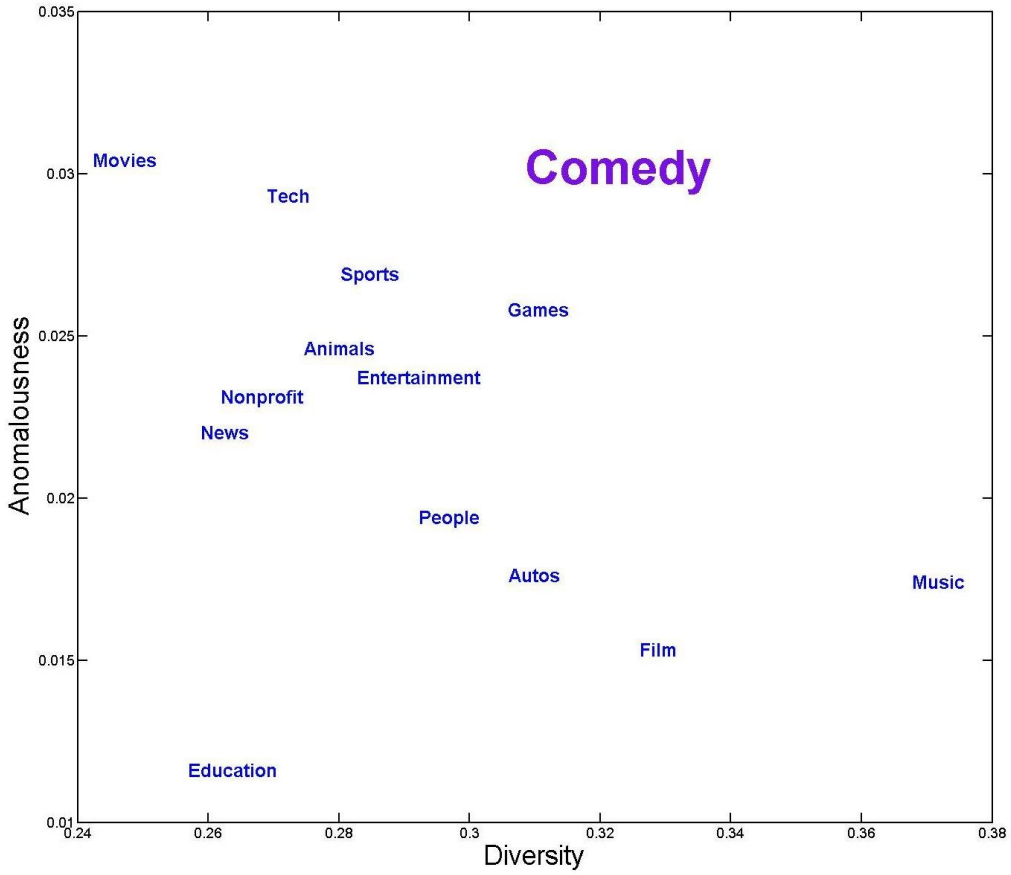

Figure 4.6: This figure maps all the 14 Video categories to the two dimensional feature space {Diversity, Anomalousness} generated during the course of our analysis. Notice that Comedy (shown in bold) is the only category that scores high on both measures.

observation lies outside the scope of this work.

4.4.1 Predicting Humor

Can the above mentioned technique of incongruity detection be used to predict humor? We conducted two binary classification experiments to investigate this line of thought. We used the Weka APIs to conduct all our experiments [42]. Note, that in all the classification experiments, equality of class distribution was obtained by randomly down-sampling the majority class. The feature space of each video consisted of two attributes $Diversity_{score}, Anomalousness_{score}$. We created two classes, namely, "Comedy" and "Others" where the "Others" class contained videos from all other categories except *Comedy*.

In the first set of experiments (results shown in Table 4.4), we attempted to discriminate the "Comedy" class from the "Others" class, by considering 50 videos from each category, and the best set of results were obtained using the Naive Bayes algorithm. We observed a high recall value for the "Comedy" class using most of the learning techniques. The fact that precision is moderate and recall is very high means that number of false positives is high, which means that most Comedy videos are incongruous but not all incongruous videos are funny which indicates that incongruity is a necessary but not a sufficient condition for humor.

However, note that the anomalousness score is zero for a majority of the videos in our dataset. To control for this potential biasing factor in the anomalousness feature, we conducted a second set of prediction experiments. We considered 25 videos each with non-zero anomalousness scores in both "Comedy" and "Others" classes. Our best set of results eventuated from the use of a Naives Bayes classifier, and are shown in Table 4.5. The discriminability achieved (though on a very small data set) is quantitatively impressive with overall F score of 0.854. These results suggest that videos with high diversity and anomalousness measures strongly predict online community identification as *Comedy*.

4.5 Discussion Of Results

In this chapter, we make the following contributions:

Table 4.4: Experiment 1

Class	Precision	Recall	F Measure
Comedy	0.69	0.94	0.7950
Others	0.59	0.67	0.6275
Overall	0.64	0.805	0.71

Table 4.5: Experiment 2

Class	Precision	Recall	F Measure
Comedy	0.875	0.933	0.903
Others	0.833	0.781	0.8062
Overall	0.854	0.857	0.854

1. We illustrated a reliable method of extracting the semantic sense of a multimedia object from user-defined tags, leveraging the semantic connectivity of the entire Web visible to Google's search crawlers.

2. We empirically show that a combination of semantic diversity and anomalousness, as measured in the space of related tags, strongly correlates with the categorization of videos as humorous on YouTube, as evidenced by high recall in category prediction.

3. We find, however, that in aggregate, this combination of semantic diversity and incongruity is a relatively weak predictor of humor in videos, as evidenced by low precision scores.

4. In light of (2) and (3) we interpret our findings as the first large-scale empirical evidence of the fact that incongruity is a necessary, but not sufficient, condition for media objects to appear humorous.

We would like to add that the generalizability of our results and their interpretations towards forming a comprehensive account of humor is necessarily constrained by the ecological specificity of our dataset. YouTube's own understanding of the causes of video 'virality' suggests [41] that videos become interesting to masses of users through a combination of influential referrals, community participation and content unexpectedness. The last factor harmonizes perfectly with our instantiation of the idea of semantic incongruity. The other two aspects are social network effects that cannot be addressed

in our analysis. Hence, it is likely that an important reason for many videos with high incongruity to remain poorly rated is the absence of positive network effects, e.g. lack of influential referrals, poor community connectivity of the video uploader etc. A further source of noise, which we have neglected to account for in our present analysis, are false positives introduced by the use of technical jargon in tagging, which will result in both high diversity and anomalousness scores, without truly being incongruous in the semantic sense. It is likely that including network effects and removing unique technical tags will result in substantial improvements in the precision of the predictions from our method.

Finally, as Richard Feynman justly observes, the highest forms of understanding we can achieve are laughter and human compassion. Perhaps not coincidentally, these most of all human qualities strongly elude scientific characterization. For this reason, we find considerable aesthetic pleasure in this demonstration of the power of computing to extract a deep conceptual sense of the world from the Web itself and improve, if by an iota, our understanding of one of the deep mysteries of the human condition.

Chapter 5

Incongruity versus Incongruity Resolution?

5.1 Background and Motivation

1. *Three vampires are sitting at a bar. Bartender asks the first one what he wants. "I think I'll have a glass of blood. "Okay, what'll you have? he asks the second vampire. "That sounds good. I'll have a glass of blood too. "And what can I get for you? he asks the third vampire. "I'll have a glass of plasma said the third vampire. "Okay, said the bartender, "That's two bloods and a blood light, then."* [54]

2. *Why do birds fly south in winter, because it's hard to walk.*

3. *If the customer is always right, why isn't everything free?* Well known joke taken from [39].

