University of Helsinki

master's programme

in data science

Data Science Portfolio

Author:
Norbert Eke

Degree:
Bachelor of Science

December 30, 2017

Contents

1 Introduction

This document will describe my contribution to 3 separate research projects that I have been part of at the University of British Columbia Okanagan. These 3 research projects have been completed during the summers of 2015, 2016 and 2017, while they were not part of my Bachelors degree. The names of the 3 projects are: Sentiment Analysis of Textual Matter, Exploratory Topic Modeling of Textual Data and finally, Feature Based Customer Opinion Mining. My roles in these projects have been Undergraduate Research Assistant for the first research project, then Undergraduate Researcher for the last 2 projects.

2 Sentiment Analysis of Textual Matter

2.1 Project description

During the summer of 2015, I volunteered as a research assistant for a Big Data Sentiment Analysis research project. The project involved performing sentiment analysis on airline customer reviews and retrieving insights from the data, while performing extensive analysis of ratings and reviews provided by airlines passengers to over 200 different airlines. I contributed in the data collection and data cleaning process, then performed keyword and model based sentiment analysis on the data set. More information about the project can be found in an early version of the paper being in progress.

2.2 Area of Data Science and Skills Required

This project belongs to the textual data analysis and natural language processing & understanding, more specifically, the sentiment analysis area of (textual) data science. Nonetheless, with larger dataset sized, this can be also considered big data analytics, as one would analyze hundreds of aspects of customer reviews about over 200 different airlines. Statistical skills, but most importantly, natural language processing and text processing knowledge and skills are required in this area of data science.

2.3 Learning Outcomes

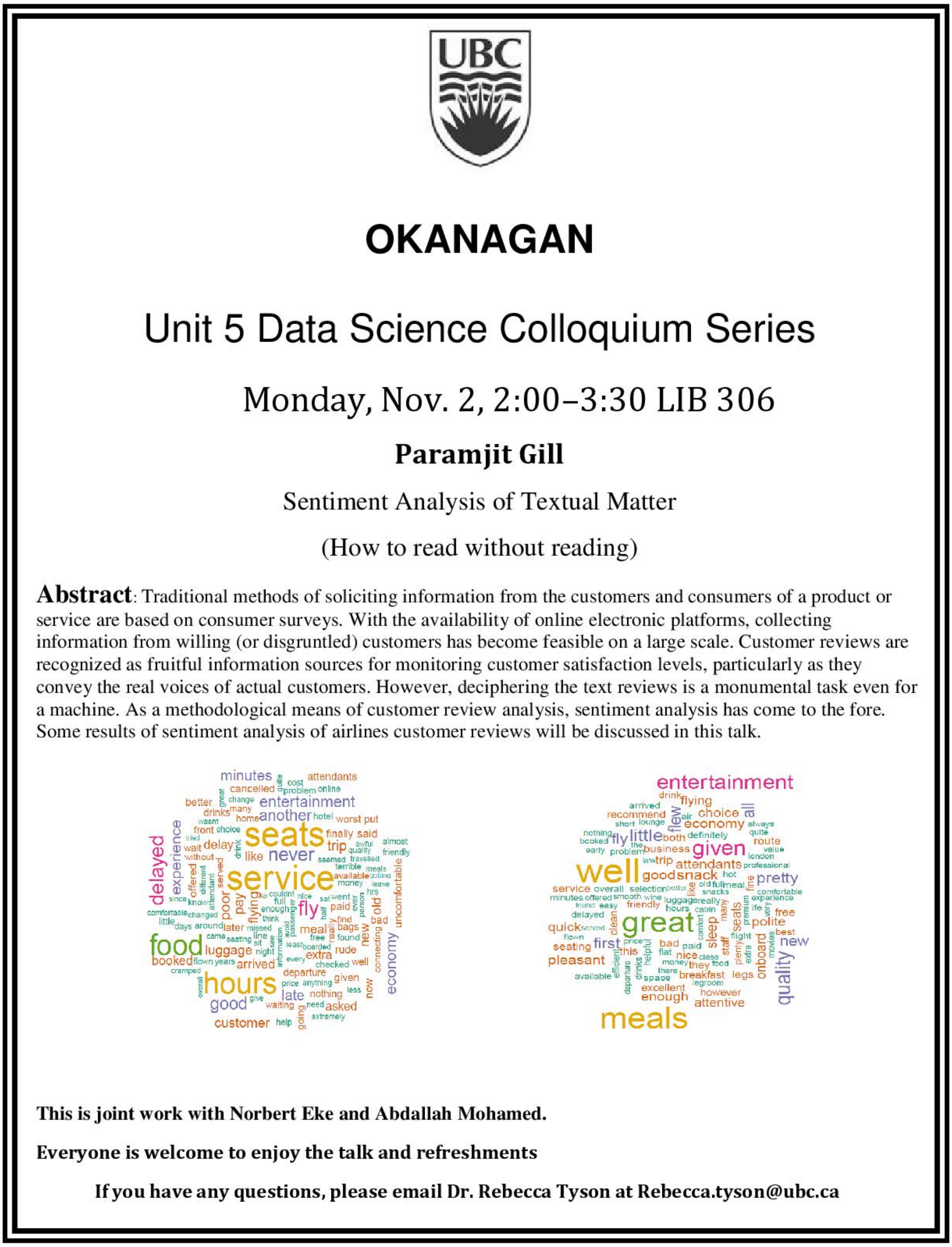
- Learn how to build a web scraper, and specific information off of hundreds of websites
- Learn how to generate structures for the scraped textual data
- Learn how to process, manipulate, and clean up textual data
- Learn how to apply techniques like word and sentence level tokenization, lemmatization, stemming, phrase detection, stop-word removal and parts of speech tagging.
- Learn how to perform keyword based and model based sentiment analysis.
- Familiarize myself with industry level sentiment analysis tools, like Semantria
- Learn the R statistical language and practicing object oriented design principles using Java

