# FEATURE SELECTION THROUGH VISUALISATION FOR THE CLASSIFICATION OF ONLINE REVIEWS

A Thesis

Submitted to the Faculty

of

Purdue University

by

Keerthika Koka

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

May 2017

Purdue University

Indianapolis, Indiana

# THE PURDUE UNIVERSITY GRADUATE SCHOOL
# STATEMENT OF THESIS APPROVAL

Dr. Shiaofen Fang, Chair

Department of Computer and Information Science

Dr. Yuni Xia

Department of Computer and Information Science

Dr. Arjan Durresi

Department of Computer and Information Science

**Approved by:**

Dr. Shiaofen Fang

Head of Departmental Graduate Program

To my dear ones

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Shiaofen Fang for his support, constant motivation and guidance during my study and research at the IUPUI. His trust and belief in my talent have led to the successful completion of this thesis. It was a great privilege for me to work with him. I would like to thank my thesis committee members: Dr. Yuni Xia and Dr. Arjan Durresi. I would like to thank my husband and family who have been very supportive. I would also like to thank my friends and members of the visualization laboratory who have made my research at the IUPUI an enjoyable and memorable one.

TABLE OF CONTENTS

Page

# LIST OF TABLES

# LIST OF FIGURES

# ABBREVIATIONS

D3 Data Driven Documents

# ABSTRACT

Koka, Keerthika M.S., Purdue University, May 2017. Feature Selection through Visualisation for the Classification of Online Reviews. Major Professor: Shiaofen Fang.

Opinion spamming is a reality, and it can have unpleasant consequences in the retail industry. As consumers and sellers take to on-line commerce increasingly, opinion spamming can pose some serious threats to the industry. While there are, several promising research works done on identifying the fake on-line reviews from genuine on-line reviews that typically involve and integrate approaches in the fields of data mining, machine learning, psychology, computational linguistics, natural language analysis; there have been surprisingly few in visual analytics. In this work, we study the possibility to analyze the reviews and classify them into fake and genuine reviews through visualization.

The purpose of this work is to prove that the visualization is at least as powerful as the best automatic feature selection algorithms. This is achieved by applying our visualization technique to the online review classification into fake and genuine reviews. The radial chart is one of the traditional representations of high dimensional data. In this work, radial chart and the color overlaps are used to explore the best feature selection through visualization for classification. Every review is treated as a radial translucent red or blue membrane with its dimensions determining the shape of the membrane. This work also shows how the dimension ordering and combination is relevant in the feature selection process. In brief, the whole idea is about giving a structure to each text review based on certain attributes, comparing how different or how similar the structure of the different or same categories are and highlighting the key features that contribute to the classification the most. Colors and saturations

aid in the feature selection process. Our visualization technique helps the user get insights into the high dimensional data by providing means to eliminate the worst features right away, pick some best features without statistical aids, understand the behavior of the dimensions in different combinations.

# 1. INTRODUCTION

## 1.1 Overview of fake and genuine on-line reviews

On-line reviews are a reality and an effective means for consumers to share their opinions or reviews about a product, brand, hotel, etc. With the advancement of the Internet, social networks, on-line retailers the users can openly share their views about various aspects with ease. On-line reviews have a significant impact in influencing the customers from buying or investing in a product. Sharing of opinions are helpful to both users and vendors. Before investing in a product, a customer would want to know the feedback from other users, and by giving a positive feedback or improvement advise the vendors can reach out to their customers for a better customer satisfaction. However, today the on-line reviews are manipulated for either promoting or demoting a product or a company. It turns out that deception in the on-line reviews has become a part of the on-line retail industry [1].

According to a research organization *mintel.com*, the 57 percent of the surveyed consumers are suspicious of products or brands having positive reviews, and 49 percent believe companies probably pay for positive reviews. The fact that the articles on spotting fake on-line reviews are frequently found on notable news sites, such as *consumerist.com*, *nbcnews.com*, *wikihow.com*, *thenextweb.com*, *cbcnews.com* to mention a few, makes it more essential to delve deep into this topic.

To gauge the severity and reality of this issue [2] market place conducted research. As part of which, they created a fake mobile fast food truck and created an account in the Yelp. They could successfully get some positive reviews posted on the Yelp site by some best-paid reviewers. They could also get their company ratings go up. They were successful in showing that the Yelp review filters could only filter the reviews based on the user accounts and IP address features. Implying that it is easy for the

companies to get their fake reviews posted on the sights passing through the filters by managing different accounts and maintaining IP addresses.

A one-star increase in the Yelp ratings leads to a 5-9 percent increase in the revenue of the respective target according to [3]. Their study also highlights several facts such as the on-line reviews are found to be prevalent in independent restaurants than the ones with chain affiliations. They show that the on-line consumer reviews substitute for more traditional forms of reputation. Customers depend upon the on-line reviews in their decision but are found to be responsive to the quality changes that are more visible than using all the information. Their study also reveals that consumers' response to the ratings are stronger when the reviews are informative. This study also highlights that the average rating of a restaurant is affected by the number of reviews but not the number of reviewers.

According to [4] the impact of on-line reviews is beyond local community unlike traditional word of mouth literature limited to social network, since the scope of Internet is vast. [5] examined the effect of consumer reviews on relative sales of books on Amazon and Barnes & Noble websites, who found that an improvement in a book's average review score could lead to an increase in comparable sales, and the 1-star reviews impacted greater than that of the 5-star reviews. Godes and Mayzlin (2003) examined that the dispersion, i.e., the extent to which conversations happened across a range of communities had the impact on the dynamic model of sales.

[6] built tool to calibrate movie revenue to forecast model based on a variation of Bass model using word-of-mouth and the traditional information sources such as *Infomediary*, offline word-of-mouth, and advertising. However, the results are mixed. Some research supports the view that on-line user reviews have an impact on sales while other research challenges such a view.

## 1.2 Text feature selection through visualization

Feature selection is one of the important steps in the data analysis. Feature selection also called as variable selection or attribute selection is the process of selecting the attributes from the data that are most relevant to the predictive modeling. From the past research work also presented in the previous research section, it is evident that visualization can be beneficial in the feature selection process to get better insights into the data. This research is an attempt to explore the feature selection process through data or text visualization. This work presents different experiments conducted and various observations. Overall there are two approaches conducted, and from the observations, it is evident that feature selection process through visualization can be at least as powerful as the best automatic feature selection algorithms.

## 1.3 Overview of visualization technique

The basic concept is to have a visual structure for each text review, based on each reviews dimensions. For this D3.js radial chart is used to represent each review as a data point with the LIWC calculated dimensions being the structure deciding points in the radial chart. The fake reviews and the genuine reviews are differentiated by the red and blue colored translucent membranes. Visualization presented in this work uses color blending and color saturations to help in presenting the dimension variations between the two categories. The more mixed the colors for a dimension appears, the less it contributes to the classification process theoretically. Our experiments also signify the importance of dimension order and combination.

## 1.4 Thesis organization

Chapter 1 is an introduction to the relevance of fake on-line reviews, text feature selection, and our visualization technique. Chapter 2 is about the previous works in the field of visual feature selection, text visual analytics, multi-dimensional data

visualization and on-line fake and genuine reviews classification. Chapter 3 is the theoretical explanation of the experiments, techniques, and approaches. Chapter 4 is an explanation of the approaches. Chapter 5 briefs about the experiments, results, and analysis. Chapter 6 is the conclusion.

# 2. PREVIOUS RESEARCH

This work is a collaboration of years of research in various fields like text analysis, classification through mining algorithms, visual analytics, opinion spam. Hence this section is classified into visual feature selection, text visual analytics, multi-dimensional data visualization, on-line fake and genuine review classification.

## 2.1 Visual feature selection

[7] uses visual feature analysis for criminal detection in digital forensics. They show that the use of visual analytics in the process of forensic investigations not only reduce the analysis time but also help pick on the slight changes of the features and their relations, thereby knowing the active features. They use a graph-based feature analysis model (GFAM) that uses self-organizing map. The visualization technique uses items on the x-axis and features on the Y axis.

In [8] an interactive visual system for subspace based analysis for HD data as informative structures in the data can be found and compared in different subspaces of a larger HD input space. This approach could effectively identify various impressive views and find the similarities of groups in data. They make use of scatter plot and parallel coordinates in their interactive visual system.

[9] proposed a visual analytics approach to analyzing related topics in a different textual corpus, using the graphs, the radial chart as fundamental visualization building unit. They manage to develop a level-of-detail visualization that balances both readability and stability.

How visualization in combination within the automated analysis by cohort construction, feature selection, model construction, result evaluation and tuning can help

in the overall analysis is shown in [10] and also dealing with spatiotemporal, multivariate visualization, pixel-based techniques, icon-based geometric projections.

## 2.2 Text visual analytics

[11] presented a visual analytics system Topic Stream for exploration and analysis of topic hierarchy in the input text streams. Dynamic Bayesian network (DBN) model to incrementally extract a new tree cut from the incoming topic tree. To present the hierarchical clusters and alignments visually over time a time-based visualization is developed. Main contributions are a streaming cut algorithm, a sedimentation based metaphor, a visual analytics system to integrate evolutionary hierarchical clustering. Theme rivers and stacks are used as the basic visualization technique.

[12] presents a visual analysis of Twitter time-series using pixel based calendar visualization technique, pixel based geo map, tag cloud and radial visualizations for sentiment representation which are applied to movie tweets, and web survey data finding interesting patterns in the customer feedback.

[13] uses theme rivers or stacked graphs to show the evolution of streaming messages from a micro-blog and also Google Maps API to show the location of the messages for situation awareness and exploration tasks which assure to handle the multiple aspects of text streams. Historical theme river helps in the historical analysis to show the evolution of text streams, while the current theme river provides real-time situation awareness.

[14] presents a visual opinion analysis for opinions in threaded discussions through integrated and interactive visualizations, which assures the ability to analyze over time aggregated opinions visually. Radial visualizations are used for this purpose. The visual interface has three views, 2 are radial views and one being a bar chart view. The first radial view shows the topics; the second one presents the popular keywords on that topic, third is a time view to present the post temporal distribution of the focusing thread.

[15] visualizes the dialogue between human and robot. Comments are classified into categories based on Latent Dirichlet Allocation (LDA). Topics of comments are visualized, and the color is used to represent the emotion to the comments. Comments are presented as nodes and the relationship between the comments by edges. It is an interactive approach by allowing users to assign colors to the node according to the emotion.

### 2.3 Multi-dimensional data visualization

Multi-dimensional data visualization is the representation of data with a high number of dimensions, visually that not only help the user make sense of the data effectively in a shorter period, but also enable interactive analysis to gain insights or knowledge. Difficulty in comprehending more than three dimensions and complex computation has made the visualization of high-dimensional data difficult for researchers according to [16]. 2d and 3d projections of the high-dimensional data have been the common way of representing high dimensional data visually as our human minds are effective in understanding these. This paper presents a framework named rank-by-feature framework for feature detection. They use Histogram Ordering for 1D projections and scatterplot ordering for 2D projections. The concept is to enable the users to find interesting histograms, scatterplots to visualize separately, which are associated interactively with other visualization views such as dendrogram, color mosaic view, tabular view, lateral coordinated view, making it possible to comprehend data from different perspectives. This helps in addressing the problems like identifying extreme values of criteria such as correlation coefficients or uniformity measures.

According to [17] the scope of visual analytics is an intersection of fields like information analysis, geospatial analysis, scientific analysis, statistical analysis, knowledge discovery, data management knowledge representation, presentation, production and dissemination, cognitive, perceptual science and interaction. Human factors play an

important role in interaction, cognition, perception, collaboration, presentation, and dissemination. The visualization of raw data does not make sense; hence it is important to do some analysis of what needs to be presented visually. "It is a goal-oriented process to gain insight into heterogeneous, contradictory and incomplete data through the combination of automatic analysis methods with human background knowledge and intuition."

[18] discusses about the visualization of multidimensional marine space data (MSD). It is one of the complex data that is hard for visualization, its principle is to pre-process the data with the most suitable preprocessing method affected by the dimensions and data type, and all the data is visualized on same projection coding system and coordinate system. They use the 3d modeling and 3d visualizations for MSD.

[19] presents multidimensional visualization (MDV) technique which is an improved form of parallel coordinates. They strive for a technique that aims for "seeing your data in a single picture." This visualization interface WinViz aims for display of multi-dimensional data and visual formulation of an interactive query. The query capabilities are compared to that of SQL.

[20] presents a visualization tool named TimeSpan which helps in exploring and analysis of temporal, multi-dimensional and multi-typed data for stroke patients, to improve the quality of care. The combination of stacked bar graph and heterogeneous embedded data attributes is used in a unified view. From this work, it is understood that not to use complex structures for visualization unless demanded.

## 2.4 Online fake and genuine reviews classification

[21] developed and compared three approaches to detect deceptive opinion spam and ultimately develop a classifier that is nearly 90 percent accurate on the gold standard. The data we use in this research work is from this research work as a standard

gold data for fake and genuine reviews is produced. They make several observations revealing a relationship between deceptive opinions and imaginative writing.

[22] explore on the extent of to which the readability, genre and writing style could predict review authenticity by conducting linguistic analysis concluding that these factors are indeed very important predictors of review authenticity.

[23] focuses on the detection of deceptive opinion spam. N-gram techniques extended using feature selection and different representation of opinions, modeled as the classification problem and Naive Bayes (NB) classifier and Least Squares Support Vector Machine (LS-SVM) are used. Different representations such as Boolean, bag-of-words, term frequency-inverse document frequency of the opinions are shown. The experiments are carried widely on the standard gold data.