It is well understood that there exists a class of stimuli which makes human subjects laugh. But what characterizes this class of stimuli, what is humor? Studies in the field of psychology and philosophy have proposed four broad classes of theories (See [35] for a review). The first class of theories, namely, superiority theories propose that humor

results from a subject's perception of his own superiority over the protagonist in the joke. The second class of theories, known as the relief theory, proposes that humor is a mechanism for release of psychological tension and pent up energy. It is worth noting that this theory is silent on the causal links between relief and humor. The third class, namely play theory, proposes that humor is simply an extension of animal play. It suggests that similarities in animal behavior between activities considered play and humor, suggest they are one and the same activity. The fourth class, namely incongruity theory [37], also currently the most dominant theoretical account of humor, postulates that humor is caused due to the playful violation of the subjective expectations of an agent; caused by the presence of incongruous stimuli in the agent's observations. For instance, in the first example humor arises due to the incongruous interpretation of the word plasma by the bartender leading to an unusual remark. In the second example, humor arises due to the incongruous use of "walk" with birds in the punch line. Similarly, in the third example, the incongruous association of "right" with "free" leads to humor. It is this ability to generate explanations for a large class of humorous stimuli, which leads to widespread acceptance and popularity of the incongruity theory.

"Incongruity Theory" encompasses a wide class of theoretical accounts which overall provide an explanation for a large class of humor inducing stimuli but most of these accounts differ in the details of their explanation. Some studies like, [44] suggest that the main point of this theory is that humor is not so much about the presence of incongruity itself but its realization and eventual resolution by a competitive explanation (For example, the inscription in first joke had two competitive explanations), hence they are called incongruity resolution theories. Some other studies [33] propose the likelihood of two valid mechanisms, first being the resolution of incongruity and the second one being the mere presence of a playful incongruity which can be eventually traced back to the presence of humor (For example, the mere suggestion of birds "walking"). Previously, there have been quite a few qualitative and quantitative studies (philosophical [33], computational [47], [29]) which have reported a strong correlation between incongruity and humor inducing stimuli, and have suggested that incongruity is most certainly a necessary (but not a sufficient) condition for humor. The question, as to which of the above two mechanisms provides a more general explanation of humor, is still an open question.

There are two possible hypotheses: one that humor arises as a motivational signal to pursue explanations for cued incongruity (incongruity detection); two, that humor arises as a satiety signal indicating the resolution of the cued incongruity (incongruity resolution). It is straightforward to find examples of humorous anecdotes that support either hypothesis. For instance, our first joke can be construed to contain an element of resolution, with the word plasma offering competing explanations from different contexts (The words plasma interpreted as a component of blood and as being related to light). In contrast, in the second case, a suggestion of birds walking in winter presents as a straightforward example of humor inducement purely through the perception of incongruity. If both possibilities are to be admitted as valid mechanisms, then a natural next question arises from there. Do these two mechanisms explain certain classes of humor inducing stimuli? In this chapter we report promising results to answer this very question.

This chapter presents an algorithmic operationalization of the two above mentioned mechanisms of humor. The scope of our current work is limited to text data. First, we provide a brief literature review of the work closely related to ours and then put forth the formal arguments behind our approach. We then describe an algorithm for measuring conceptual incongruity of a piece of text and then finally present empirical results on three different datasets. We conclude with a short discussion of our approach, results and future implications. As per the best of our knowledge, this work presents the first known quantitative comparison between incongruity and incongruity resolution, as possible mechanisms of humor.

5.2 Related Work and Formalism

5.2.1 Literature Review

There have been quite a few instantiations of the incongruity theory of humor in the domain of text data. The most influential of which is by Raskin [34] whose "General Theory Of Verbal Humor" (GTVH) proposes that it is the presence of opposing words like good/bad, pretty/ugly etc. inside the same piece of text which creates a semantic discord that ultimately leads to humor. [36] presents a remarkable literary criticism of the GTVH and puts forth the idea of presence of incongruity clusters rather than

stand-alone incongruities as a possible pattern, which explains a lot of jokes. The work most relevant to our investigation, [33] presents a literary criticism of the incongruity theory of humor and brings forth two important ideas, first that incongruity is a necessary but not a sufficient condition for humor, and second that both the mere presence of incongruity and its eventual resolution are likely explanations of humor generation for different kinds of jokes. In a work closely related to ours, [29] have attempted to operationalize a computational model for incongruity detection in humorous episodes. They formulated the problem by having a set-up line which had four possible follow ups with just one of them being funny and the task was to be able to detect the funny line. They used a dataset of 150 such expert curated setup follow-up combinations. Their model used corpus related measures like LSA, semantic similarity measures based on Wordnet and other domain based heuristics to quantify the divergence of the follow-up lines from the setup lines to detect incongruity. Conceptually, a work which complements their effort was done by us in chapter 4, where instead of attempting to find incongruity in humor inducing stimuli, we test whether incongruity predicts the existence of humor. Our dataset consisted of 700 most highly watched videos from YouTube from 14 different video categories, some of which happened to be funny. We introduced an algorithm to detect incongruous episodes in the "tag-space" of a given video and eventually they reported a stronger presence of incongruity in comedy videos than in any other kinds of videos like News or Music. We also attempted to classify comedy videos from general videos and conclude that semantic incongruity is a necessary but not a sufficient condition for humor. Their empirical results support Veale's proposal [33] that incongruity is a necessary but not a sufficient condition for humor.

Overall, all the above studies, have found a strong correlation of humor with incongruity and vice-versa, but they leave open the question whether humor is truly caused by the mere presence of incongruity or its eventual resolution. Our work in the subsequent sections will introduce an algorithm to investigate this very line of thought and reports some intuitively promising results on various data-sets.

5.2.2 Central Research Questions

Any piece of text can be represented by the set of words W constituting it. Without any loss of generality, let us assume that we can represent these words in a notional

semantic space where the distance between any two words, is simply the similarity measure between any two words based on their co-usage in the universe. The anecdotes presented in section 1 have been displayed in one such notional semantic space (Figures 1 and 2).

Figure 1, shows the presence of a statistical anomaly (plasma), which is far away, from two closely knit clusters and hence seems to resolve the conceptual dissonance introduced by these disparate clusters. Quantitatively, this pattern in a given semantic space should correspond to the qualitative idea of an incongruity (like plasma) being introduced by one cluster (a cluster of words which make sense together); which eventually gets resolved by another cluster. In effect, jokes which are easily explained by "incongruity resolution" should have a higher statistical frequency of displaying the above pattern. On the contrary Figure 2, shows one tightly knit cluster and the presence of just one statistical anomaly (walk) adding the necessary incongruity in the humor inducing stimuli. Quantitatively, this given pattern in a semantic space should correspond to the qualitative idea of the presence of a plane incongruity being capable of adding humor to a text. In effect, jokes which are easily explained by incongruity presence should have a higher statistical frequency of displaying this given pattern.

Intuitively, it seems that both these can be admitted as valid mechanisms and the presence of these patterns should vary from joke to joke. Do both these valid mechanisms explain different classes of jokes? If yes, which class of humor inducing stimuli is better explained using incongruity detection and which class of stimuli is better explained using incongruity resolution? Is this a generalizable result? In order, to answer these very questions we developed an algorithm which was can detect the above mentioned patterns (Figure 1 and Figure 2) in the semantic space of a joke. A more succinct description these patterns are given in the following section. We eventually tested it on three different datasets, which are briefly described below.