2.4 Results & demonstration of newly learned skills

A research paper with the title *A Sentiment Analysis of Customers Reviews of Airlines* is in the process of being published. This paper reports the results of an extensive analysis of ratings and reviews provided by airlines passengers.

Since this project was somewhat successful, at least one seminar talk was given on this topic. Figure 1 shows a seminar talk's poster, which was used to promote the research talk. The slides used for this research talk can be found on GitHub [1].

As for demonstration of newly learned skills, some source code was produced during this project in the following 2 Github Repositories: Web-Scraper and Data Manipulation scripts. An early version of my own write-up can be also found on Github, as proof of contribution to the project.



OKANAGAN

Unit 5 Data Science Colloquium Series

Monday, Nov. 2, 2:00–3:30 LIB 306

Paramjit Gill

Sentiment Analysis of Textual Matter

(How to read without reading)

Abstract: Traditional methods of soliciting information from the customers and consumers of a product or service are based on consumer surveys. With the availability of online electronic platforms, collecting information from willing (or disgruntled) customers has become feasible on a large scale. Customer reviews are recognized as fruitful information sources for monitoring customer satisfaction levels, particularly as they convey the real voices of actual customers. However, deciphering the text reviews is a monumental task even for a machine. As a methodological means of customer review analysis, sentiment analysis has come to the fore. Some results of sentiment analysis of airlines customer reviews will be discussed in this talk.

minutes quite cost attendants
cancelled problem online
better great change entertainment
drinks many home another hotel worst put
wasnt front choice seats finally said
tried wait delay drink trip awful quality almost friendly
delayed experience without like never seemed travelled
offered different attendant served terrible meals
service available money leave uncomfortable
since know poor pay flying couldnt nice sat went hrs
comfortable changed full enough passenger fly ever person paid old
little days around later think missed avoid half find bad
food luggage came seating line sit see information meal really bags new
booked flown years arrived night least boarded free found rude
cramped every extra checked well connecting economy
overall hours departure given
price anything less now
good give late nothing
waiting need asked
customer help going extremely
entertainment
drink flying
arrived recommend flew air choice all
short lounge economy always
nothing booked fly little both definitely quite
early the problem business given route value
late trip london
well attendants professional
good snack hot pretty
service overall selection better like old full meal fine
minutes offered smooth wine snacks comfortable experience
found easy friendly luggage really many premium life very
delayed clean hours cabin seats free
quick served great comfort sleep polite
flown first price helpful flight best
seating this bad paid staff plenty extra movies new
pleasant efficient drinks flat nice class onboard
available departure money they food quality
there space breakfast legs
legroom
excellent however
enough attentive
meals

This is joint work with Norbert Eke and Abdallah Mohamed.