[24] also work experiments in the detection of fake reviews, through some selected features. Six-time sensitive features are proposed to highlight the fake reviews as early as possible. These works happen to have promising results in identifying the fake reviews with high precision and recall.

# 3. VISUAL REPRESENTATION AND FEATURE SELECTION FOR TEXT REVIEWS

## 3.1 Representation of text reviews as high dimensional data set

### 3.1.1 Source and form of the reviews

Though publicly available on-line reviews are ubiquitous, there is no way to understand what are genuine and fake reviews. The filtered-out reviews by Yelp, Google or Amazon, could have been an alternative. However, as discussed in the previous work section, it is evident that they are not reliable. Luckily [21], in their research work on opinion spam analysis contributed a gold standard dataset of fake and genuine reviews. They have created a 20 truthful and 20 deceptive opinions for each of 20 chosen hotels with a total of 800 opinions. 5-star truthful reviews are mined from the 20 most popular hotels on TripAdisor, following the work of Yoo and Gretzel(2009) for comparing the truthful and deceptive positive reviews. AMT is used to collect the deceptive opinions.

Data characteristics consist of 3130 non-5-star reviews, 41 non-English reviews, 75 reviews with fewer than 150 characters (according to the paper deceptive are at least 150 characters long) 1607 reviews from first-time authors balanced truthful and deceptive reviews by selecting 800 each. The number of truthful and deceptive reviews are balanced, which are of similar length using log-normal distribution as suggested in the work of Serrano et al. (2009). There is a total of 800 negative polarities and 800 positive polarity reviews, completing 1600 reviews.

### 3.1.2 Data preprocessing

Text reviews are converted to a data point with n dimensions. In this work, the number of dimensions ranges from 100 to 200. Since each review needs to be in the form of a data point with attributes, we made use of Linguistic Inquiry Word Count (LIWC) tool for sentiment analysis purpose. LIWC2015 is a gold standard in automated text analysis, that has been developed based on years of scientific research. LIWC reads in text and outputs the frequency percentage for linguistic categories predefined in its dictionary as well as word count. The data that is fed into the visualization consists of a record for every review with over 93 to 200 attributes each that we get as an output from the LIWC. Below are two sample reviews one Deceptive and other Truthful.

***Deceptive review*** *We stayed at the Chicago Hilton for four days and three nights for a conference. I have to say; normally I am very easy going about amenities, cleanliness, and the like...however our experience at the Hilton was so awful I am taking the time to write this review. Truly, DO NOT stay at this hotel. When we arrived in our room, it was clear that the carpet hadn't been vacuumed. I figured, "okay, it's just the carpet." Until I saw the bathroom! Although the bathroom had all the superficial indicators of housekeeping having recently cleaned (i.e., a paper band across the toilet, paper caps on the drinking glasses, etc., it was clear that no ACTUAL cleaning took place. There was a spot (probably urine!) on the toilet seat and, I kid you not, the remnants of a lip-smudge on the glass. I know people who have worked many years in the hotel industry, and they always warned that lazy housekeeping would make things "appear" clean, but in fact, they make no effort to keep things sanitary. Well, the Hilton was proof. I called downstairs and complained, and they sent up a chambermaid hours later. Frankly, I found the room disgusting. The hotel itself, outside the rooms, was cavernous and unwelcoming, with an awful echo in the lobby area that created a migraine-inducing din. Rarely have I been so eager to leave a place like this. When I got home, I washed all my clothes whether I had worn them or not;*

*such was the skeeziness of our accommodations. Please, do yourself a favor and stay at a CLEAN hotel.*

***Truthful Review*** *My $200 Gucci sunglasses were stolen out of my bag on the 16th. I filed a report with the hotel security and am anxious to hear back from them. This was such a disappointment, as we liked the hotel and were having a great time in Chicago. Our room was nice, with two bathrooms. We had two double beds and a comfortable hideaway bed. We had a great view of the lake and park. The hotel charged us $25 to check in early (10 am).*

The table 3.1 and 3.2 are a sample output from the LIWC for the above mentioned deceptive and truthful from section 3.1.2. This is output with 93 attributes, which are the keywords provided by the LIWC2015 dictionary and their relative frequencies in the entire text of each review.

The LIWC2015 program reads the text and outputs the percentage of words reflecting different emotions, thinking styles, social concerns, and even parts of speech. LIWC was developed by researchers from social, clinical, health and cognitive psychology. The categories aim to capture people's social and psychological states.

It has a text analysis module and dictionary. It is a Java based application. The dictionary is a predefined list of words categorized into psychologically relevant ways. The dictionary can also be user defined. Typically, the text processing module compares each word in the input text and compares against the inbuilt dictionary words or user-defined dictionary, mapping the words to relevant psychologically-relevant categories. Eventually, it calculates the percentage of total words that are mapped. The text analysis module identifies and categorizes words based on the dictionary.

In this work, LIWC2015 version is used which has a dictionary composed of 6,400 words and stems and selected emoticons. For every dictionary word, there is a corresponding category such as sadness, negative emotion, verb, past focus, etc. The words that go into the dictionary are a result of intense research work, determining the core of each word category. This involved the use of large data sets to test how

every word of the dictionary is related to others in the same category in a statistically valid way.

Table 3.1.
LIWC Output for the reviews

| # | Attribute | Deceptive | Truthful | # | Attribute | Deceptive | Truthful |
|---|---|---|---|---|---|---|---|
| 1 | WC | 273 | 85 | 47 | differ | 3.3 | 1.18 |
| 2 | Analytic | 64.27 | 87.04 | 48 | percept | 2.2 | 2.35 |
| 3 | Clout | 45.62 | 82.67 | 49 | see | 1.47 | 1.18 |
| 4 | Authentic | 68.01 | 12.57 | 50 | hear | 0.37 | 1.18 |
| 5 | Tone | 32.04 | 95.55 | 51 | feel | 0 | 0 |
| 6 | WPS | 16.06 | 12.14 | 52 | bio | 2.2 | 0 |
| 7 | Sixltr | 15.75 | 11.76 | 53 | body | 1.1 | 0 |
| 8 | Dic | 87.91 | 80 | 54 | health | 0.73 | 0 |
| 9 | function | 54.95 | 55.29 | 55 | sexual | 0 | 0 |
| 10 | pronoun | 14.29 | 11.76 | 56 | ingest | 0.37 | 0 |
| 11 | ppron | 9.16 | 10.59 | 57 | drives | 5.86 | 9.41 |
| 12 | i | 5.13 | 3.53 | 58 | affiliation | 1.83 | 5.88 |
| 13 | we | 1.83 | 5.88 | 59 | achieve | 1.47 | 0 |
| 14 | you | 0.73 | 0 | 60 | power | 0.73 | 0 |
| 15 | shehe | 0 | 0 | 61 | reward | 1.47 | 2.35 |
| 16 | they | 1.47 | 1.18 | 62 | risk | 0.37 | 1.18 |
| 17 | ipron | 5.13 | 1.18 | 63 | focuspast | 9.16 | 7.06 |
| 18 | article | 11.36 | 11.76 | 64 | focuspresent | 6.59 | 3.53 |
| 19 | prep | 10.26 | 14.12 | 65 | focusfuture | 0.73 | 0 |
| 20 | auxverb | 8.06 | 9.41 | 66 | relativ | 14.65 | 9.41 |
| 21 | adverb | 5.49 | 3.53 | 67 | motion | 2.56 | 0 |
| 22 | conj | 6.96 | 5.88 | 68 | space | 8.79 | 7.06 |
| 23 | negate | 2.2 | 0 | 69 | time | 4.4 | 3.53 |
| 24 | verb | 17.58 | 11.76 | 70 | work | 2.93 | 1.18 |

Table 3.2.
LIWC Output for the reviews

| # | **Attribute** | **Deceptive** | **Truthful** | # | **Attribute** | **Deceptive** | **Truthful** |
|---|---|---|---|---|---|---|---|
| 25 | adj | 4.4 | 8.24 | 71 | leisure | 2.93 | 4.71 |
| 26 | compare | 1.47 | 1.18 | 72 | home | 3.3 | 4.71 |
| 27 | interrog | 1.47 | 0 | 73 | money | 0 | 0 |
| 28 | number | 0.73 | 5.88 | 74 | relig | 0 | 0 |
| 29 | quant | 1.1 | 1.18 | 75 | death | 0 | 0 |
| 30 | affect | 5.49 | 9.41 | 76 | informal | 0.73 | 0 |
| 31 | posemo | 2.93 | 7.06 | 77 | swear | 0 | 0 |
| 32 | negemo | 2.56 | 2.35 | 78 | netspeak | 0 | 0 |
| 33 | anx | 0 | 1.18 | 79 | assent | 0.37 | 0 |
| 34 | anger | 0 | 0 | 80 | nonflu | 0.37 | 0 |
| 35 | sad | 0 | 1.18 | 81 | filler | 0 | 0 |
| 36 | social | 6.96 | 8.24 | 82 | AllPunc | 20.51 | 15.29 |
| 37 | family | 0 | 0 | 83 | Period | 7.69 | 8.24 |
| 38 | friend | 0 | 0 | 84 | Comma | 8.06 | 2.35 |
| 39 | female | 0 | 0 | 85 | Colon | 0 | 0 |
| 40 | male | 0 | 0 | 86 | SemiC | 0 | 0 |
| 41 | cogproc | 9.89 | 1.18 | 87 | QMark | 0 | 0 |
| 42 | insight | 1.1 | 0 | 88 | Exclam | 0.73 | 0 |
| 43 | cause | 1.47 | 0 | 89 | Dash | 0.73 | 0 |
| 44 | discrep | 0.37 | 0 | 90 | Quote | 1.47 | 0 |
| 45 | tentat | 1.1 | 0 | 91 | Apostro | 0.73 | 0 |
| 46 | certain | 3.3 | 0 | 92 | Parenth | 1.1 | 2.35 |

The tables above list the categorical dimensions of the default LIWC. LIWC in its default dictionary has around 93 categories where each category has several English language words. The words in the dictionary can be overlapped in distinct categories. For example, the word *I* can fall under personal as well as I category. All the variables except WC (Word Count) and WPS (Words per sentence) indicate the percentage of total words. For instance, the sample insights gained from the above table are that how deceptive review has significantly higher use of informal words while it is almost 0 in the truthful. Or the percentage usage of the comma in the deceptive review is fairly higher than that used in the truthful, etc.

***Note**: A text of 10,000 words is claimed to yield far more reliable results than the one with 100.*

The text reviews are hence transformed into data set of 1600 rows and 93 columns, where each category of LIWC output is considered a dimension.

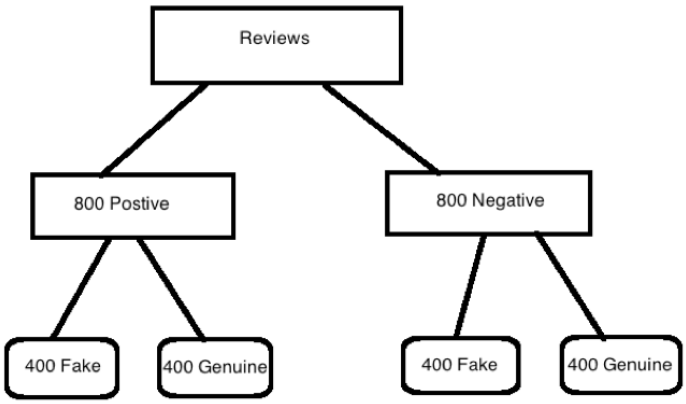



Fig. 3.1. Total reviews classification

All the attribute values are normalized to range between 0 to 1 about the columns max and min value (value - min/max - min ). All the values are scaled equally for plotting on radial chart axis.

### 3.1.3 Why use LIWC

LIWC is a transitional text analysis, shifting from traditional language analysis program to a new era of language analysis, which could be able to analyze more complex language structures. The first version of it is developed as part of exploratory language study Francis, 1993; Pennebaker,1993 as part of an exploratory study of language. The updated version with the updated dictionary and more modern design [25]. The most recent versions LIWC2007 and LIWC2015 have significantly altered both the dictionary and design. LIWC works by analyzing the frequencies of the select categorical words, while there are others which work on n-grams, groups of two or more words together. Language style information is one of the best-proven ways to understand a person's state of mind just as in LIWC. LIWC efficiently summarizes the dimensions that reflect emotional state, social relationships, thinking styles and individual differences as quoted in [26] and hence rightly captures the function and emotion words. From the fact that it is cited and used by 2379 research papers in their respective studies and research experiments, and from above-mentioned functionality, LIWC is chosen for the text analysis and conversion of text reviews to n-dimensional data.

## 3.2 Radial chart visualization

### 3.2.1 What is radial chart and significance in this work

Radial visualization is a practice of displaying data in a circular or elliptical pattern according to [27]. Radial visualizations are first coined in the paper Hoffman et al. in the 1990s.Though, there are many techniques evolved today the underlying concepts are firmly rooted in the statistical graphics literature of 19th century. Some common radial visualization techniques are the pie chart, star plots, socio-grams, polar-plot, and ring based.

In this work, we use a simple radial graph to plot each of 1600 reviews and their attributes. Radial Chart also knew as web chart, spider chart, star-plot, cobweb chart, irregular polygon or polar chart is a graphical method of displaying multivariate data in a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point, as defined. It is plotted on a polar coordinate system, rather than on a Cartesian one.

Radar charts can represent multivariate data without obscuring the details of the variables such as highest, lowest scoring variables, similar ranging variables, and outliers among each variable. Since our data is also multi-variable and is to be analyzed efficiently, radar chart has been on top of the list for the visualization technique.