5.2.3 Datasets

We used three different kinds of data-sets in our empirical analysis. Our first data-set, introduced by [29], consists of 150 funny one-liners as described earlier. Typically, a one-liner is a short sentence which produces comic effects with the use of interesting word play, unusual adjectives, rhyme, alliterations etc. and usually lacks a complex

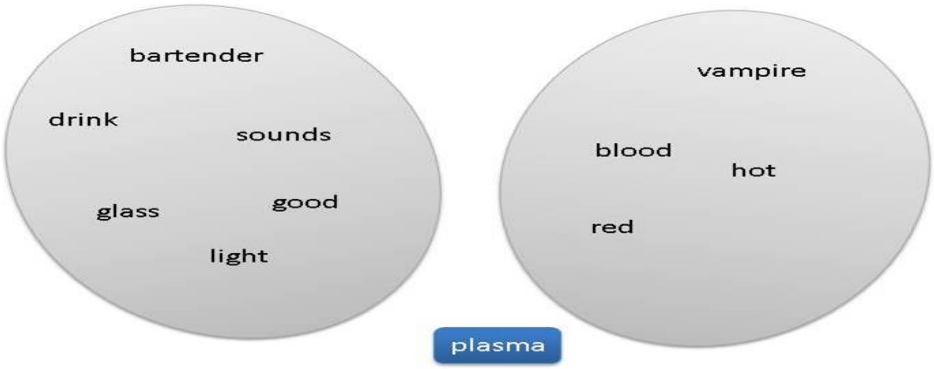

Figure 5.1: This figure shows a semantic deconstruction of our first anecdote in a notional semantic space. Notice the presence of two disparate contexts as well as an anomaly, in accordance with predictions of incongruity resolution theories of humor.

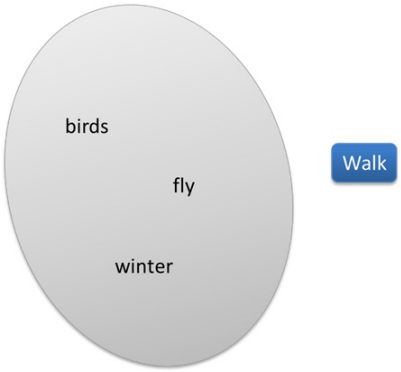

Figure 5.2: This figure shows a semantic deconstruction of our second anecdote in a notional semantic space. Notice the presence of one tightly knit cluster as well as the presence of one anomaly, in accordance with predictions of presence of incongruity leading to humor.

narrative structure or too many hidden meanings. This observation was confirmed by the description provided by the authors and our manual inspection of the data-set. Based on the above evidence, we label this dataset as a representative set of "simple stimuli" which produce a humorous response.

Our second data-set was introduced by [54]. [54] conducted one of the largest known scientific surveys to find the funniest jokes from all over the world. We picked up a subset of 50 jokes as our working set in the following manner. Firstly, we picked the top joke for each country and then picked up a random sample from the rest 1001 top jokes. As mentioned by [54], most of the jokes which are regarded as "top" jokes of an entire community usually have a complex narrative structure and the source of fun could be attributed to multiple reasons. For example, our first anecdote, a very highly rated joke, contains many elements of humor like the presence of the incongruous inscription, presence of multiple competitive explanations of the situation etc. This observation was confirmed by the description provided by the authors and our manual inspection of the data-set. Based on the above evidence, we can label this dataset as a representative set of "complex stimuli" which produce humorous response.

Our third data-set, introduced by us in chapter 4, consisted of the tag-sets of the 700 most watched videos from 14 different video categories of YouTube which also includes the category "Comedy". More details regarding process of data collection, cleaning etc. can be found in chapter 2. We show empirically that the comedy was the most incongruous of all video categories. As the purpose of our investigation is not to compare "Comedy" with other categories but rather to compare two different mechanisms on a given humor inducing stimuli, we just selected the 50 videos from the comedy category as our dataset. Unlike the above two data-sets, the tag sets of these videos represent very different kinds of videos, and hence represent a very heterogeneous set of stimuli. We cannot categorize these videos into a specific class of humor like simple or complex. The objective of our investigation on this dataset is purely to explore the generalizability of our technique on noisy online data.

5.2.4 De-constructing Semantic Space Using Normalized Google Distance

In order to detect the presence of a statistical anomaly, a cluster of anomalies, multiple clusters or any other such pattern in a piece of text, it is necessary to use a semantic measure which can spit out the similarity between two words in the corpora. We decided to use "Normalized Google Distance" (NGD henceforth) [5] for reasons described earlier. Given two terms (could be words, phrases or clauses) x and y the NGD between them is given by equation 5.1. $f(x)$ and $f(y)$ denotes the number of pages returned by Google containing x and y respectively and $f(x, y)$ denotes the number of pages containing both (x, y) . N denotes the total number of pages indexed by Google. [1]

$$NGD(x, y) = \frac{max(logf(x), logf(y)) - log(f(x, y))}{logN - min(logf(x), logf(y))} \tag{5.1}$$

The range of this measure is (0,∞) where 0 indicates complete similarity between two terms and ∞ indicates complete dissimilarity.

5.2.5 Dealing with non-metric space using a similarity matrix

NGD is a symmetric measure. (Equation 5.2)

$$NGD(x, y) = NGD(y, x) \tag{5.2}$$

NGD is not a metric, which means the triangle inequality (Equation 5.3) doesn't always hold. This presents us with a slight technical challenge because most statistical and machine learning techniques only work on metric spaces. Hence, in order to get past this hurdle we created a similarity matrix and only made use of techniques which are agnostic to metric space assumption.

$$NGD(x, z) <= NGD(x, y) + NGD(y, z) \tag{5.3}$$

The similarity matrix S for a given joke was built as follows. Let W denote the set of words constituting the joke $W = \{w_1, w_2, ..w_N\}$ where w_i is the i_{th} word. The cardinality of the set W is assumed to be N. We then construct a N*N matrix where each entry of the matrix is the NGD between any two words (w_i, w_j).

[1] We used wikipedia as our corpus for the first two data-sets (150 one liners and Laughlab Dataset) and Google for the third data-set (YouTube Videos, as it was highly noisy).

5.3 Operationalization of incongruity theories of humor

In order to answer the question; is humor better explained by the mere presence of incongruity or its eventual resolution quantitatively, we need to be able to detect patterns similar to those displayed in Figures 1 and 2 in the semantic space of a joke. Jokes whose comic effect is better characterized by the mere presence of an incongruity should contain at least one statistical outlier and a single tightly knit cluster (Figure 1), which corresponds to the qualitative explanation that, humor is produced by the presence of a semantic incongruity in an otherwise perfectly reasonable statement. Let us say, the number of outliers in a joke is denoted by A and the number of clusters is denoted by C. $(A > 0)$ denotes the event that there exists at least one anomaly per joke and $(C = 1)$ denotes the event that there exists just one cluster per joke(clustering is done after removing the anomalies)[2] . The probability of presence of incongruity, denoted by $P(I_{Presence})$ is given by the conditional probability shown in equation 5.4. We substitute, the sample estimates of $(C == 1)$ and $(A > 0)$, to estimate this conditional probability.

$$P(I_{Presence}) = P(C = 1/A > 0) \tag{5.4}$$

Jokes whose comic effect is better characterized by resolution of the incongruity should contain at least one statistical anomaly and should contain at least more than one cluster (Figure 2), which corresponds to the qualitative explanation that, humor is produced when the incongruity observed in one explanation is eventually resolved by another competitive explanation. Using the above notations, we can say that the probability of this pattern, denoted by $P(I_{Res})$ is given by equation 5.5.

$$P(I_{Res}) = P(C > 1/A > 0) \tag{5.5}$$

In order to estimate the above conditional probabilities, we first need to estimate the presence of statistical anomalies and estimate the number of clusters in the semantic space of a joke. First, we explain our rationale behind conducting these two tasks and then we sketch out the overall incongruity characterization algorithm.