Everyone is welcome to enjoy the talk and refreshments

If you have any questions, please email Dr. Rebecca Tyson at Rebecca.tyson@ubc.ca

Figure 1: Sentiment Analysis Projects Seminar Talk Poster

3 Exploratory Topic Modeling of Textual Data

3.1 Project Description

During the months of May and June of 2016 I initiated a Statistical Machine Learning project, involving research in Topic Modeling and Natural Language Processing. The objective of the project was to explore the applicability of machine learning algorithms in learning and interpreting preprocessed textual data. My contributions included the design of an algorithm to combine topic modeling and deep learning models to extract insight from the data.

3.2 Area of Data Science and Skills Required

This project is a combination of statistical textual analysis, deep learning and natural language processing. The statistical topic models try to identify topics within the data, while the deep learning language models try to 'interpret' the textual data. Some efforts were made to explore the high dimensional document and word embeddings coming out of the Doc2Vec and Word2Ved deep learning language models. Dimensionality reduction using Classical Multidimensional Scaling and Principal Component Analysis and Cluster Analysis using k-means clustering and hierarchical clustering were performed, as exploratory data analysis.

Skills, knowledge and application of statistical topic models like LDA, LSI and HDP is required, while applying deep learning language models like Doc2Vec and Word2Vec is crucial. It is also beneficial to understand and know how to apply various text processing techniques, while having a solid background in natural language processing, to be able to handle subjective textual data. The knowledge and application of various data mining and data analysis tools and techniques is also bonus, since one can unleash the powers of statistics and data analytics by performing exploratory data analysis.

3.3 Learning Outcomes

- Learn how to apply various statistical topic models (LDA, LSI, HDP) to subjective, unstructured, messy textual data

- Learn how to make various deep learning language models interact with topic models (Word2Vec and Doc2Vec interacting with LDA and LSI).
- Learn how to combine natural language processing techniques with deep learning language models and statistical topic models.
- Learn what are the best text preprocessing techniques to be used alongside deep learning language models and statistical topic model
- Learn how to experiment with dimensionality reduction and clustering analysis to extract insight from the data.
- Learn the programming language called Python and practice its use in data science and analytics