Each variable is provided its axis and scaled and distanced equally. The plotting for the values are made on each respective axis with the help of the grid lines that connect the axes, as shown in figure 3.2, and by connecting all the points on the chart, we get a polygon. Though radar charts are inherent with some drawbacks, they are still chosen as the basic visualization technique in this work due to the enhancements proposed in this work.

### 3.2.2 Radial visualization of our data with D3.js

The idea is to plot 1600 polygons for each of the review data records on to the D3.js based radial chart. The drawbacks of radar chart are: having too many polygons makes it difficult to make sense of the polygons, and too many variables make it difficult to perceive the analysis. Since there is a limit to the perception of a human eye, after a fair number of trials it is decided to present at most 30 attributes of each record at a time on the radial chart screen. Still, these many polygons seem not to make any sense. Hence, our proposed approach makes use of 2 categorical colors and color saturation to overcome this drawback.

The radial chart takes the dimensions of each data point in our case each review and its dimensions and plots the normalized values on the polar coordinate plane,

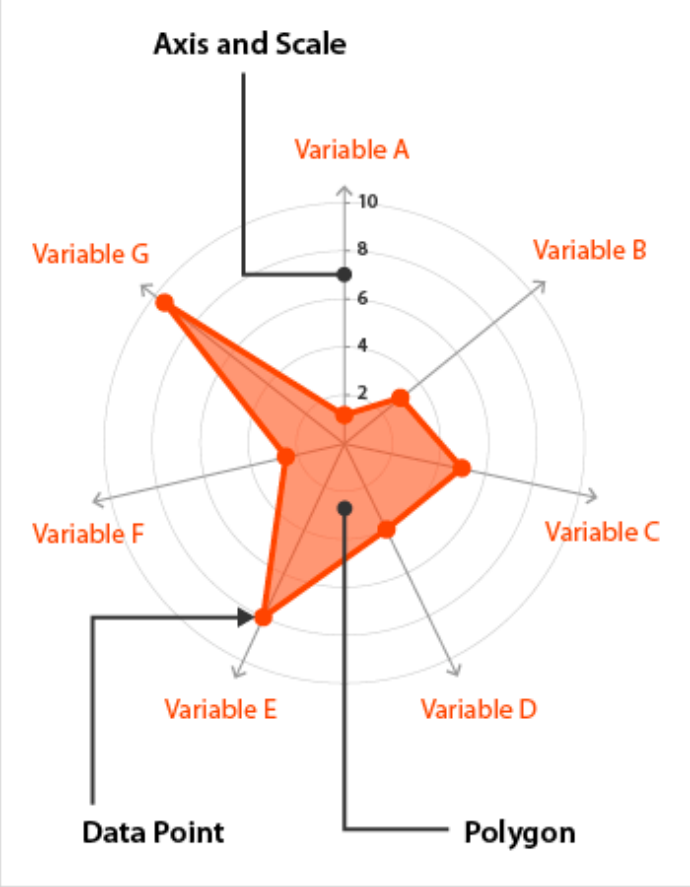
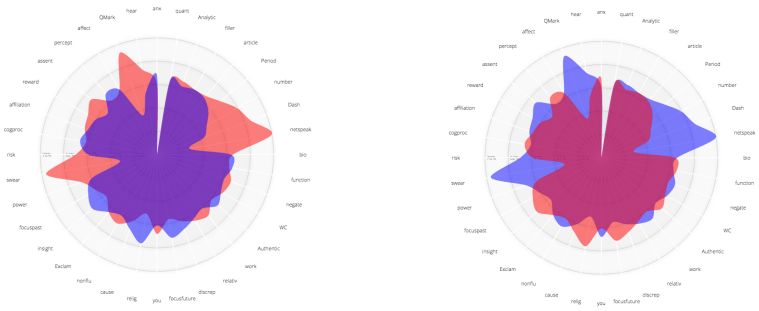



Fig. 3.2. Radar chart anatomy *Credits datavizcatalogue.com*

thereby giving a structure to each review. Every fake review is assigned a red color while every genuine review a blue color or vice versa. Hence for every review, a radial plot with the 30 attributes spreads itself like a red or blue translucent membrane.

Fig 3.3 is a sample figure to explain the radial membranes colored membranes representing the fake and genuine reviews plotted. The figures are to show the differences of red over the blue membrane and blue over a red membrane. The overlapped area represents the impure colors and hence not an interesting region to concentrate on. However, the pure regions like *netspeak*, *Qmark*, swear are the dimensions that show a large pure colored region which might prove to be interesting features.



(a) Red for Genuine (b) Blue for Genuine

Fig. 3.3. Radial blue red membranes representing fake and genuine reviews

### 3.3 Color Overlap for purity or impurity of the dimensions

D3.js natural color overlap behavior (Saturation)

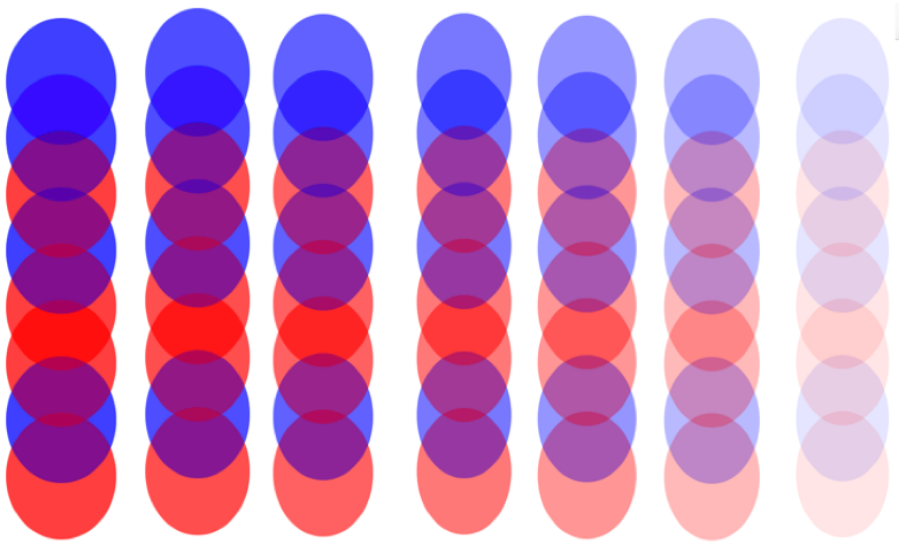


Fig. 3.4. Color blending palette

In figure 3.4 from top to bottom, the colored circles are blue, blue, red, blue, red, red, blue, red respectively with each column varying in opacity ranging between 0 and 1. These overlapping circles are to demonstrate how the colors change naturally in D3.js for different saturation values.

Reason to choose Blue and Red as our two colors are completely empirical. After several rounds of trials for various color combinations in D3.js, blue and red color combination seemed to differentiate the categories better. This area can sure be explored further, as the probability of success is bound to improve with the correct color combinations.

As shown in figure 3.4 the saturation is the lightness or brightness of the color. D3.js styling saturation can have values ranging between 0 and 1. In this research, a saturation value of 0.01 is chosen for each color membrane. Since 0.01 * 100 = 1, technically, an overlap of over 100 figures can reach a saturation point. The higher range of saturation values for the values can impact positively on the success rates of our feature selection process.

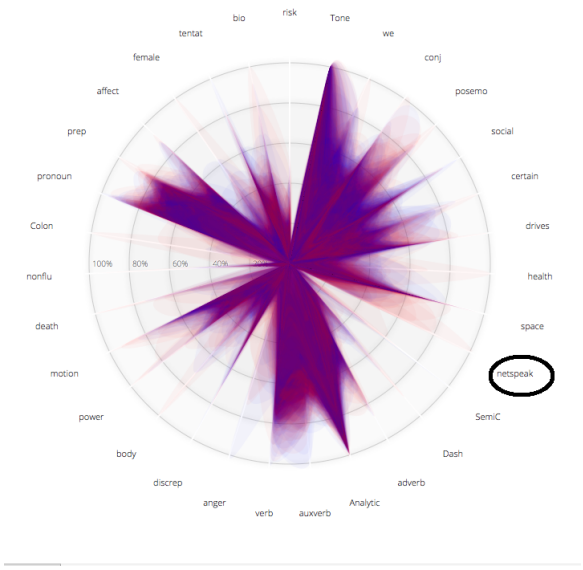
### 3.3.1 How to pick the best or worst dimensions based on colors

Case1: The more the red membranes overlap, the purer red manifests, or similarly, the more the blue membranes overlap, the purer blue manifests. If our perception can pick more red hues or blue hues for an attribute, it could be marked as an important attribute. In other words, the purer the colors are, the more significant the attribute is.

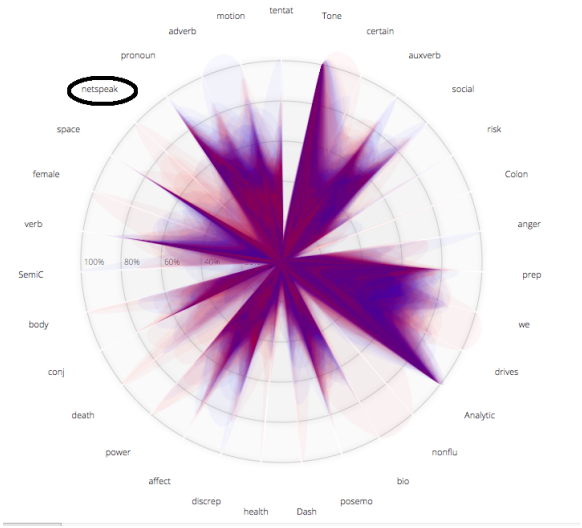
Fig 3.5 shows different iterations for the same set of dimensions. The highlighted *netspeak* dimension particularly shows pure red in both the iterations. It is scored higher in the feature selection process.

Case2: In another case, there is a clear blue, red patterns for an attribute. For example, more blue hues in a lower part of the attribute axis and clear red hues in the upper part of the attribute axis can be a good pattern and so that attribute can be considered to score high in feature selection process.

In figure 3.6, the attribute *OtherP* manifests more of red hues in the bottom layer and more of blue in the top in both the iterations. Irrespective of this behavior being




(a) First Iteration

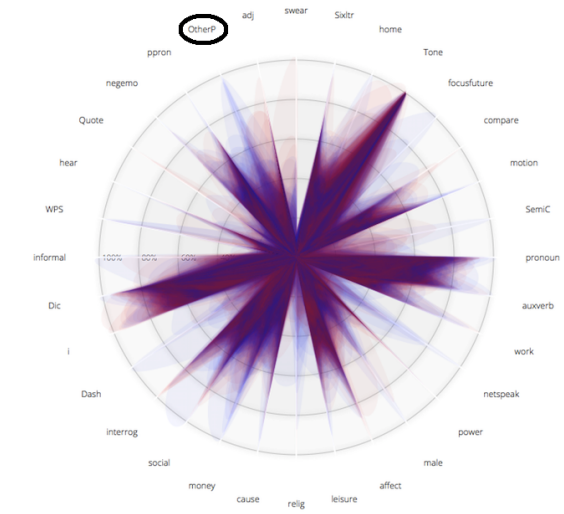



(b) Second Iteration

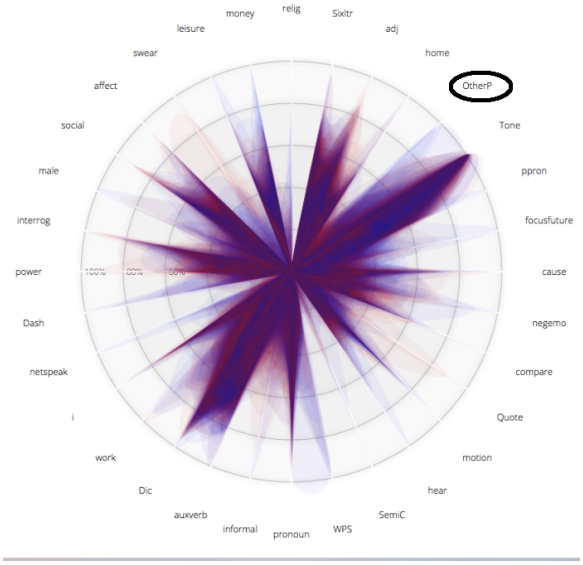
Fig. 3.5. Demonstration of pure color dimension

consistent in other iterations, this pattern can be considered a scoring factor for that particular iteration.