[2] Clustering techniques work better once the anomalies are removed [4]

5.3.1 Detecting Statistical Anomalies Using Local Outlier Factor

In order to detect the patterns shown in Figures 1 and 2, in the semantic space, the first thing we need to detect is the presence of statistical anomalies. We used a well-known density based anomaly detection algorithm called "Local Outlier Factor" (LOF henceforth) [51] to achieve this. The key idea behind LOF is the concept of local density, where the locality of a point is defined by a set of its nearest neighbors, whose distance from the given point, is used to measure a point's local density. The ratio of local density of a point to the local density of its neighbors can help us identify regions of similar density and regions of substantially lower density. Points with substantially lower densities than its neighbors are eventually labeled anomalies. Figure 3 illustrates the main advantage of this technique graphically. Point p2, would be labeled "normal" by standard nearest neighbor based techniques, but LOF would correctly label it as an outlier. LOF has been successfully applied in the past, in many domains like network intrusion detection and information retrieval [52]. The LOF algorithm is sketched briefly below.

1. Firstly we define the $k$distance(X) of a point X, to be the distance of the point X to its k_{th} nearest neighbor. In case of ties, there could be more than k nearest neighbors. Let the set of k nearest neighbors of point X, be denoted by $S_k(X)$. $d(X, Y)$ denotes the distance between any two points. Based on the above definitions, the reachability distance (rdistance henceforth) between any two points (X, Y) is computed as follows:

$$rdistance(X, Y) = max(kdistance(X, Y), d(X, Y)) \tag{5.6}$$

2. Secondly, we compute the local reachability density (lrd henceforth) of a point X as follows. $|S_k|$ denotes the cardinality of set S_k.

$$lrd(X) = \frac{1}{\frac{\sum_{Y \in S_k(X)} rdistance(X,Y)}{|S_k|}} \tag{5.7}$$

3. Finally we compare the lrd of a given point with its neighbors to compute its LOF factor. An LOF factor significantly higher than 1 denotes an outlier, a value close

to 1 denotes a normal point and value lesser than one denotes a dense locality.

$$LOF(X) = \frac{\sum_{Y \in S_k(X)} \frac{lrd(Y)}{lrd(X)}}{|S_k|} \tag{5.8}$$

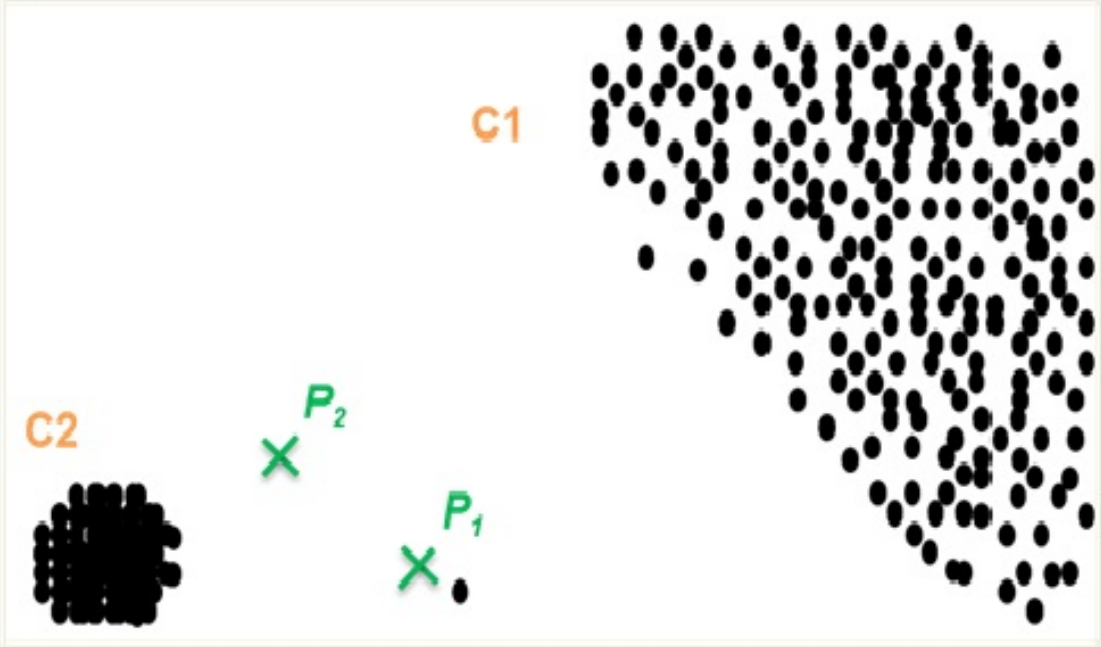

Figure 5.3: This figure shows a notional feature space highlighting the density based rationale of LOF. Notice that point p2 due to its proximity with the lower cluster will not be flagged anomalous by most nearest neighbors based techniques. LOF will flag it as an anomaly correctly. This underscores the importance of a density approach especially in our case, as words might form clusters of completely varying density.

5.3.2 Estimating The Number Of Disparate Clusters

After detecting the statistical anomalies and removing them, the next thing we need to do, is the estimation of number of clusters in the semantic space. Finding the number of clusters in a given data-set is a traditionally ambiguous problem, and the machine learning literature points to many interpretations and solutions. We used the "Monte Carlo Cross Validation" technique. It fits a Gaussian mixture model to the data using Expectation Maximization [53] and then measures the divergence between true

probability density and the new probability density post clustering into k components. It then picks the value of k which results in minimum divergence. Let $f(x)$ denote the true probability density of x. Let us say we divided the data-set into two partitions, training and test. We learn a finite Gaussian mixture model with k components using maximum likelihood estimate technique on the training set. We denote the learned set of models as: $f_k(x/\phi_k)$ for a given value of k. The log likelihood for the k_{th} model is defined as follows (assuming the test set is independent of training set):

$$L_k(Test_{Data}) = \sum_{i=1}^{N} log f_k(x_i/\phi_k) \tag{5.9}$$

Equation 5.10 [53], gives an estimate of the expected log likelihood on the test set which is agnostic to the independence assumption between test and training set (a more likely real world scenario) where C is a constant and E stands for expected value.

$$E[L_k] = -E[\int f(x) * log(\frac{f(x)}{f_k(x_i/\phi_k)})dx] + C \tag{5.10}$$

The first term on the RHS is simply the KL-Divergence between mixture model and the true density. It has been shown to be a function of parameter k only. Hence, it serves as an estimator for comparing mixture models with different components. Thus, the value of k which minimizes this divergence serves as the closest estimate of the number of clusters.

We then carry out cross validation to avoid over fitting. The data is divided into M folds and the Gaussian mixture model with parameter k is learned using $M - P$ folds. The rest P folds are used to calculate the divergence which serves as a measure for the goodness of fit. The key distinction between normal cross-validation and Monte Carlo cross validation is that different test subsets are chosen randomly which are not always completely disjoint from the training set. This technique has shown better results than most popular estimation techniques (estimation techniques for number of clusters in data) like AIC, BIC etc. on both real and synthetic data sets. [53]

5.3.3 The Algorithm

After covering the important subroutines in the above two sections we now formally lay out our algorithm. For a given joke we first removed all the stop words and punctuations.