3.4 Results & demonstration of newly learned skills

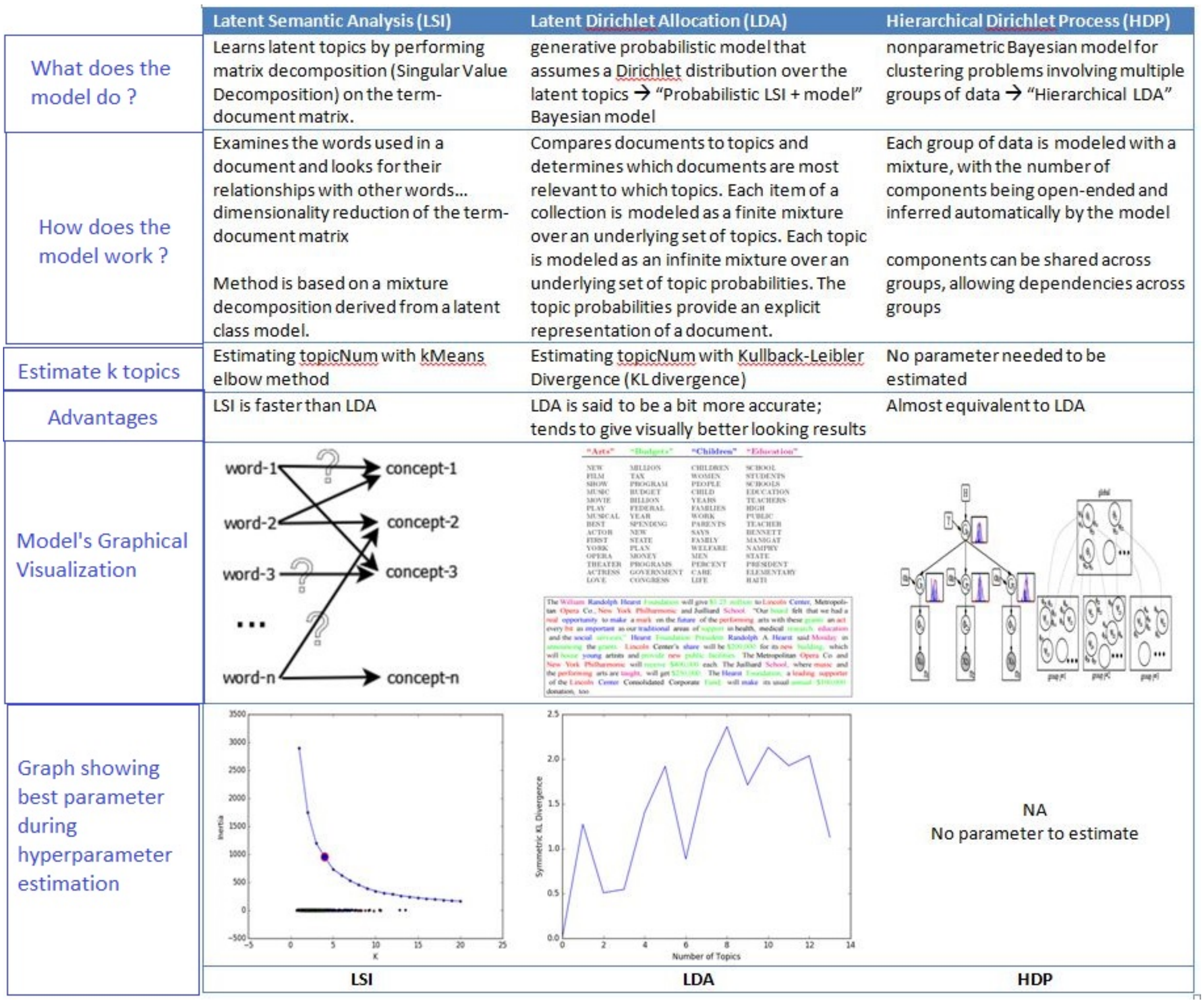
As previously mentioned, this project was an attempt at information retrieval from subjective textual data like customer reviews by combining topic models, deep learning and natural language processing using exploratory text mining.

With the limited amount of time during the summer of 2016, some hidden characteristics and features have been extracted from the textual data using the combination of topic models and deep learning language models, but nothing of major significance was found by the end of the project. Unfortunately, my preliminary results from exploratory text mining were not convincing enough for a potential publication, or even a formal write-up.

My conclusion for this project was that it is definitely possible to extract insight from textual data using deep learning language models and statistical topic models, but it was not clear how exactly they will interact with natural language processing techniques in order to provide insight from data. This project's results only served as playground for my 2017 research project described in section 4, as the project described in section 4 was built upon the lessons learned from this exploratory project.

Some preliminary status reports from the exploratory research work can be found on my GitHub page, with 2 different meeting notes. Most of the unstructured, raw research code can be also found on my GitHub page, on the projects Github repository[5].

Figure 2 shown below represents a quick overview of my understanding of topic models: what are they, how they work, what is each model's advantage, and how to estimate the number of topics within some target documents.



	Latent Semantic Analysis (LSI)	Latent Dirichlet Allocation (LDA)	Hierarchical Dirichlet Process (HDP)
What does the model do ?	Learns latent topics by performing matrix decomposition (Singular Value Decomposition) on the term-document matrix.	generative probabilistic model that assumes a Dirichlet distribution over the latent topics → "Probabilistic LSI + model" Bayesian model	nonparametric Bayesian model for clustering problems involving multiple groups of data → "Hierarchical LDA"
How does the model work ?	Examines the words used in a document and looks for their relationships with other words… dimensionality reduction of the term-document matrix Method is based on a mixture decomposition derived from a latent class model.	Compares documents to topics and determines which documents are most relevant to which topics. Each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is modeled as an infinite mixture over an underlying set of topic probabilities. The topic probabilities provide an explicit representation of a document.	Each group of data is modeled with a mixture, with the number of components being open-ended and inferred automatically by the model components can be shared across groups, allowing dependencies across groups
Estimate k topics	Estimating topicNum with kMeans elbow method	Estimating topicNum with Kullback-Leibler Divergence (KL divergence)	No parameter needed to be estimated
Advantages	LSI is faster than LDA	LDA is said to be a bit more accurate; tends to give visually better looking results	Almost equivalent to LDA
Model's Graphical Visualization			
Graph showing best parameter during hyperparameter estimation			NA No parameter to estimate
	LSI	LDA	HDP