Fig 3.7 demonstrates the case where a dimension manifests a clear divide of the colors. The dimension *power* manifests more of red color in the first iteration while



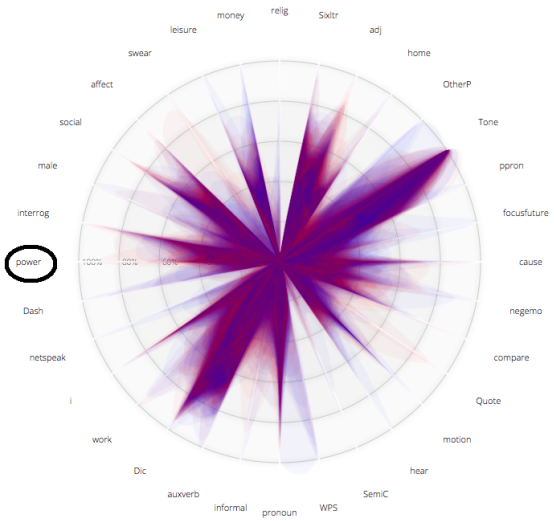

(a) First Iteration



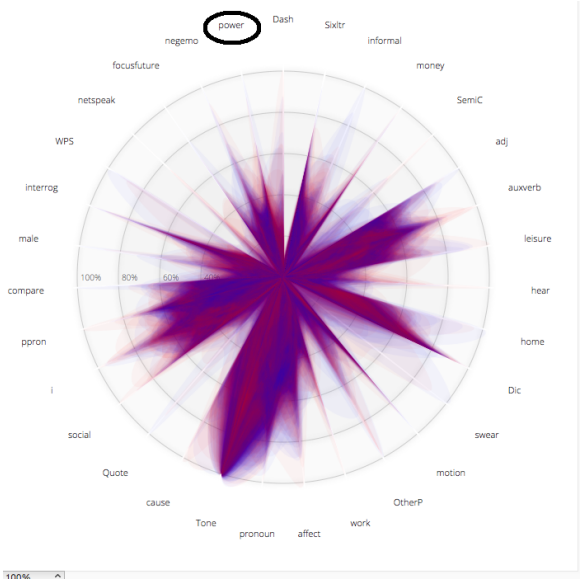

(b) Second Iteration

Fig. 3.6. Demonstration1 of clear color divide pattern

it manifests a clear divide in the second. It can be deduced that this attribute can be a potential feature for classification purpose.



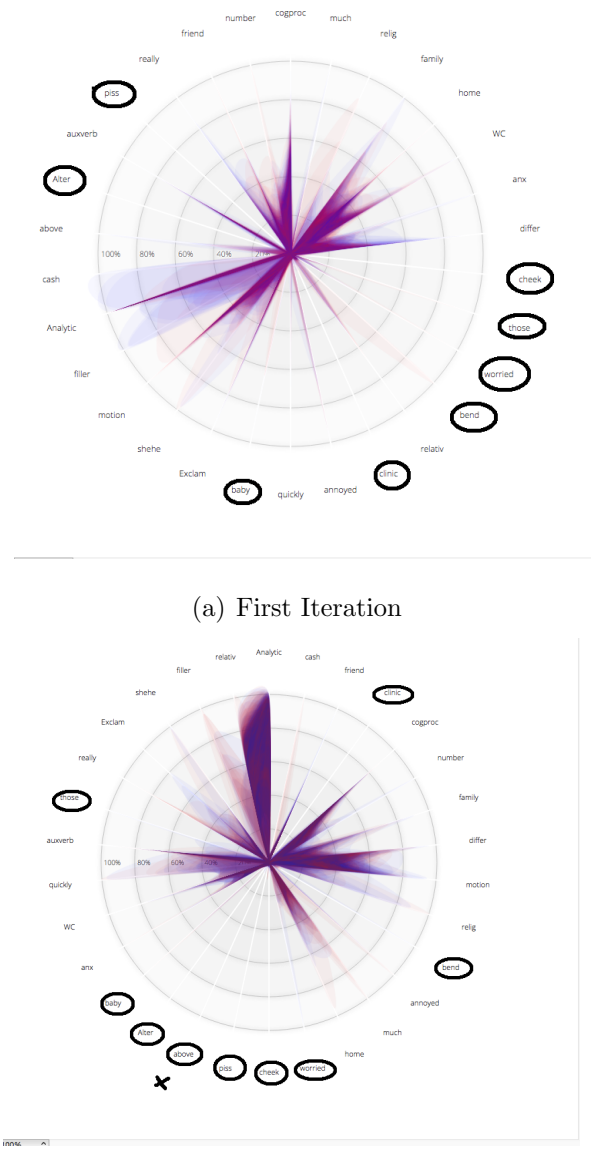

(a) First Iteration




(b) Second Iteration

Fig. 3.7. Demonstration2 of clear color divide pattern

Case3: Some have very light colors, hard to perceive or classify into one color are often values that have many zero values in the data. These attributes could be eliminated straight away.




(a) First Iteration




(b) Second Iteration

Fig. 3.8. Demonstration of obviously worthless attributes

In figure 3.7 the selected attributes in both the iterations exhibited similar characteristics of having no color at all, except for the *above* attribute. Though this attribute gets a negative score in one iteration, it does not receive one in the next. Similar observations are made in simultaneous iterations and the scoring either decreases or increases the worth of a particular attribute.

**Color Combinations**

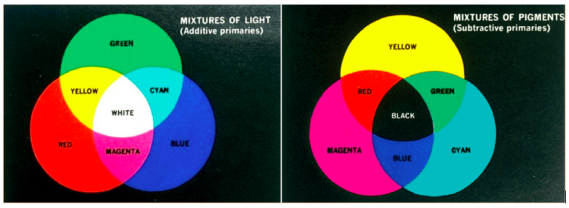



Fig. 3.9. Sample color blending palette

Figure 3.10, 3.11, 3.12, 3.13 are a demonstration of sample trials of assorted color combinations.

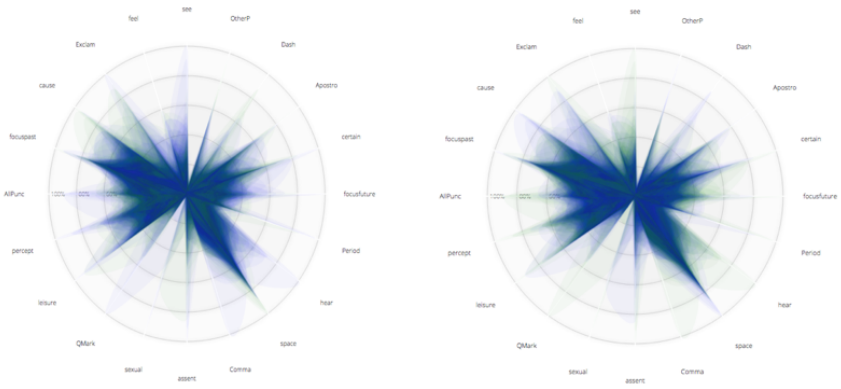


Fig. 3.10. Green + Blue = Cyan

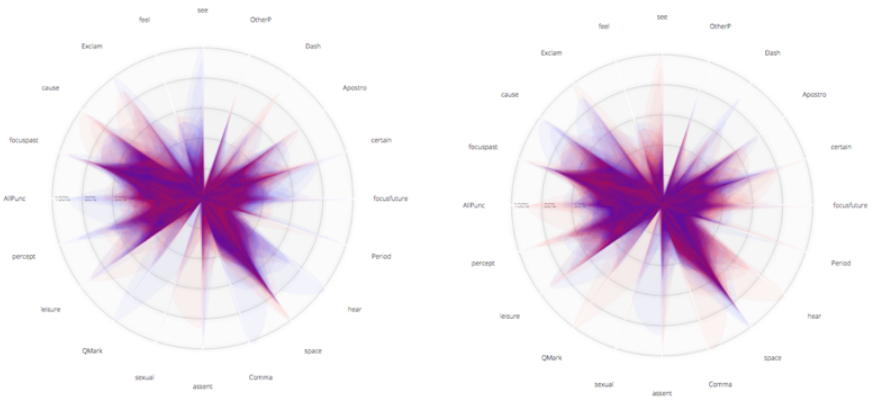



Fig. 3.11. Red + Blue = Magenta

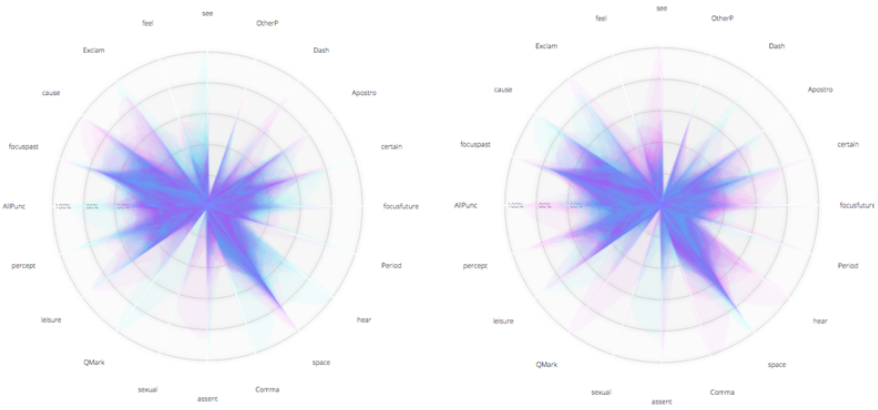



Fig. 3.12. Magenta + Cyan = Blue

## 3.4 Dimensions ordering and grouping

### 3.4.1 Significance of dimensions' order in the radial chart

After connecting the variable values on the radial chart, the output is a polygon. This implies that the shape of the resultant polygon is bound to change the positions of the attributes axis change. This inherent characteristic of the radial chart makes the

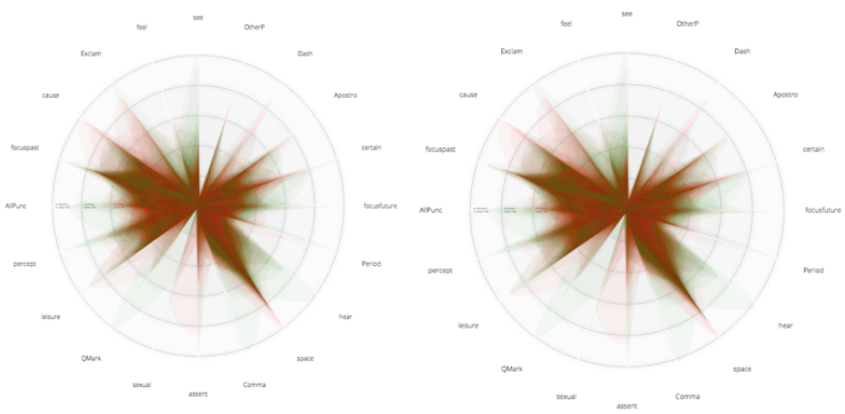


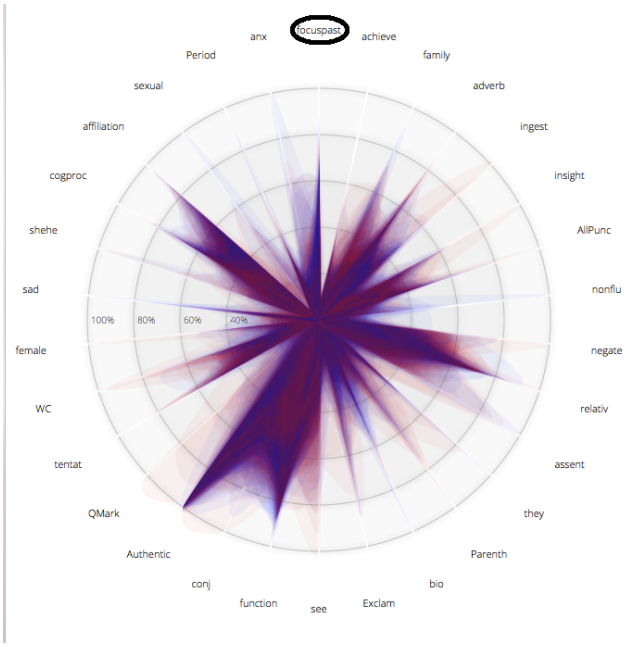

Fig. 3.13. Red + Green = Yellow

dimension or attribute ordering significant in this work and can be demonstrated by the following example. In the fig 3.14, the dimension or attribute *focupast* manifests blue color in one order while it manifests more of red in the second. This is because the radial polygon membranes differ in shape at various positions of the dimensions axes. Hence the number of dimension orderings can result in more accurate observations.

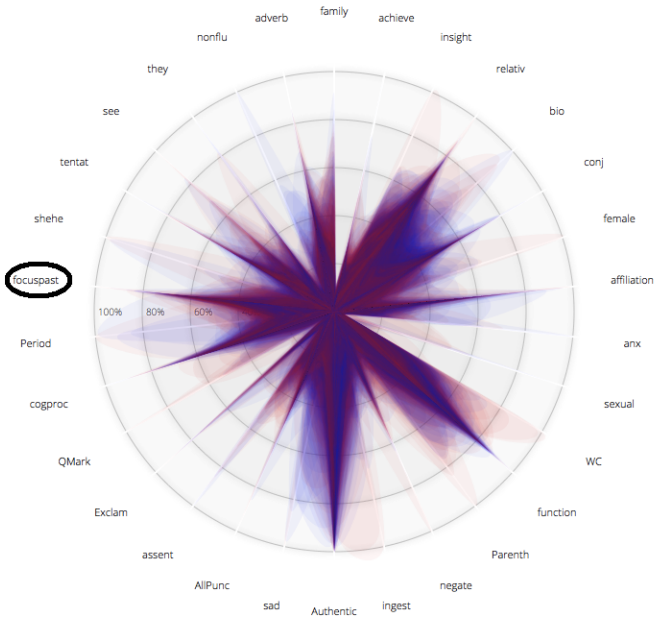
### 3.4.2 Significance of dimensions' grouping

Dimension selection is about picking different combinations of dimensions to project at once on the radial chart. As mentioned in the previous section, the behavior of an attribute is bound to exhibit varying behaviors in the presence or absence of different dimensions. Hence it is very crucial to take observations of behavior in different combinations sets.

In the figure 3.15, the two figures demonstrate the behavior of the attribute *family* in different combinations of dimensions. As obvious from the first figure, it shows the *family* seems not to exhibit any color at all. However, in the second figure, the same dimension seems to have the blue color. Hence it is important to have as many sets of different combinations of dimensions as possible for experimenting.