Then we passed the text through a stemmer (we made use of porter-stemmer), and then removed the duplicate stem words. Finally we were left with a set of N keywords, which represents the meaning of the joke. Thereafter, we constructed a $N*N$ similarity matrix as explained above and fill it with the pairwise NGD of two words. Then we use this pairwise matrix to find the LOF factor of each point in the matrix. The points above a parametric threshold *thresh* were removed from the matrix W and added to set of anomalies A. The resultant matrix is then subjected to Monte Carlo Cross Validation to detect the number of clusters. The results obtained on the different datasets are discussed below. *Note: All the above measures were normalized properly to nullify the effect of number of words per joke..*

We hasten to add that the efficacy of our technique relies heavily upon two factors. First, is the threshold chosen while calculating LOF. Even mild variations in the threshold could result in slightly different statistics. We used just one threshold (which may not be a wise choice from the anomaly detection stand point) for all our data-sets to facilitate a fairer comparison among them. Second, the efficacy of estimation of number of clusters using Monte Carlo Cross Validation depends on the random generator used by the program. We used Matlab's random number generator to conduct the Monte Carlo cross validation. We have observed that upon using other packages, like C++ STL or numpy the distribution of number of number of clusters varies slightly but not too much.For example, for the 150 one-liners data-set, on using matlab the number of jokes with one cluster was found to be 148 but it was found to be 150 on using numpy.

5.4 Results

5.4.1 Testing Readability

Several times, the usage of specialized, infrequent and hard to read words, could give the false impression of the presence of an anomaly, in the semantic space. For example, in the following set of words "liposuction, procedure, doctor, documentary", the first word would be flagged as an anomaly as it is not frequently co-used with rest of the words in the society. But, in reality these words make perfect sense together[3] . Hence, the

[3] These words most likely talk about a documentary where a doctor demonstrates the liposuction procedure

Algorithm 1 The Incongruity Characterization Algorithm

Input: A piece of text T, threshold of LOF Θ

Output: Set of anomalies A, number of cluster k

Remove stop words and punctuations from T

S= Conducting stemming, remove the repeated word from T

Create the similarity matrix W where $W(i,j)=NGD(i,j)$

{Calculate LOF value for each point}

    While $i < length(T)$ **do**

    $L(i)=LOF(W,i)$

    End while{A is the set of anomalies}

While $L(i) \geq \Theta$ **do**

Append $S(i)$ to A

Remove row/col for i from W

End while

k=MonteCarloCrossValidation(W)

Output k, A

Figure 5.4: The Incongruity Characterization Algorithm

presence of overtly unreadable words in a piece of text could produce semantic anomalies which don't always correspond to any conceptual incongruity. This would clearly defeat our purpose as we intend to implicitly infer the presence of conceptual incongruity by detecting outliers in the semantic space. In order to ensure that the jokes in our dataset were easily comprehensible, we conducted the Flesch Kincaid readability test [57] on the raw text of the jokes. The average readability of the first two data-sets is shown below in Table 5.1[4] . The Flesch Kincaid readability score indicates the ease of reading a piece of text. A high score implies an easy text. For example, comic books and children's book score around 90 whereas legal documents and scientific journals (hard to read) usually score below 20. Table 5.1 shows that both the datasets score above 85. Hence, we can safely assume that the semantic anomalies would most likely not result due to the presence of overtly unreadable/specialized words. After this small sanity test, we executed our algorithm on three datasets which described briefly above in Section 1.2. We will now describe the results obtained on each of these datasets individually.

[4] The readability of the third data set was not calculated as it consisted of user-defined tags not raw unstructured text.

Table 5.1: Flesch Kincaid Readability Score Of The Datasets

Serial Number	Dataset Name	Score
1	One Liner Simple Jokes	86.8
2	Top Jokes From LaughLab	88.7

Table 5.2: Distribution Of Number Of Clusters In First Dataset

Number of Clusters	Number of Jokes	Percentage
1	148	0.9867
2	2	0.01

5.4.2 150 One-Liners: Simple Stimuli

As described earlier, this data set contains 150 one-liners, with one setup line and multiple punch-lines with just one of them being funny. We picked the human annotated funny line for each setup line which completed the joke, for each of the 150 one-liners. The number of jokes containing at least one anomaly was equal to 110 and the average number of anomalies per joke was equal to 1.26 (close to one anomaly per joke). After removing the anomalies the number of clusters in each joke was estimated and the average number of clusters per joke was equal to 1.02 (very close to 1). The frequency distribution of the number of clusters per joke is shown in Table 5.2. We then calculated the conditional probabilities mentioned in equations 5.4, 5.5. The probability of Incongruity Presence was equal to: $P(C = 1/A > 0)$=0.98 The probability of Incongruity Resolution was equal to: $P(C > 1/A > 0)$=0.02. Clearly, incongruity presence explains most of the One-Liner Jokes. The fact that this probability is resoundingly high supports the explanation that the comic effect in one-liners is usually produced by the sudden introduction of an incongruity like clever wordplay, non-contextual words etc.

estimation of techniques relies heavily upon the choice of random number generator

5.4.3 Jokes from Laughlab: Complex Stimuli

As described earlier our subset of this data set contains 50 jokes, consisting of the "top" joke of each country and a random sample from the rest 1001 jokes. We treat this set as a representative of "complex humor inducing stimuli". The number of jokes

Table 5.3: Distribution Of Number Of Clusters In LaughLab Jokes

Number of Clusters	Number of Jokes	Percentage
1	16	32
2	28	56
3	6	12

containing at least one anomaly was equal to 40 and average number of anomalies per joke was equal to 1.94 (almost equal to 2 anomalies per joke). After removing the anomalies the number of clusters in each joke was estimated and the average number of clusters per joke was equal to 1.8 (close to 2). The frequency distribution of the number of clusters per joke is shown in Table 5.3. We then calculated the conditional probabilities mentioned in equations 5.4, 5.5. The probability of Incongruity Resolution was equal to: $P(C > 1/A > 0)$=0.7. The probability of Incongruity Presence was equal to: $P(C = 1/A > 0)$=0.3. Clearly incongruity resolution explains a higher number of the laughlab jokes. The fact that this probability is not resoundingly in favor of one mechanism suggests the validity of both the mechanisms in complex narrative structures. Another interesting observation was that 15/18 of the top jokes contained at least one incongruity and the average number of clusters per joke was equal to 1.67, which clearly shows that most top jokes have at least one incongruity but their number of clusters (which is below the average) is quite variable.

5.4.4 YouTube Videos

As described earlier our subset of this dataset consists of the tag-sets of 50 most watched comedy videos crawled from YouTube. The number of tags per video ranges between $(0-85)$ and the median number of tags per video was equal to 9. The number of videos containing at least one anomaly was equal to 42. The average number of anomalies per joke was equal to 3.2. It is worth noting that this number is higher than both the previous datasets but it matches our expectations as this a noisy online dataset. After removing the anomalies the number of clusters in each joke was estimated and the average number of clusters per joke was equal to 1.3. The frequency distribution of the number of clusters per joke is shown in Table 5.4. We then calculated the conditional probabilities mentioned in equations 5.4, 5.5. The probability of Incongruity Resolution

Table 5.4: Distribution Of Num Of Clusters in YouTube Videos

Number of Clusters	Number of Jokes	Percentage
1	35	0.7
2	15	0.3

was equal to: $P(C > 1/A > 0)$=0.35. The probability of Incongruity Presence was equal to: $P(C = 1/A > 0)$=0.65. We discuss the results in greater detail in the next section

5.5 Discussion

Before we begin, we would like to reiterate, that the objective of this investigation was not to verify the correlation of incongruity with humor inducing stimuli (verified computationally by [47] [29]), nor was it to classify humorous stimuli from non-humorous ones in any way (initial attempt made by [47]). Instead we begin with the assumption that incongruity does account for humor generation and take up the debate further and verify if it's the mere presence of incongruity or its eventual resolution which leads to humor. Hence, the central focus of our work was not to come up with the best possible incongruity detection algorithm but rather to come up with an algorithm which can quantitatively verify the presence of patterns corresponding to incongruity presence and incongruity resolution in various kinds of humor inducing stimuli.