word-1
word-2
word-3
...
word-n
concept-1
concept-2
concept-3
concept-n
"Arts"
"Budgets"
"Children"
"Education"
NEW FILM SHOW MUSIC MOVIE PLAY MUSICAL BEST ACTOR FIRST YORK OPERA THEATER ACTRESS LOVE
MILLION TAX PROGRAM BUDGET BILLION FEDERAL YEAR SPENDING NEW STATE PLAN MONEY PROGRAMS GOVERNMENT CONGRESS
CHILDREN WOMEN PEOPLE CHILD YEARS FAMILIES WORK PARENTS SAYS FAMILY WELFARE MEN PERCENT CARE LIFE
SCHOOL STUDENTS SCHOOLS EDUCATION TEACHERS HIGH PUBLIC TEACHER BENNETT MANIGAT NAMPHY STATE PRESIDENT ELEMENTARY HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 2: Comparison of 3 different Topic Models

4 Feature Based Opinion Mining

4.1 Project Description

In March 2017 I received an Undergraduate Research Award from University of British Columbia Okanagan to work on my own research project. This project involved deep learning using word embeddings applied in feature based opinion mining. The goal was to explore the possibility of designing an automated technique to detect opinion phrases in subjective textual data.

4.2 Area of Data Science and Skills Required

This project is a combination of textual data analysis, deep learning, data mining and natural language processing & understanding. Deep learning language models try to 'interpret' the textual data, then data mining techniques try to find hidden links within the high dimensional data, which can be used to extract features from the text, then use statistical classification models to classify whether a certain word-pair is an an opinion phrase or not.

Skills, knowledge and application of statistical predictive models, deep learning language models and data mining techniques is required. It is also beneficial to understand and know how to apply various text processing techniques, while having a solid background in natural language processing, to be able to handle subjective textual data.

4.3 Learning Outcomes

- Learn how to design a modern approach to the feature based opinion mining problem, using less natural language processing, and more deep learning. Figure 3 shows the designed technique's composition in one diagram.

- Learn how to apply deep learning models to messy, unstructured, subjective customer reviews (various Word2Vec models)

- Learn how to perform dependency parsing on messy, unstructured, subjective textual data (using Stanford Dependency Parser and SpaCy (industry-level) Dependency parser)

- Learn how to train, evaluate and validate the performance of various types of statistical classification models (SVM, LASSO, LDA, QDA, Random Forest, Bagging, Gradient Boosting) on Feature based Opinion Mining benchmark datasets
- Learn how to investigation of the relationship between feature words and descriptor words, which resulted with the discovery of feature-descriptor relation vectors
- Learn how to perform feature extraction and apply various data mining techniques to find hidden features within the high dimensional word embedding data
- Learn how to perform binary sentiment classification on opinion phrase polarity, to decide whether an opinion was used in a positive or negative context
- Learn how to approach the problem of feature based opinion mining in a new, innovative, modern and unique way
- Gain more practical knowledge of statistical packages in R and machine learning libraries in Python

4.4 Results & demonstration of newly learned skills

On June 17th, 2017 I gave a conference talk at the Canadian Undergraduate Computer Science Conference (CUCSC 2017), organized at University of Toronto. My talk included ongoing research work on designing *A Modern Approach to Feature Based Customer Opinion Mining*, presenting partial results and hopes of improvements on the technique. Slides from this conference talk can be found on my Github page [7]. Video recording of the talk can be found on my LinkedIn page [2].