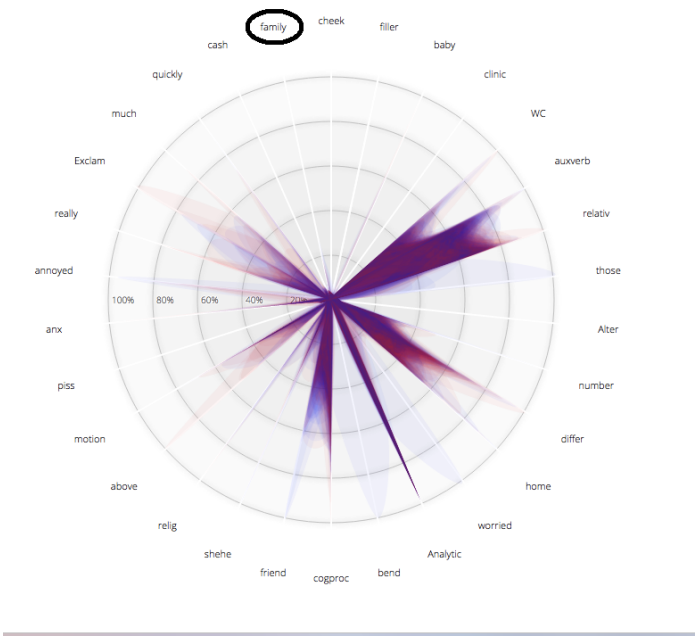



(a) First Iteration

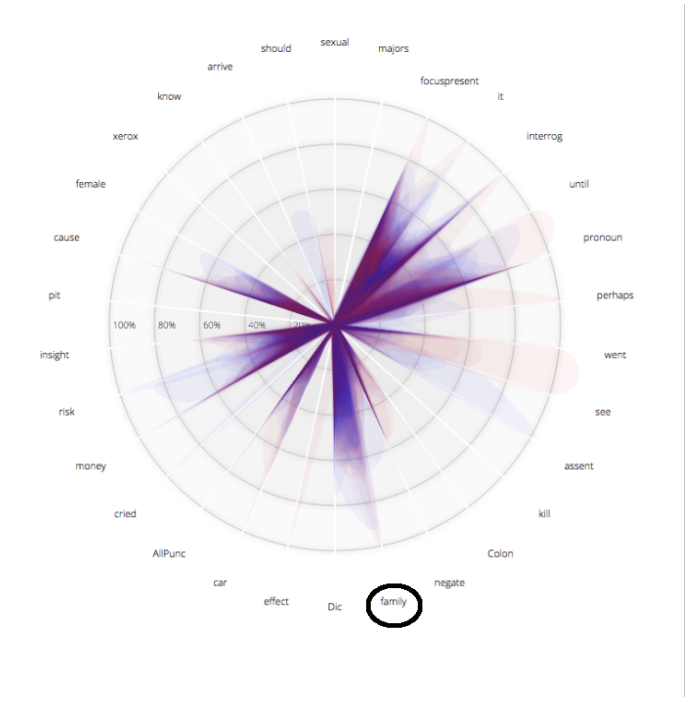



(b) Second Iteration

Fig. 3.14. Demonstration of dimension order significance




(a) First Set




(b) Second Set

Fig. 3.15. Demonstration of dimension combination significance

## 3.5 Process of feature selection

### 3.5.1 Shuffle and pick for dimensions' grouping

Since our data consists of 1600 reviews, there are 1600 polygons on each radial chart with at most 30 variables or axis. To pick different combinations of these dimensions, for every experiment, the dimensions are divided by 30 to get n number of sets. The dimensions are selected from the whole set randomly to fill in these n sets. To ensure that each set consists of different combinations of dimensions this process is repeated thrice. So if the whole data set consists of m number of dimensions, then the number of smaller sets = m/30 * 3.For example, if the total number of dimensions is 90 the number of small sets = 90/30 * 3 = 9 small sets that are fed in for the radial visualization.

### 3.5.2 Shuffle for dimension orderings and visualize

After dividing the whole data set into the smaller sets of around 30 dimensions each set is shuffled to change the positions of the dimensional axis. This is done multiple times ranging from 5 to 10 in different rounds of this experiment. For instance, if there are nine small sets with ten permutations of the same set we get 90 different radial charts or in general terms m/30*3*10.

### 3.5.3 Score the dimensions and final elimination or selection of features

For each of these visualizations, the observations are noted. Microsoft Excel is used for this purpose. The observation list consists of m columns (number of total dimensions) and m/30*3*10 rows to note down the observations. If it is indecisive for a dimension in a specific iteration, it is noted 0, if its pure blue it is positive 10 or if its pure red it is negative 10. But if its clear -divide it is 100. An Ideal feature cannot be close to 0. Some blues have more positive multiples of 10, the number of reds the more negative multiples of 10. If a certain dimension exhibits blues and red,

it is bound to have values close to 0. However, such dimensions tend to exhibit clear divide and so get picked up as potential features.

All the values having values nearer to 0 can be eliminated. After repeating the process couple of times with the filtered attributes based on the initial number of dimensions, the final list of features can be selected or can be used just for the dimension reduction purpose. For dimension reduction, the least scored attributes can be eliminated straight away.

# 4. VISUAL FEATURE SELECTION

This chapter explains the different approaches that are experimented. Before discussing the different approaches, it is essential to understand the other possible brute force approaches. One of the approaches that might seem like an attractive option is to combine all the fake reviews into one enormous chunk of fake review text and the combining all the genuine reviews into one large piece of genuine review text. These reviews are then preprocessed through LIWC that outputs data with two records and default 93 dimensions. The visual feature selection process would not make much sense because the resulting features would only be the top dimensions that have the higher variance of the values for both fake and genuine. This process is like picking the attributes that have the most differing values, which can be proofed with some observations in the following chapter.

## 4.1 Approach1, Identification of best features

The initial approach is to follow the theoretical assumptions explained in the previous section and use that to pick the possible features. Approach1 aims at observing the trends like pure color hues and clear divides to score and rank each dimension. After several rounds of observations and iterations, the final score for each dimension is calculated. Based on the rankings gained by each of the dimension and certain threshold, the final features are selected. The focus of this approach is to choose the best attributes i.e.; observations are made with the only aim of picking the attributes or dimensions that most fit the constraints of an ideal feature.

After preprocessing of the data through LIWC, the visual feature selection process is done. The total data consists of 1600 records. 1400 records are selected initially for the training purpose. The total number attributes are divided by 30 to get smaller

sets of data attributes. As explained in the section 3.5.1, there will be around m/30*3 small subsets of data attributes, where m is the total number of dimensions in that experiment. 1400 records are split into random subsets ensuring there is a good shuffle for dimensions' grouping. The subsets are shuffled individually for ensuring the dimension order and are visualized over the radial chart. The observations involve scoring each dimension based on how ideal the dimension is for that observation. As the behavior of a dimension is bound to change in the presence or absence of other dimensions a higher number of observations would yield in higher accuracy. This approach involves observing almost all the dimensions to see if it falls under the ideal feature category.

## 4.2 Approach2, Elimination of worst features

Approach2, unlike the approach1, focuses on eliminating obviously hard to analyze dimensions. Approach2 is an easier process compared to approach1 as it is less straining on the user. All the user needs to do is just select the dimensions in every iteration that seem to be of no value in the visualization. With every selection made, the score of the dimension keeps increasing. In the final scoring, the dimensions with more number of scores are the ones that are selected the most number of times as indecisive. Hence those dimensions are eliminated.

## 4.3 Extension of Dimensions

The default output of LIWC has 93 categories, which has its group of keywords in it. But since in practical cases, we may want to extend these categories to our defined set of categories with keywords, there is an option provided by LIWC for this purpose. This functionality helps define user defined dictionaries to add as many numbers of categories as possible. Thereby extending the dimensions. In this work, a dictionary of nearly 100 attributes or dimensions is created, increasing the total dimensions to around 200.

Initial experiments are conducted on the default LIWC 93 dimensions. These 93 attributes are categories that aid in the psychological analysis of the reviews and are themselves a best-reduced set of features. Since our experiment is to prove that our proposed visualization techniques help in the process of feature selection for classification purpose, it makes more sense to experiment with the more diluted set of attributes than on the default 93 attributes that are already best-proved feature set. It was important to experiment on the best 93 attributes to explore the difficulty levels.

LIWC provides functionality to create our dictionary with our categories. For this purpose, some least popular categories among the default 93 categories are picked, and the words that fall under that category are each given a category in the user defined dictionary. In this way, the initial 93 attributes could be increased to 200 with the addition of these extra categories.

# 5. EXPERIMENTS, RESULTS, AND ANALYSIS

## 5.1 Experiments

### 5.1.1 Approach1

Approach1 is performed on 1400 records that consist of 93 dimensions each. The 93 columns are the default output of LIWC. The data consists of normalized attributes values of corresponding to the LIWC output of the reviews. The remaining 1200 reviews are left aside purely as a test set. This experiment involves all the four steps as discussed in section 3.5. Approach1 being the very first experiment it helps exploring the technique and its observations. Approach1 is repeated thrice with all the four steps. Shuffle and pick, shuffle and order, score the dimensions, final ranking of attributes.

### 5.1.2 Approach2

Approach2 which focuses on dimension reduction through worst attributes elimination is performed on both 93 dimensions and 200 dimensions.

- On 93 dimensions: The same four steps are performed on the 93 attributes of 1400 records.
- On 200 dimensions: The same four steps are performed on the 200 dimensions with 1400 records.

## 5.2 Results

### 5.2.1 Approach1

#### Iteration1

The 93 attributes are divided into three sets displaying nearly 30 attributes at any given point on the radial chart. Each round has different dimension ordering permutations of the selected dimensions. Roughly around ten rounds and three shuffles for different combinations of these dimensions are performed. Total plots of observations for this experiment are around 90. The primary intention of this iteration is to observe the behavior of the attributes on the radial charts. The attributes that have a better purity in colors are considered important in our work, which resulted in selecting almost 80% of the attributes, and the selected 80% had even lower efficiency than the first 93 attributes. This iteration has been exploratory.

*Key Leanings*:

1. The order of the dimensions' matter: A dimension which looked redder might not be likely to appear with same purity in color for other combinations or permutations of the dimensions and a dimension that might have seemed colorless might be appearing with a prominent one in the next order.

2. The combination of the dimensions' matter: A dimension which has a significant shape plotted on the graph might not have an obvious one in combination with other dimensions. And a dimension might seem to appear one color might appear as another color in combination with other dimensions.

3. The color patterns matter: If the observation is obviously blue or red, that dimension may be considered valuable. In some cases, the dimension has an apparent transformation from one color to the other. The lower part might be prominently one color while the upper part is prominently another. In some

other dimension ordering or combination, a dimension might reveal a clear divide in the two colors.

**Iteration2**

Observations based on the key learnings from iteration1 are applied during the feature selection process. Could pick up to 49 attributes with an efficiency of 68.375% in *LibSVM Weka*.

**Iteration3**

Picked 52 features with a classification efficiency of 71.875% in *LibSVM Weka*.

### 5.2.2 Approach2

**On 93 Dimensions**

Approach2 primarily focuses on dimension reduction through the elimination of apparently worst dimensions. This experiment is to explore if elimination of worst instead of the selection of best features could work. Iteration1 could reduce the 93 attributes to 80 attributes with an efficiency of 76.25% achieved, best so far. Iteration2 could reduce the 93 to 68 attributes with an efficiency of 75.833%.

**On 200 Dimensions**

The number of dimensions is increased to 200 by adding some more keywords that might individually not have a significant impact on the classification process. The primary 93 attributes are supposed to be the best attributes produced by the LIWC. As a result, the difference in the different set of attributes efficiencies is more or less the same. The primary 93 attributes are diluted by adding 100 other attributes which are keywords, unlike the 93 primary attributes that are more categorical.

LIWC has an option of creating user-defined dictionaries. The original 93 dimensions are the categories that have some words under each category. The additional set of 100 dimensions include the words that are categorized under the least popular categories. The efficiency is 70.75% for all the 200 dimensions included. Iteration1 could reduce the number of attributes to 106 from 200, and the efficiency of the classification increased to 76.75%, which is same as that of the features selected from Information Gain feature selection method on 200 attributes.

### 5.3 Analysis

This work, as mentioned earlier, aims to explore the possibility of visual feature selection. Hence in this section, analysis and learning from different outcomes of the experiments performed are discussed.

Figure 5.3 is a table of 93 default LIWC attributes and the features selected from the different approaches and iterations. Figure 5.4 is a visual presentation of the results of various outcomes. To show how the initial 93 LIWC attributes are themselves close to being the best features, different comparisons are drawn. For example *All Attributes Minus Info Gain* column is the resultant set of all 93 attributes minus the features selected from Information Gain (InfoGain) feature selection method in Weka, which has given the worst percentage thus far of LibSVM classification resulting in 67% with 63 number of features. The *InfoGain* in Weka method has picked 29 attributes which have given the LibSVM classification accuracy of 74.84% which is ideally our benchmark to compare our results to.*93 Attributes Included* column is the result of running the LibSVM classification with all the 93 attributes included which is 76%, even better than that of *InfoGain* result. *XXX attributes* column is the LibSVM Weka classification accuracy results of the features proved best by [21], which are selected from several mining and analysis methods, which has also given the accuracy of 76% equaling the result of all the 93 attributes included, with only 26 features, by far the best result.

As presented in figure 5.4, the Approach1 used for picking the best features could take up the results from 67% to 71.88% in the iteration2 confirming the possibility of visually identifying the best features. However, not being able to equal the accuracy results gained by *InfoGain* point out that this approach is not satisfactory, at least with our current visualization technique. It can also be observed that the number of features being selected matter in improving the accuracy, as in iteration1 of approach1 the number of attributes picked are only 49, while in the iteration2 the number is 52 with a difference of 3.8% in the accuracy.