For reasons and references discussed earlier, we consider our first dataset as a representative of "'simple stimuli" and the subset of jokes from laughlab as representative of "complex stimuli". Simple jokes were better explained by the mere presence of incongruity. The average number of clusters per joke (in our simple dataset) was equal to 1.08 and the average number of anomalies per joke was equal to 1.26, which comes close to the pattern shown in Figure 2. The conditional probability (0.98) was resoundingly in favor of one mechanism over the other. Hence, we conclude that simpler jokes are better explained by the mere presence of the anomaly than its eventual resolution. This makes perfect intuitive sense because the comic effects in a one liner are usually produced by the subtle introduction of a simple incongruity in the form of effects like clever wordplay, misspelled words, puns, etc.

The average number of clusters per joke in our complex dataset was equal to 1.8

and the average number of anomalies per joke was equal to 1.94, which comes close to the pattern expected by us. (i.e. presence of at least one statistical anomaly and the presence of at least more than one cluster). The conditional probability (0.7) was in favor of the mechanism of incongruity resolution, though it should be duly noted that this probability is not resoundingly in favor of one mechanism. This could imply two possible things. First, a joke might not derive its narrative complexity merely from the presence of disparate cluster of words, it could also happen that many terms which generate a competitive explanation are not physically present in the joke but are implicitly inferred by the reader. Second, it is also likely that one single explanation contains multiple contextual or social explanations, whose corresponding terms aren't present in a joke but are implicitly understood. Hence, based on these results, we conclude that jokes containing a complex narrative structure are statistically more likely to be explained by the eventual resolution of an anomaly.

We feel, however, that there further deeper constraint on the types of humorous episodes available on YouTube, introduced by the richness of the medium of expression. Videos are a much richer medium of communication as compared to a piece of writing. The tag-set of a video merely represents a noisy semantic compression/description of the content of the video. It is more likely to simply contain some of the incongruous words rather than the all the words constituting all the competitive narratives which lead to the eventual resolution of the apparent incongruity. For example, the set of tags "baby, panda, cute, angry, keyboard", is the tag-set of a funny video where a cute baby panda unable to send an email correctly, displays its anger towards its keyboard. In this case, the unusual behavior of a panda creates an incongruity which is resolved by its apparent cuteness, but our algorithm would clearly flag "keyboard" as an anomaly and would support incongruity presence as explanation of humor. Due to the absence of enough number of words describing the entire scenario, our technique will be unable to find disparate clusters. While admittedly entering the realm of speculation on this front, we therefore propose a testable hypothesis that, comparing equally sized compendia of short videos and written anecdotes, the ratio of number of data points admitting incongruity resolution-based explanations to the number of total data points will be higher in the written set of anecdotes. The numbers we report in above section (Section YouTube Videos) are in line with our prediction but we hasten to add that large scale

experiments are required to test this hypothesis further.

Summarily in this chapter, we make the following contributions:

1. We frame the problem of testing two valid mechanisms of humor generated by the presence of incongruity as a quantitative equivalent of measuring the simultaneous presence of a statistical anomaly and diverse concept clusters in a semantic space.

2. We demonstrate a methodology of measuring two different kinds of incongruity, leveraging the semantic connectivity of the entire Web visible to Google's search crawlers.

3. We empirically show that a combination of semantic anomalousness and semantic diversity (number of disparate clusters), helps us verify two different mechanisms of humor on different datasets.

4. We report strong conclusive evidence, that simple humor inducing stimuli correlate strongly with the mere presence of incongruity and complex ones correlate significantly with the resolution of incongruity.

This work also makes a minor contribution in the larger context of social computing research. The use of Google Distance allows us to extract, in a manner of speaking, the sense the Internet hive mind has of various contexts. The use of this measure of semantic distance allows us to quantify the sense of ideas like semantic diversity and anomalousness which have caused severe rhetorical disagreements (see e.g., [44] for a representative example) amongst previous humor researchers. The ability to compile statistics on these quantities obviates, to a certain degree, the need to depend on any one frame of interpreting the linguistic terms under consideration. As such, we feel that this research further buttresses the value of computational investigations in resolving *language games* that hinder the reconciliation of different theories in the social sciences, and hence is a useful example of the value of computational social science research.

Chapter 6

Conclusion and Discussion

6.1 Summary Of Contributions

The principal contribution of this work is the quantification of the qualitative idea of "interestingness" derived from the incongruity theory, in the domain of text data. First, we propose an information theoretic definition of incongruity in text. Then, we provide the necessary mathematical proofs and arguments supporting our proposal. Second, based on this we provided an anomaly detection algorithm, incorporating semantic context and hence bettering the state-of-the-art of text anomaly detection algorithms. Third, we provided a computational model of humor which replicated qualitative expectations set by previous researchers. Incongruity showed strong correlation with humor but it wasn't highly predictive of humor, as surmised by many researchers previously. Fourth, we provided a general algorithmic framework to detect incongruity in various other kinds of data displaying "interestingness", like creative words, popular media objects etc. We found that incongruity does display a monotonic U-shaped behavior as surmised earlier. Finally, as this entire work is based on the idea of characterizing incongruity in the semantic space; we conducted a data-driven investigation verifying the presence of different kinds of incongruity in different kinds of humorous stimuli. We presented an algorithm for doing so and found that simpler stimuli like slapstick jokes, puns, one-liners etc. were better explained by the mere presence of incongruity, whereas more complex stimuli; like high-quality jokes, jokes with hidden meanings etc. contained a greater amount of conceptual dissonance and hence, were better characterized using the

idea of resolution of incongruity. A quick snap-shot of the summary of contributions made in each chapter is given below.

- Chapter 2 shows that the use of external semantic information can reduce the false positive rate in existing systems by explaining away spurious statistical deviations. Introducing contextual information from a much larger corpus helps us find semantic explanations for anomalies and in filtering out false positives. We can also flag previously suspect data points as anomalies with greater confidence. Our approach also allows for threshold manipulation based on operator input and can lead to better anomaly detection system design. (For details, please check Chapter 2)

- Chapter 3 provided information theoretic motivations behind incongruity eventuating to interestingness in the semantic space and also showed how the two measures designed by us for doing so were the equivalent of computing entropy and conditional entropy respectively in any given semantic space. We described methods to compute $Diversity, Anomalousness$ in a semantic space under varying degrees of noise. We found that incongruity correlates strongly with subjective human judgment of interestingness and that that high and low incongruity classes correlate strongly with high and low interestingness. We also provided empirical validation (in semantic space) of the fact, that incongruity is a necessary but not sufficient condition for interestingness.

- Chapter 4 illustrates a reliable method of extracting the semantic sense of a multimedia object from user-defined tags and empirically show that a combination of semantic diversity and anomalousness, as measured in the space of related tags, strongly correlates with the categorization of videos as humorous on YouTube. We find, however, that in aggregate, this combination of semantic diversity and incongruity is a relatively weak predictor of humor in videos. We interpret our findings as the first large-scale empirical evidence of the fact that incongruity is a necessary, but not sufficient, condition for media objects to appear humorous.