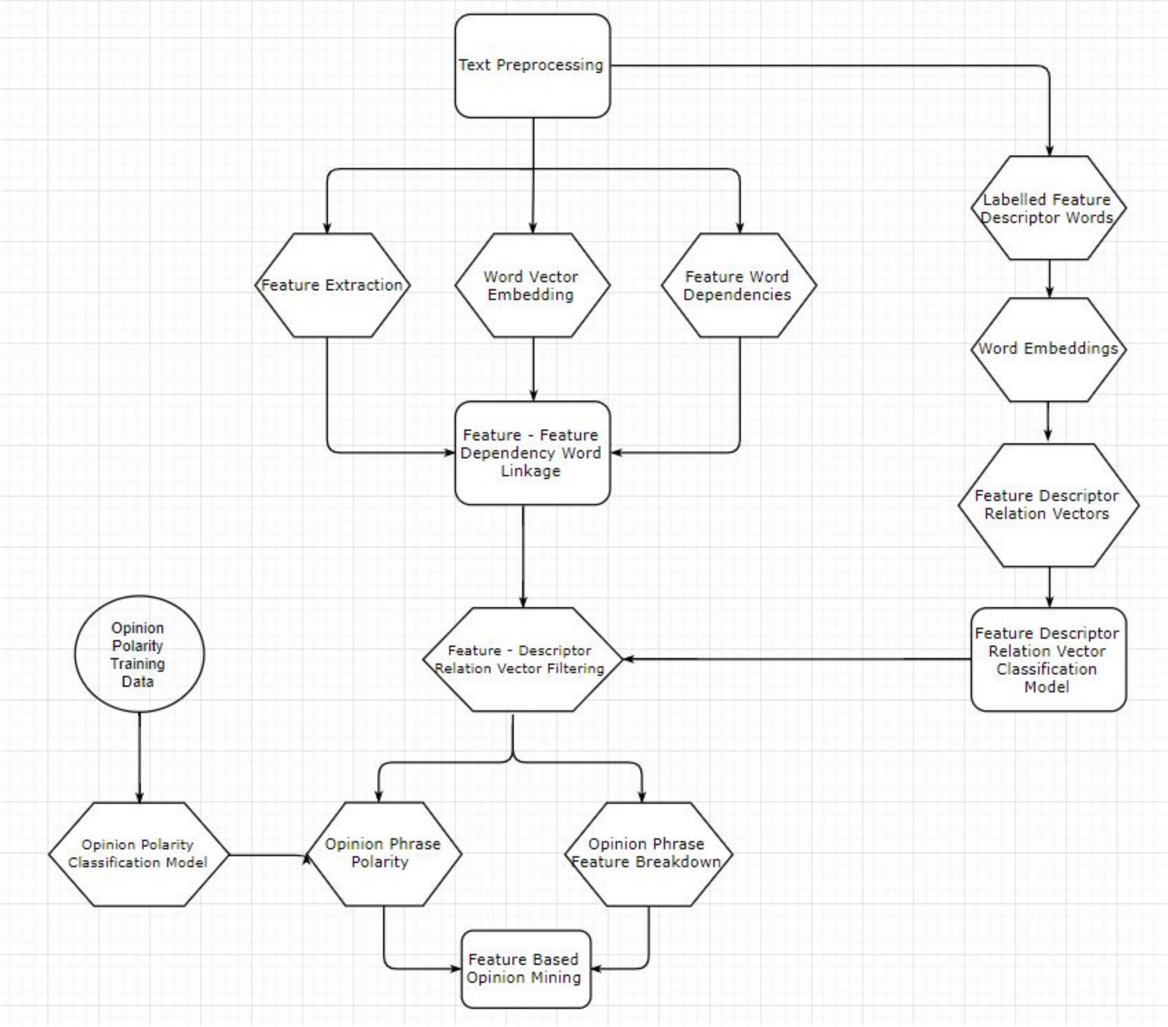
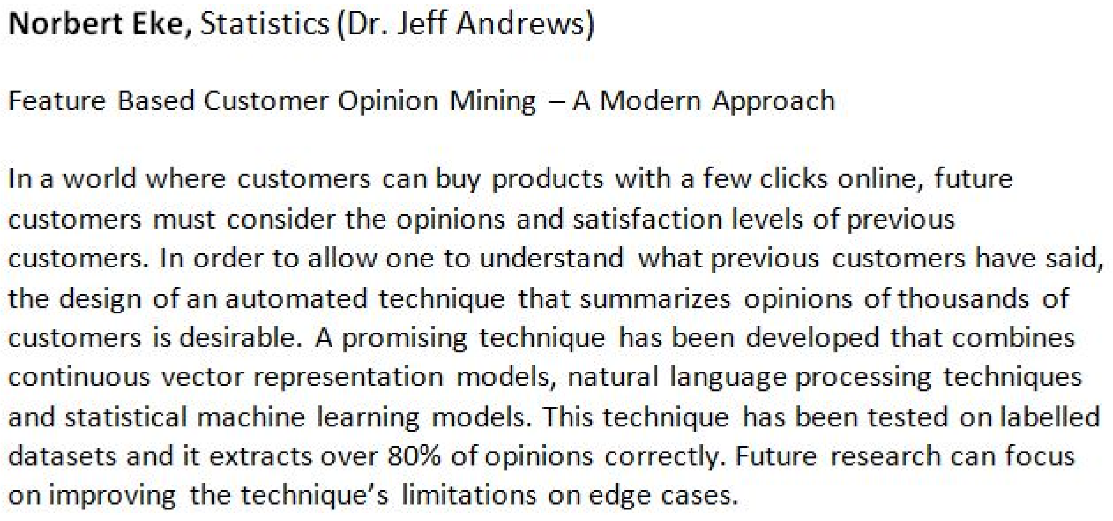