Approach 2 is apparently successful in getting an improved efficiency over *InfoGain* accuracy. Though in the iteration1 the accuracy enhanced by only a small amount, it is still significant in our case due to the nature of our data. Even better is with the iteration2, in which we could gain the best accuracy by far, of 76.62%, 2 percent more than that of InfoGain. This approach is promising as it could even achieve improved results even in comparison to some of the best research methods. The *XXX attributes* results re-assure this.

Figure 5.5 is the table presenting all the 200 attributes and the features selected in different experiments. In this experiment, all our approaches showed promising results. *200Attributes* column in figure 5.6 shows the LibSVM Weka classification results of all the 200 attributes included, which is 70.25%. *InfoGainOn200Attributes* column presents the LibSVM Weka classification result of Weka InfoGain feature selection method which led to 79% of accuracy with 31 features.

Approach2 could significantly reduce the number of dimensions from 200 to 106 with an overall improvement in the LibSVM Weka classification accuracy of 76.75%, which is an improvement from 70.25% with all the 200 attributes included. However, since the InfoGain on 200 attributes is 79%, it surely needed another round of iteration, but eliminating some more attributes might risk, the elimination of a few best attributes. Hence, performed Weka InfoGain feature selection process for the selected 106 features. The result presented by the *InfoGainOn106Attributes* column is 79%

same as the original InfoGain on 200 attributes but with reduced number of features 27, as compared to 31 attributes picked by the initial InfoGain.

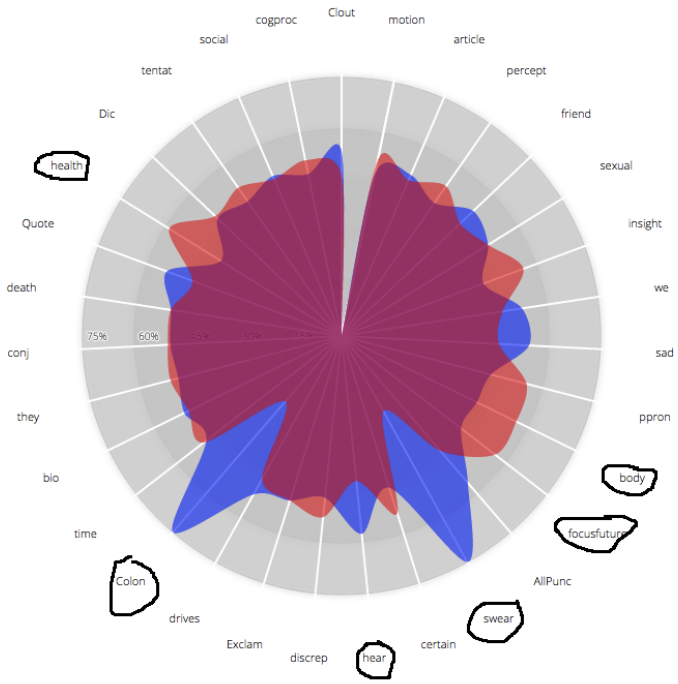
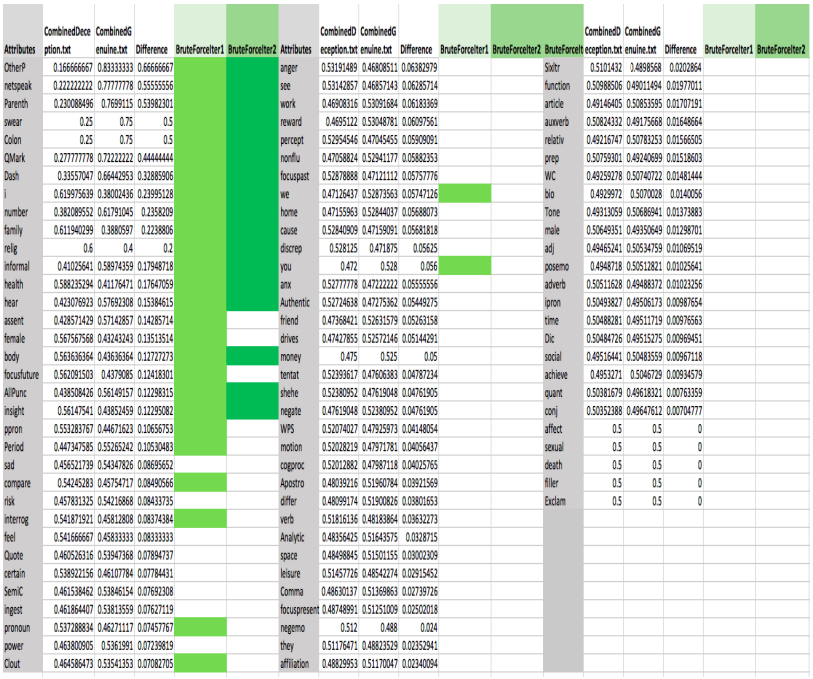



Fig. 5.1. Selected features from experiments on the 93 attributes



| Attributes | CombinedDeception.txt | CombinedGenuine.txt | Difference | BruteForceIter1 | BruteForceIter2 |
|---|---|---|---|---|---|
| OtherP | 0.166666667 | 0.83333333 | 0.66666667 | | |
| netspeak | 0.222222222 | 0.77777778 | 0.55555556 | | |
| Parenth | 0.230088496 | 0.7699115 | 0.53982301 | | |
| swear | 0.25 | 0.75 | 0.5 | | |
| Colon | 0.25 | 0.75 | 0.5 | | |
| QMark | 0.277777778 | 0.72222222 | 0.44444444 | | |
| Dash | 0.33557047 | 0.66442953 | 0.32885906 | | |
| I | 0.619975639 | 0.38002436 | 0.23995128 | | |
| number | 0.382089552 | 0.61791045 | 0.2358209 | | |
| family | 0.611940299 | 0.3880597 | 0.2238806 | | |
| relig | 0.6 | 0.4 | 0.2 | | |
| informal | 0.41025641 | 0.58974359 | 0.17948718 | | |
| health | 0.588235294 | 0.41176471 | 0.17647059 | | |
| hear | 0.423076923 | 0.57692308 | 0.15384615 | | |
| assent | 0.428571429 | 0.57142857 | 0.14285714 | | |
| female | 0.567567568 | 0.43243243 | 0.13513514 | | |
| body | 0.563636364 | 0.43636364 | 0.12727273 | | |
| focusfuture | 0.562091503 | 0.4379085 | 0.12418301 | | |
| AllPunc | 0.438508426 | 0.56149157 | 0.12298315 | | |
| insight | 0.56147541 | 0.43852459 | 0.12295082 | | |
| ppron | 0.553283767 | 0.44671623 | 0.10656753 | | |
| Period | 0.447347585 | 0.55265242 | 0.10530483 | | |
| sad | 0.456521739 | 0.54347826 | 0.08695652 | | |
| compare | 0.54245283 | 0.45754717 | 0.08490566 | | |
| risk | 0.457831325 | 0.54216868 | 0.08433735 | | |
| interrog | 0.541871921 | 0.45812808 | 0.08374384 | | |
| feel | 0.541666667 | 0.45833333 | 0.08333333 | | |
| Quote | 0.460526316 | 0.53947368 | 0.07894737 | | |
| certain | 0.539922156 | 0.46107784 | 0.07784431 | | |
| SemiC | 0.461538462 | 0.53846154 | 0.07692308 | | |
| ingest | 0.461864407 | 0.53813559 | 0.07627119 | | |
| pronoun | 0.537288834 | 0.46271117 | 0.07457767 | | |
| power | 0.463800905 | 0.5361991 | 0.07239819 | | |
| Clout | 0.464586473 | 0.53541353 | 0.07082705 | | |

| Attributes | CombinedDeception.txt | CombinedGenuine.txt | Difference | BruteForceIter1 | BruteForceIter2 |
|---|---|---|---|---|---|
| anger | 0.53191489 | 0.46808511 | 0.06382979 | | |
| see | 0.53142857 | 0.46857143 | 0.06285714 | | |
| work | 0.46908316 | 0.53091684 | 0.06183369 | | |
| reward | 0.4695122 | 0.53048781 | 0.06097561 | | |
| percept | 0.52954546 | 0.47045455 | 0.05909091 | | |
| nonflu | 0.47058824 | 0.52941177 | 0.05882353 | | |
| focuspast | 0.52878888 | 0.47121112 | 0.05757776 | | |
| we | 0.47126437 | 0.52873563 | 0.05747126 | | |
| home | 0.47155963 | 0.52844037 | 0.05688073 | | |
| cause | 0.52840909 | 0.47159091 | 0.05681818 | | |
| discrep | 0.528125 | 0.471875 | 0.05625 | | |
| you | 0.472 | 0.528 | 0.056 | | |
| anx | 0.52777778 | 0.47222222 | 0.05555556 | | |
| Authentic | 0.52724638 | 0.47275362 | 0.05449275 | | |
| friend | 0.47368421 | 0.52631579 | 0.05263158 | | |
| drives | 0.47427855 | 0.52572146 | 0.05144291 | | |
| money | 0.475 | 0.525 | 0.05 | | |
| tentat | 0.52393617 | 0.47606383 | 0.04787234 | | |
| shehe | 0.52380952 | 0.47619048 | 0.04761905 | | |
| negate | 0.47619048 | 0.52380952 | 0.04761905 | | |
| WPS | 0.52074027 | 0.47925973 | 0.04148054 | | |
| motion | 0.52028219 | 0.47971781 | 0.04056437 | | |
| cogproc | 0.52012882 | 0.47987118 | 0.04025765 | | |
| Apostro | 0.48039216 | 0.51960784 | 0.03921569 | | |
| differ | 0.48099174 | 0.51900826 | 0.03801653 | | |
| verb | 0.51816136 | 0.48183864 | 0.03632273 | | |
| Analytic | 0.48356425 | 0.51643575 | 0.0328715 | | |
| space | 0.48498845 | 0.51501155 | 0.03002309 | | |
| leisure | 0.51457726 | 0.48542274 | 0.02915452 | | |
| Comma | 0.48630137 | 0.51369863 | 0.02739726 | | |
| focuspresent | 0.48748991 | 0.51251009 | 0.02502018 | | |
| negemo | 0.512 | 0.488 | 0.024 | | |
| they | 0.51176471 | 0.48823529 | 0.02352941 | | |
| affiliation | 0.48829953 | 0.51170047 | 0.02340094 | | |

| Attributes | CombinedDeception.txt | CombinedGenuine.txt | Difference | BruteForceIter1 | BruteForceIter2 |
|---|---|---|---|---|---|
| Sixltr | 0.5101432 | 0.4898568 | 0.0202864 | | |
| function | 0.50988506 | 0.49011494 | 0.01977011 | | |
| article | 0.49146405 | 0.50853595 | 0.01707191 | | |
| auxverb | 0.50824332 | 0.49175668 | 0.01648664 | | |
| relativ | 0.49216747 | 0.50783253 | 0.01566505 | | |
| prep | 0.50759301 | 0.49240699 | 0.01518603 | | |
| WC | 0.49259278 | 0.50740722 | 0.01481444 | | |
| bio | 0.4929972 | 0.5070028 | 0.0140056 | | |
| Tone | 0.49313059 | 0.50686941 | 0.01373883 | | |
| male | 0.50649351 | 0.49350649 | 0.01298701 | | |
| adj | 0.49465241 | 0.50534759 | 0.01069519 | | |
| posemo | 0.4948718 | 0.50512821 | 0.01025641 | | |
| adverb | 0.50511628 | 0.49488372 | 0.01023256 | | |
| ipron | 0.50493827 | 0.49506173 | 0.00987654 | | |
| time | 0.50488281 | 0.49511719 | 0.00976563 | | |
| Dic | 0.50484726 | 0.49515275 | 0.00969451 | | |
| social | 0.49516441 | 0.50483559 | 0.00967118 | | |
| achieve | 0.4953271 | 0.5046729 | 0.0093458 | | |
| quant | 0.50381679 | 0.49618321 | 0.00763359 | | |
| conj | 0.50352388 | 0.49647612 | 0.00704777 | | |
| affect | 0.5 | 0.5 | 0 | | |
| sexual | 0.5 | 0.5 | 0 | | |
| death | 0.5 | 0.5 | 0 | | |
| filler | 0.5 | 0.5 | 0 | | |
| Exclam | 0.5 | 0.5 | 0 | | |

Fig. 5.2. Brute force approach demonstration

### 5.3.1 Key findings in a nutshell

#### a. Brute force approach doesn't need visualization

All the fake reviews are combined into one sizable chunk of Fake review text and all the genuine reviews into one big piece of Genuine review text. These two texts are preprocessed by running through the LIWC and normalizing. The resultant output is data with only two records and default 93 dimensions. The visual feature selection process involves picking the dimensions with pure colors or eliminating the ones which are impure. Since there are only two rows, each represented by blue and red color, there are just peaks of pure colored dimensions that can be picked making the visualization look very simple and easy to pick the dimensions as shown in figure 5.1.