- Chapter 5 presents an algorithm of measuring two different kinds of incongruity in text data. We empirically show that it helps us verify two different mechanisms

of humor on different datasets. We report strong evidence, that simple humor inducing stimuli correlate strongly with the mere presence of incongruity and complex ones correlate significantly with the resolution of incongruity.

Finally, we hope this work triggers a greater interest in incorporating intrinsic motivations in the act of knowledge discovery.

6.2 Future Directions

There could be many possible applications of the contextual anomaly detection algorithm described in Chapter 2. We showed that this idea could be implemented at various levels of abstraction. We operationalised it at word level and topic level. A possible future work in this direction would be to be able to do this exercise at document level as well. This could be a useful system design exercise. Extension of our approach to online settings could significantly improve existing techniques of sentiment extraction being researched using social network feeds [16]. Also, since anomaly detection here occurs at a topic level, it is possible to implement privacy-preserving tracking of intra-organizational communication using systems built around our basic concept.

The investigation done in Chapter 3 (The algorithmic framework for characterizing interestingness) points towards developing a new interdisciplinary class of knowledge discovery algorithms and could result in many useful applications. As demonstrated earlier, the augmentation of our feature space with the classical bag of words feature space could better the classification accuracy of popularity ratings. Similarly, there could be many other such potential applications where the augmentation for our theory driven feature space could help knowledge discovery algorithms. For example, our technique could be used for finding influential nodes in a network (like twitter) [49] by identifying nodes with high connectivity and outreach who also generate "interesting" content. A second possible application of our technique could be in the field of recommendations. Most recommendation systems recommend future items based on either the agent's past history or based the preferences shown by agents similar to the the given agent (collaborative filtering). Let's say, based on the a user's past preferences we can figure out how incongruous a given observation is likely to be and then augment this incongruity as a part of the existing recommendation feature space, then we

could possibly predict items which the user or his collaborative locality has not been exposed to in the past, but is very likely to be interested in, in the future. For example, person who mostly watches comedy movies, might find a hardcore action movie absolutely uninteresting but might find a "parody action" movie highly satisfying. A third possible application could be in the area of advertising. Two possible directions could be pursued as follows. First, mass media advertisers try to write a script which is both relevant to the campaign and yet interesting to most people. Our method for detecting an interesting piece of text could be used as an initial filter to pick out the the most likely interesting advertisements from a large initial pool of advertisements given a training set of most representative previous advertisements in a given domain. Secondly, most advertisement placement software mine a given user's personal history and show the most relevant ads. Users are more likely to be interested in clicking on an ad if it uses a creative choice of words and hence looks interesting to them based on their personal threshold of incongruity. As described earlier, we could detect sets of words considered creative by human subjects. A similar implementation could serve as a good personal filter for the advertisements. Though our work is purely data-driven in nature, it raises a couple of important questions which could lead to theoretical advancements in social and behavioral sciences. It clearly suggests that psychological and physiological responses can be predicted based on information seeking patterns. This brings up the natural next question; are all information seeking behaviors similar in nature, if no what are the major sub-categories of such behavior? Can different kinds of humor and creativity be characterized by using the idea of incongruity alone? Do implicit environmental factors play an important role in determining levels of interest shown by an agent? The good news is that in the era of web 2.0, the deluge of human interaction data available across different content generation platforms presents a unique opportunity to convert all the above and many other such questions into testable hypotheses.

The model proposed in Chapter 4 (The Computational Model Of Humor) generates a plausible account of the correlation between incongruity and humor. But our research still leaves open the question of causality. In particular, our present analysis (in Chapter 5) has only made an initial attempt in differentiating between two possible mechanisms of humor. (Incongruity versus Incongruity Resolution). The much deeper question on

whether incongruity is the root cause or is it accompanied with something deeper is still open. True cause for humor inducement still remains undiscovered, since incongruity cannot be causally implicated in a simple mechanistic way. Resolving this mystery would greatly deepen our understanding of the link between information foraging and humor inducement, and will be a primary focus for our future investigations in this area.

The algorithm illustrated in Chapter 5 (Comparison between Incongruity and Incongruity Resolution) has two primary limitation. Firstly, we completely rely on the words present in the joke to construct our semantic space. In any standard conversation and hence in a joke, the implicit presence or absence of some words adds meaning to a joke. Ideally, we require all those words to construct the complete semantic space, and as mentioned earlier the likelihood of presence of implicit terms is much higher in a complex narrative structure. Secondly, a limitation which is fundamental to all computational models of humor is that it tends to ignore the social content. Jokes about a dictatorial boss, an inept colleague, a hard to use device may not exactly be identified as incongruities semantically, but they are incongruous if we also incorporate the social context in our explanation. We also hasten to add that the generalizability of our results and their interpretations towards forming a comprehensive account of humor is necessarily constrained by the ecological specificity of our dataset. In spite of the above mentioned limitations, the approach illustrated in Chapter 5 can result in immediate practical applications in domains like marketing and advertising. We list a couple of likely future applications below.

1. Advertisement campaigns regularly look for funny catch lines and make use of different kinds of jokes in different places. For example, billboards normally contain a funny one-liner but the narrative on a pamphlet tends to be slightly more complex in order to leave a longer impression on the reader. Our technique could easily be used as a filter to categorize different kinds of humor inducing stimuli and hence lead to better manual selection.

2. On-line advertisements are usually tailored based on the browsing preferences of users in order to increase the click through rate. Stimulus complexity could serve as an important signal in tailoring these advertisements. For example, a person who shows heavy interest in browsing through a certain class of stimuli could be

more interested in clicking on funny textual advertisement belonging to that class of stimuli.

6.3 Epilogue

"If you are out to describe the truth, leave elegance to the tailor."- Albert Einstein. This quote from Einstein summaries the true feelings of the author at the end of this investigation. Most of our findings, may or may not be surprising enough, but we hope they are useful enough to aid future data-driven research in this area. Finally, we feel, we have only scratched the surface of some of the biggest mysteries of human condition. And to the path that lies ahead, the author would like to quote some lines, which have always inspired him.

Though much is taken, much abides; and though;
We are not now that strength which in old days
Moved earth and heaven, that which we are, we are;
One equal temper of heroic hearts,
Made weak by time and fate, but strong in will
To strive, to seek, to find, and not to yield.
–Lord Tennyson

References

[1] Agovic A., Shan H., Banerjee A. Analyzing aviation safety reports: From Topic Modeling to Scalable multi-label Classification, In Proceedings of the Conference on Intelligent Data Understanding, 2010

[2] Blei D., Ng A., Jordan M. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003 3, 993-1022.

[3] Bollegala D., Matsuo Y., Ishizuka M. Measuring the similarity between implicit semantic relations from the web. In Proceedings of the 18th International Conference on World Wide Web, ACM, 2009, pp. 651- 660

[4] Chandola V., Banerjee A., Kumar V. Anomaly detection: a survey . ACM Computing Surveys 2009, 41 (3), pp. 1-58.

[5] Cilibrasi R., Vitanyi P. The Google Similarity Distance, In IEEE Transactions on Knowledge and Data Engineering, 2007, 19 (3), pp. 370-383.

[6] Guthrie D., Guthrie L., Allison B., Wilks Y. Unsupervised anomaly detection. In Proceedings of the twentieth international joint conference on artificial intelligence, 2007 pp. 1626-1628.