The rounded dimensions are the ones that are marked for scoring in that particular iteration.As mentioned in section 4 this approach is nothing more than selecting

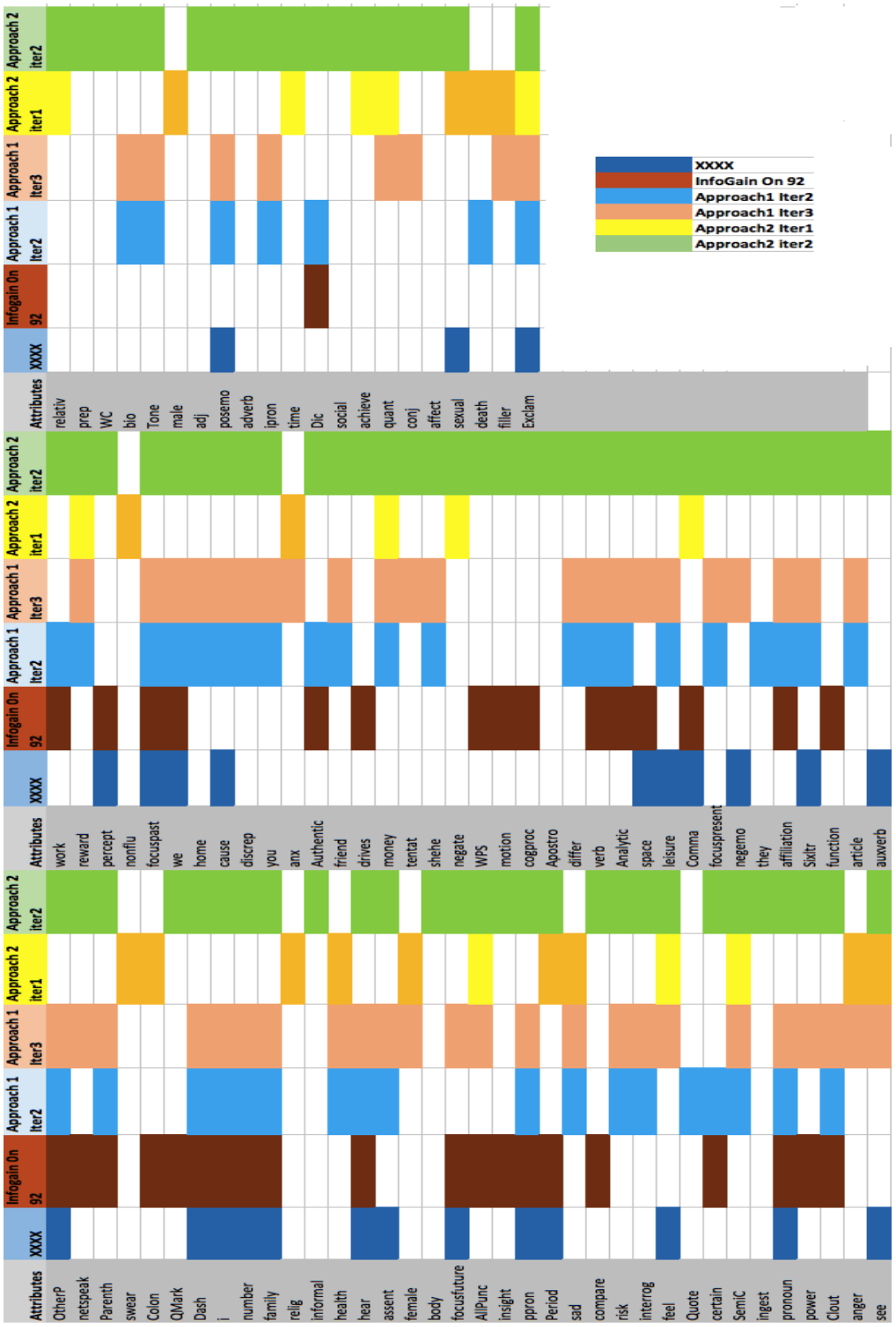
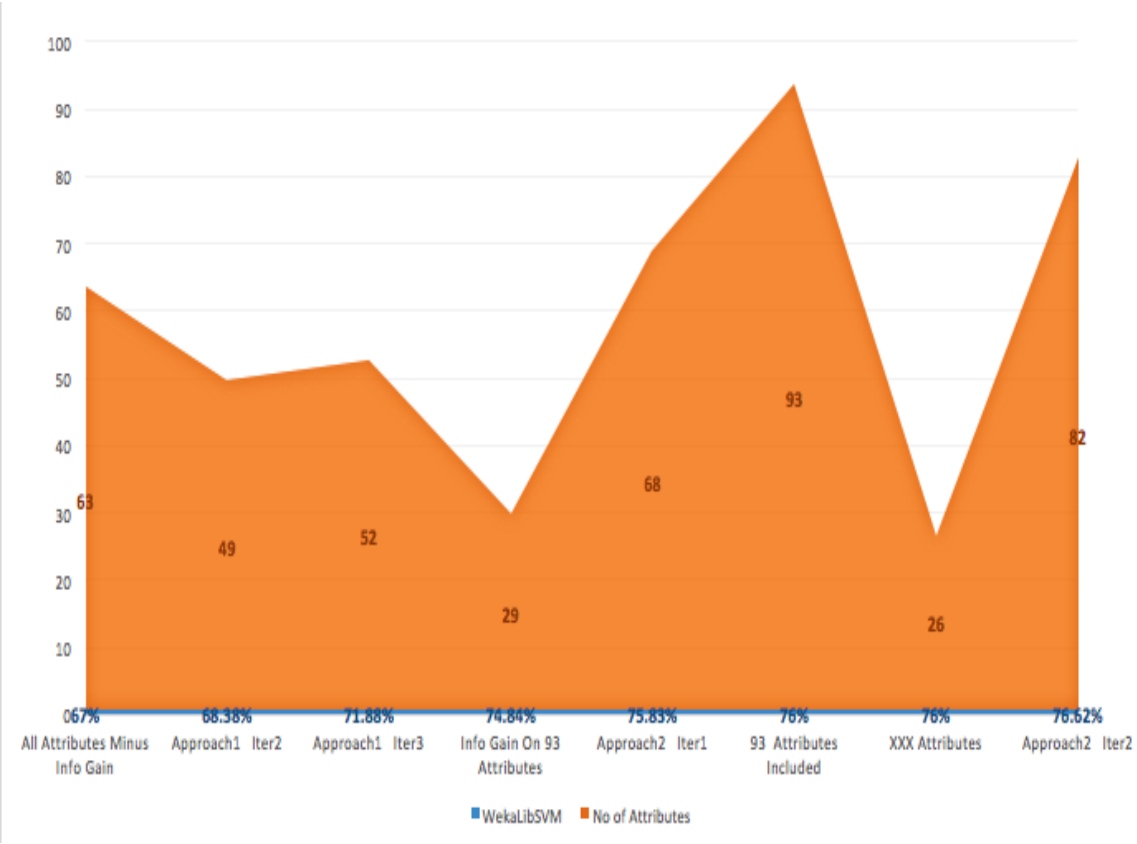
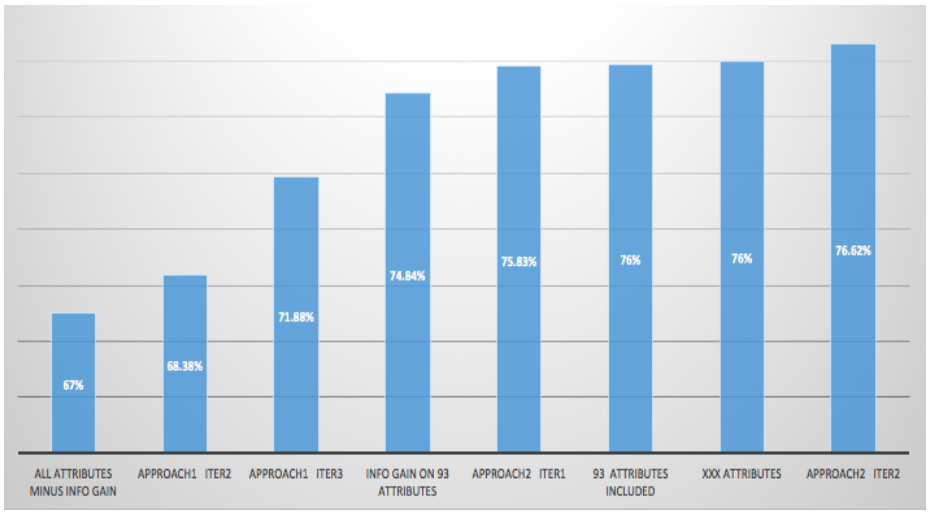



Fig. 5.3. Selected features from experiments on the 93 attributes




(a) Number of Attributes




(b) Weka LibSVM Accuracy

Fig. 5.4. Results of experiments on 93 attributes

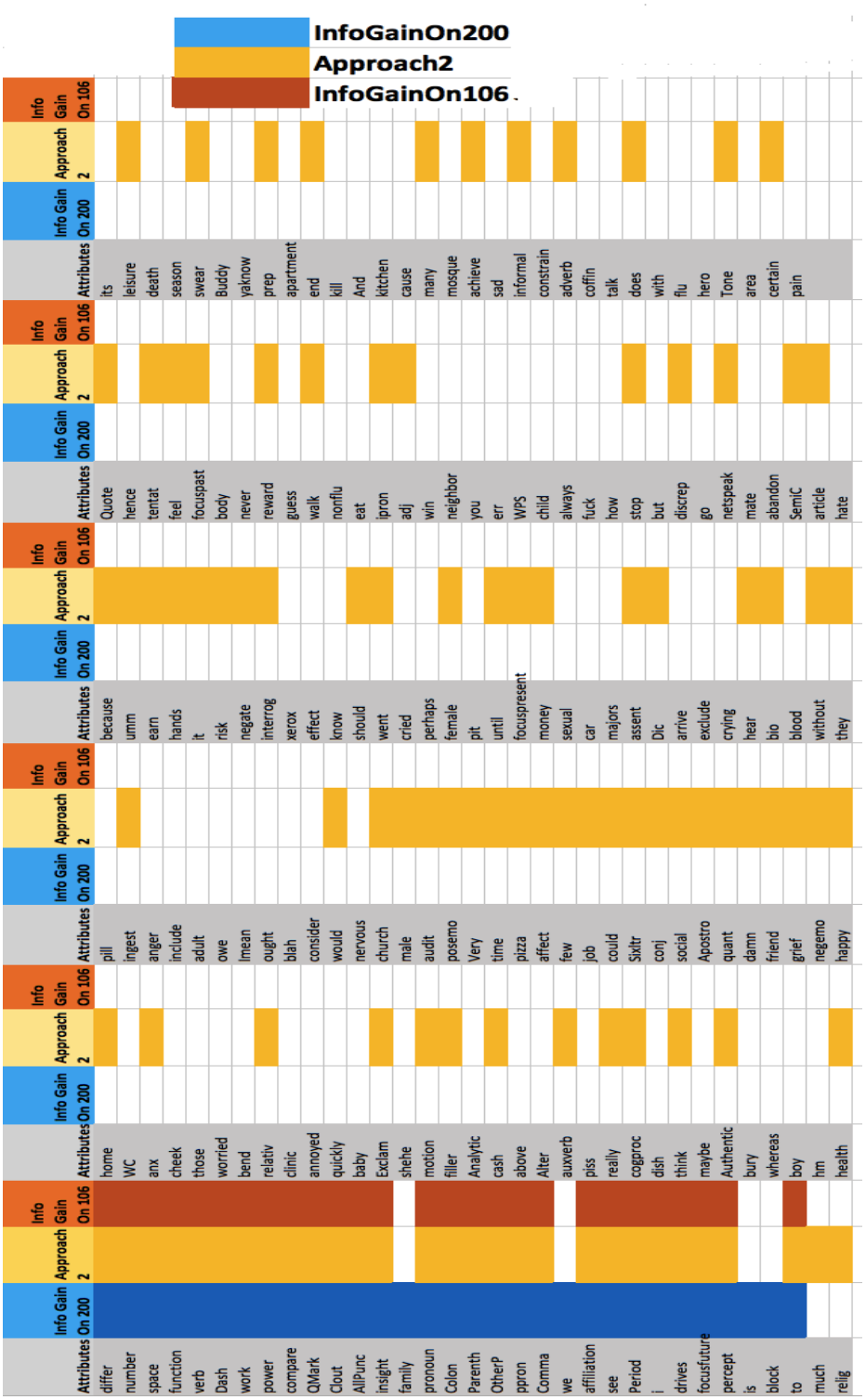
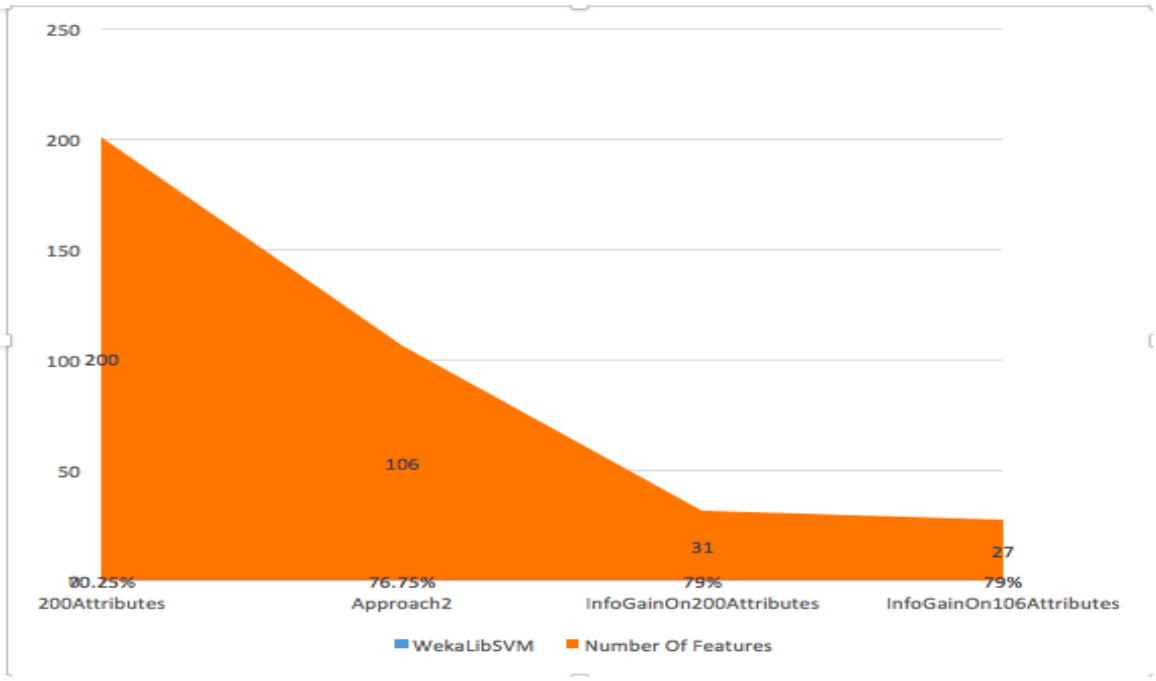
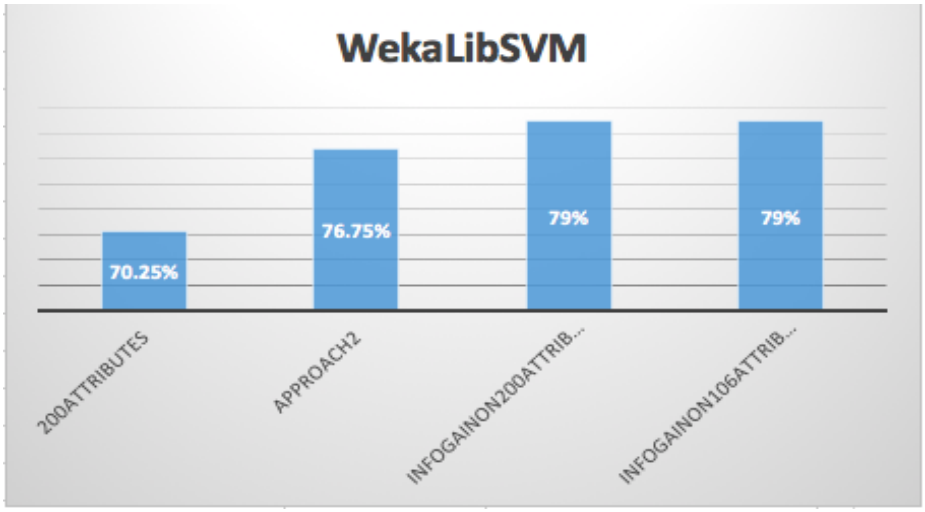



Fig. 5.5. Selected features from experiments on the 200 attributes




(a) Number of Attributes




(b) WekaLibSVM Accuracy

Fig. 5.6. Results of experiments on 200 attributes

the dimensions whose values differ a lot in the two categories. In figure 5.2 the dimensions are arranged in the descending order from higher difference values to lower difference values. In the figure 5.2 the columns *BruteForceIter1* and *BruteForceIter2*

the colored cells are the ones selected through visual feature selection for this brute force. Evidently, the features selected through visual features selection and the attributes presented in figure 5.2 confirm that simple mathematical approach makes no sense. The result of the first iterations is 74.53% with 26 attributes, while the second iteration is 72.86% with 17 attributes selected.

Although this has a higher efficiency rate than that of the approach1, it does not make sense to visualize this data, as the text reviews are a lot more complicated than mere average or combined higher differing dimensions. Each of the reviews is a unique text written by different individuals, which makes more sense to study the single review structure like done in our approach1 and approach2. The whole idea to present the reviews in a visualization is to be able to achieve the involvement of each review structure.

**b. Pros and cons of Approach1**

1. Though Approach1 could help pick most the best attributes it could also eliminate some, a reason attributed to the fact that, users tend to get confused while picking the best features in this visual technique. The values of such dimensions are concentrated revealing confusing impure colors.

2. This can be overcome by ensuring that every dimension especially the ones not manifesting obvious color purity or patterns are observed in several possible groupings and orderings to lower the risk of elimination of potential features.

3. Approach1 takes more time and is also more stressful from the user point of view because of the number of observations and time taken to observe each dimension in every visualization.

### c. Approach2

1. Approach2 has helped us in reaching best results so far. On 93 attributes the approach2 could achieve higher classification accuracy. On the 200 attributes, approach2 could improve the accuracy of the initial 200 dimensions included, reducing the number of dimensions to nearly half and with automatic feature selection process on the reduced set of dimensions could perform even better, with the least number of features yet, in this experiment.

2. This approach is easier from the user point of view, as it does not involve observation of each dimension. Only the dimensions that are obviously useless are eliminated.

### d. Noteworthy points

1. In either approach, there is no threshold on how many features are to be selected or eliminated. Since visual feature selection is about the perception of the user, it is recommended to eliminate the obviously less useful than pushing too hard to eliminate more.

2. The nature of the dimensions impact the visual feature selection process. Experimenting with the 93 attributes and 200 attributes is an example for this. As the 93 attributes already represent best attributes the visual feature selection process on them was not efficient nor secure. But for the 200 attributes, which are much more diluted, the process was very efficient as well as secure.

3. This work has no particular threshold value on the number of shuffles for dimension orders nor on the number of iterations for dimension groupings.

# 6. CONCLUSIONS

For some applications, visual feature detection is preferred as human expertise is valuable or even necessary in interpreting the potential features (e.g. in medical diagnosis). This thesis shows that with a carefully designed visual representation, visual feature detection can be as effective as the best automatic feature detection methods. Gold standard on-line fake reviews and genuine reviews obtained from [21] Are used to explore feature selection for classification through the presented approaches. The statistical, data mining algorithms are often used to select the best features that contribute to the classification purpose. However, from this work, it is evident that visual feature detection can be as effective as the best automatic feature detection methods, though it cannot completely replace the statistical, mathematical or mining algorithms. This work presents different approaches of visualizing the on-line reviews with the failure and success cases.

Main contributions in this work are:

1. The visualization technique: The concept and implementation of visual feature selection by presenting in a radial format through radial color overlaps. 2. Application of the technique to online review classification: Giving structure to the text reviews, applying the visualization technique for the best feature selection. 3. Demonstrated that visualization technique is at least as powerful as the automated feature selection methods.

The implementation of visualization using radial chart and color blending to give a structure to on-line reviews to help in the feature selection process has been exploratory and could present some key observations and results. Overall, it is evident from this work that manual intervention through visualization during the classification process is promising, though it cannot replace the statistical, mathematical or mining algorithms. In this work, we could prove that visual feature selection is at

least as powerful as the best automatic feature selection methods. From this study, the following conclusions are drawn:

1. The ordering of the dimensions has a significant impact on the best feature selection i.e.; dimensions behave differently in combination with different dimensions

2. As the number of shuffling for the dimensions increase, the more accurate the observations can be.

3. At most 30 dimensions look better on screen for visualizing in this technique at any given point in time.

4. The color combinations to represent the categories have a significant impact.

5. The saturation levels for the color blending also plays a key role.

6. The best feature selection is difficult, as the purity or impurity or colors for a specific dimension gets challenging to our eyes and hence cannot easily differentiate the slight changes in the color nuances.

7. The elimination of the worst dimensions is much easier, as it is not very challenging to our eyes. Users tend to eliminate the most useless dimensions easily than picking the best as the colors in the worst dimensions are not too concentrated or confusing.

8. Visualization can be used best for dimension reduction than to pick the best.

9. Brute force approach for feature selection is explored by combining all the fake and genuine reviews as one big review each and attempted to select the top differing features. It is evident from our experiments that classification is more complex than mere brute force approaches for feature selection.

10. The fact that brute force approach cannot work is also attributed to the fact that each review is unique in style as they are written by different users and hence combining all of them would not give a holistic analysis of individual reviews.

This visualization uses the D3.js radial chart and the natural color blending when two or more translucent colors overlap. The following areas can be studied to improvise this work:

1. This work uses natural color mixing available in D3.js; it would make more sense to explore better color blending approaches so that can improve the purity and impurity color perception of users and in turn give better results.

2. The random picking off 30 dimensions in one visualization, shuffling and dimension ordering are done manually using the Excel and D3.js. This can be automated for better usability.

3. The feature ranking process is also done manually using MS Excel. An automated approach may be built.

4. The best color combination for the categories can be further studied.

5. The number of permutations of 30 dimensions is beyond the number of groupings taken into consideration in this work. The threshold for the minimum number of permutations can be further studied.

REFERENCES

# REFERENCES


[1] A. Press, "Fake Online Reviews: Here Are Some Tips for Detecting Them," http://www.nbcnews.com/business/consumer/fake-online-reviews-here-are-some-tips-detecting-them-n447681, 2015.

[2] cbcnews, "Online Reviews Faking It," http://www.cbc.ca/marketplace/episodes/2014-2015/online-reviews-faking-it, 2014.

[3] M. Luca, "Reviews, reputation, and revenue: The case of yelp. com," 2016.

[4] N. Hu, P. A. Pavlou, and J. Zhang, "Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication," in *Proceedings of the 7th ACM conference on Electronic commerce*. ACM, 2006, pp. 324–330.

[5] A. Davis and D. Khazanchi, "An empirical study of online word of mouth as a predictor for multi-product category e-commerce sales," *Electronic Markets*, vol. 18, no. 2, pp. 130–141, 2008.

[6] C. Dellarocas, N. Awad, and X. M. Zhang, "Using online reviews as a proxy of word-of-mouth for motion picture revenue forecasting," 2004.

[7] W. B. Wang, M. L. Huang, J. Zhang, and W. Lai, "Detecting criminal relationships through som visual analytics," in *Information Visualisation (iV), 2015 19th International Conference on*. IEEE, 2015, pp. 316–321.

[8] A. Tatu, F. Maaß, I. Färber, E. Bertini, T. Schreck, T. Seidl, and D. Keim, "Subspace search and visualization to make sense of alternative clusterings in high-dimensional data," in *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, 2012, pp. 63–72.

[9] X. Wang, S. Liu, J. Liu, J. Chen, J. Zhu, and B. Guo, "Topicpanorama: A full picture of relevant topics," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 12, pp. 2508–2521, 2016.

[10] Y. Zheng, W. Wu, Y. Chen, H. Qu, and L. M. Ni, "Visual analytics in urban computing: An overview," *IEEE Transactions on Big Data*, vol. 2, no. 3, pp. 276–296, 2016.

[11] S. Liu, J. Yin, X. Wang, W. Cui, K. Cao, and J. Pei, "Online visual analytics of text streams," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 11, pp. 2451–2466, 2016.

[12] M. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D. A. Keim, L.-E. Haug, and M.-C. Hsu, "Visual sentiment analysis on twitter data streams," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, 2011, pp. 277–278.

[13] N. Médoc, M. Stefas, M. Ghoniem, and M. Nadif, "Visual analytics of text streams through multiple dynamic frequency matrices," in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on.* IEEE, 2014, pp. 381–382.

[14] Y. Chen, "Visual opinion analysis of threaded discussions," in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on.* IEEE, 2015, pp. 646–651.

[15] Y.-S. Chen, L.-H. Chen, T. Yamaguchi, and Y. Takama, "Visualization system for analyzing user opinion," in *System Integration (SII), 2015 IEEE/SICE International Symposium on.* IEEE, 2015, pp. 646–649.

[16] J. Seo and B. Shneiderman, "A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections," in *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on.* IEEE, 2004, pp. 65–72.

[17] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in visual data analysis," in *Information Visualization, 2006. IV 2006. Tenth International Conference on.* IEEE, 2006, pp. 9–16.

[18] Y. Song, J. Gong, Z. Zuo, L. Zhang, and D. Wang, "Data integration and visualization: Dealing with massive and multi-dimensional marine spatial data," in *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol. 4. IEEE, 2010, pp. 1620–1623.

[19] H.-Y. Lee, H.-L. Ong, E.-W. Toh, and S.-K. Chan, "A multi-dimensional data visualization tool for knowledge discovery in databases," in *Computer Software and Applications Conference, 1995. COMPSAC 95. Proceedings., Nineteenth Annual International.* IEEE, 1995, pp. 26–31.

[20] M. H. Loorak, C. Perin, N. Kamal, M. Hill, and S. Carpendale, "Timespan: Using visualization to explore temporal multi-dimensional data of stroke patients," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 409–418, 2016.

[21] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.* Association for Computational Linguistics, 2011, pp. 309–319.

[22] S. Banerjee and A. Y. Chua, "Applauses in hotel reviews: Genuine or deceptive?" in *Science and Information Conference (SAI), 2014.* IEEE, 2014, pp. 938–942.

[23] R. Patel and P. Thakkar, "Opinion spam detection using feature selection," in *Computational Intelligence and Communication Networks (CICN), 2014 International Conference on.* IEEE, 2014, pp. 560–564.

[24] Y. Lin, T. Zhu, H. Wu, J. Zhang, X. Wang, and A. Zhou, "Towards online anti-opinion spam: Spotting fake reviews from the review sequence," in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on.* IEEE, 2014, pp. 261–264.

[25] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[26] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[27] G. M. Draper, Y. Livnat, and R. F. Riesenfeld, "A survey of radial methods for information visualization," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 5, pp. 759–776, 2009.