[7] Jiang J.J., Conrath D.W. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics, 1997

[8] Lin D. (1998). An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, 1998, pp. 296-304.

[9] Manevitz L., Yousef M., Document classification on neural networks using only positive examples, Proc. 23rd Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval, vol. 34, pp. 304 - 306, 2000.

[10] Manevitz L., Yousef M., One-class SVMs for document classification, Journal of Machine Learning Research, 2002, 2.

[11] Mangalath P., Quesada J., Kintsch, W. Analogy-making as predication using relational information and LSA vectors. In K.D. Forbus, D. Gentner & T. Regier (Eds.), Proceedings of the 26th Annual Meeting of the Cognitive Science Society, Chicago: Lawrence Erlbaum Associates, 2004.

[12] Medelyan O., Milne D., Legg C., Witten I.H. Mining Meaning from Wikipedia. International Journal of Human-Computer Studies, 2009, 67(9), pp. 716-754.

[13] Liu D., Hua X., Yang L.,Wang L., Zhang H. Tag ranking. In Proceedings of the 18th international conference on the World Wide Web, 2009

[14] Newman D., Asuncion A., Smyth P., Welling M. Distributed inference for latent Dirichlet allocation.In Proceedings of NIPS 20. MIT Press, Cambridge, MA, 2008.

[15] Pedersen T., Patwardhan S., Michelizzi J. WordNet: Similarity - measuring the relatedness of concepts. In Proceedings of the 19th National Conference on Artificial Intelligence, 1997; pp.144–152.

[16] Petrovi S., Osborne M., Lavrenko V. Streaming first story detection with application to Twitter, In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010; pp.181-189.

[17] Sontag D., Roy D. Complexity of inference in Latent Dirichlet Allocation. NIPS, pp. 1008-1016, 2011.

[18] Srivastava A., Zane-Ulman B. Discovering recurring anomalies in text reports regarding complex space systems. In Proceedings of IEEE Aerospace Conference, IEEE Computer Society Press, Los Alamitos, 2005.

[19] Srivastava N, Srivastava J. A hybrid-logic approach towards fault detection in complex cyber-physical systems. In Proceedings of the Annual Conference of the Prognostics and Health Management Society, 2010.

[20] Resnik P. Using information content to evaluate semantic similarity. In Proceedings of the 14th International Joint Conference on Artificial intelligence, 1995; pp. 448-453.

[21] Gligorov R., Kate W., Aleksovski Z., Harmelen F. Using Google distance to weight approximate ontology matches, In Proceedings of the 16th international conference on the World Wide Web, 2007.

[22] Wagstaff K., Rogers S., Schroedl S. Constrained K-means clustering with background knowledge. In Proceedings of the International Conference on Machine Learning, 2001; pp. 577 - 584 .

[23] WordNet *http://wordnet.princeton.edu/*

[24] WordNet: Similarity *http://marimba.d.umn.edu/*

[25] Topic Modelling toolbox *http://psiexp.ss.uci.edu/research/programsdata*

[26] Frank A., Asuncion A. UCI Machine Learning Repository *http://archive.ics.uci.edu/ml*. University of California, Irvine, CA; 2010.

[27] Lin D. (1998). An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning (ICML-98), Madison, WI, pp. 296-304

[28] Mahapatra A.,Srivastava N.,Srivastava J.(2012). Contextual Anomaly Detection In Text Data.Text Mining Workshop, Proceedings of the Twelth SIAM International Conference on Data Mining, Anaheim,CA, April 26-April 28.

[29] Mihalcea R., Strapparava C. and Pulman S. (2010).Computational models for incongruity detection in humour. Lecture Notes in Computer Science, 6008/2010, 364374.

[30] Schalekamp F., Zuylen A. V. (2009) Rank aggregation: Together were strong. In Proc. of 11th ALENEX, pages 3851. SIAM.

[31] Weinberger K., Slaney K., and Zwol R.(2008) Resolving tag ambiguity. In Proc. ACM Conf. Multimedia, pages 111229.

[32] Wyer R. S., Collins J. E. A theory of humor elicitation. Psychological Review, Vol 99(4), Oct 1992, 663-688.

[33] Veale T. (2004). Incongruity in humor: Root cause or epiphenomenon? HUMOR:International Journal of Humor Research, 17 (4), pp. 419428.

[34] Raskin V.(1985). Semantic Mechanisms of Humor. Dordrecht and Boston & Lancaster:D. Reidel Publishing Company

[35] Morreall J. (ed.) 1987. The Philosophy of Laughter and Humor. New York: State University of New York Press.

[36] Brock A.(2004). Analyzing scripts in humorous communication. HUMOR:International Journal of Humor Research, 17 (4), pp. 353360

[37] Keith-Speigel P.(1972)Early conceptions of humor: Varieties and issues. In Goldstein and McGhee, pages 339.

[38] Krikmann Arvo (2007) Contemporary linguistic theories of humour. Folklore: Electronic Journal of Folklore 33, 2757

[39] Fun Fare: A Treasury of Reader's Digest Wit and Humor. Pleasantville, NY: The Reader's Digest Association, 1949 (p. 84).

[40] Smuts A. (2006) 'Humor', The Internet Encyclopedia of Philosophy, accessed 12 May, 2012.

[41] TED Talk:Kevin Allocca: Why Videos Go Viral

[42] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I..H (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[43] Ranjan A., Srinivasan N. (2010). Dissimilarity in creative categorization. Journal of Creative Behavior, 44, 71-83

[44] Ritchie G.,(1999) Developing the Incongruity-Resolution Theory. Pp. 78-85 in Proceedings of AISB Symposium on Creative Language: Stories and Humour, Edinburgh, April 1999

[45] Loewenstein G. (1994). The psychology of curiosity: A review and reinterpretation. Psychological Bulletin, 116, 7598.

[46] Mahapatra A., Srivastava N., Srivastava J. (2012). Contextual Anomaly Detection In Text Data. Algorithms. 2012; 5(4):469-489

[47] Mahapatra A., Srivastava N., Srivastava J. (2012). Characterizing The Internet's Sense Of Humor. International Confernece on Social Computing (SocialCom) (pp. 579-584). IEEE.

[48] Oudeyer P. -Y. and Kaplan F. (2006). What is intrinsic motivation? A typology of computational approaches. Frontiers Neurorobot., vol. 1.

[49] Weng J., Lim E., Jiang J., He Q. (2010). Twitterrank: Finding topicsensitive influential twitterers. ACM international conference on web search and data mining.

[50] Blahut R.E. (1991) Principles and Practice of Information Theory. Reading,MA: Addison-Wesley

[51] Breunig M. M., Kriegel H. -P., Ng R. T., Sander J. (2000) LOF: Identifying Density-based Local Outliers ACM SIGMOD Record 29: 93.

[52] Lazarevic A., Ozgur A., Ertoz L., Srivastava J., Kumar V. (2003). A comparative study of anomaly detection schemes in network intrusion detection. 3rd SIAM International Conference on Data Mining: 2536.

[53] Smyth (1996). Clustering using monte carlo crossvalidation. Knowledge Discovery and Data Mining, pages 126133, 1996.

[54] *http://www.richardwiseman.com/LaughLab/home.html* (2002)

[55] Buijzen M., Valkenburg P. M. (2004). Developing a typology of humor in audiovisual media. Media Psychology, 6(2), 147-167.

[56] Extended abstracts of the (3rd INTERNATIONAL) workshop on computational humor Amsterdam, June 8, 2012

[57] Kincaid J. P. (1975) Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Media Psychology, 6(2), 147-167.