Text Preprocessing
Labelled Feature Descriptor Words
Feature Extraction
Word Vector Embedding
Feature Word Dependencies
Word Embeddings
Feature - Feature Dependency Word Linkage
Feature Descriptor Relation Vectors
Opinion Polarity Training Data
Feature - Descriptor Relation Vector Filtering
Feature Descriptor Relation Vector Classification Model
Opinion Polarity Classification Model
Opinion Phrase Polarity
Opinion Phrase Feature Breakdown
Feature Based Opinion Mining

Figure 3: Diagram of the whole system of algorithms interacting with each other to create a modern Feature based Opinion Mining system

By September 2017 the project reached its end, and the technique was successfully designed, with promising results showing 80 to 85% accuracy on successfully (using an automated way) identifying opinion phrases from labeled benchmark data. Unfortunately the project's source code at the moment is just messy, not-perfectly-structured research code, but the good news is that it is open source, and it can be found in one of my GitHub repositories [6]. Detailed documentation for the source code is still on my (never ending) *To Do list*.

On September 20th, 2017 I gave a research talk at the Undergraduate Research Symposium, at University of British Columbia Okanagan. Slides from the research symposium talk can be found on my Github page [8]. Video recording of the talk can be found on my LinkedIn page [3]. A screen-shot of my abstract from the symposium, at University of British Columbia Okanagan can be seen in figure 4.



Norbert Eke, Statistics (Dr. Jeff Andrews)

Feature Based Customer Opinion Mining – A Modern Approach

In a world where customers can buy products with a few clicks online, future customers must consider the opinions and satisfaction levels of previous customers. In order to allow one to understand what previous customers have said, the design of an automated technique that summarizes opinions of thousands of customers is desirable. A promising technique has been developed that combines continuous vector representation models, natural language processing techniques and statistical machine learning models. This technique has been tested on labelled datasets and it extracts over 80% of opinions correctly. Future research can focus on improving the technique's limitations on edge cases.

Figure 4: Abstract submitted to Undergraduate Research Symposium at UBCO

A brief summary of my research project, and my findings can be found on my GitHub page [9].

The culmination of my undergraduate research project was the writing of my first research paper. At the time of this application, the paper is in the final review stages, being overlooked by my summer research supervisor, Jeffrey Andrews, an Assistant Professor in the Statistics department at University of British Columbia Okanagan. After one more extensive editing round, hopefully I could get the paper published. The current and up to date version of my research paper can be viewed on my GitHub page [4].

5 Conclusion

As a final point, I have an interesting curiosity towards what sorts of insight lie behind data. I enjoy performing various analyses, trying out different models or techniques, and eventually retrieving insight from the data. I call it passion for data science and analytics. As my research suggests, I am very passionate about applied machine learning, natural language understanding, data analytics, and most importantly data mining. I truly believe that a master's program in Data Science will help me learn more about my passions and it will catapult me towards my career goals of analyzing as much interesting data as possible.

References

[1] GitHub repo of the Sentiment Analysis Project Documents (2015)

[2] LinkedIn Article on Canadian Undergr. Computer Science Conference (2017)

[3] LinkedIn Article about Undergraduate Research Symposium talk (2017)

[4] Feature Based Opinion Mining Research Paper (2017)

[5] GitHub repo of Topic Modeling of Textual Data Exploratory Research (2016)

[6] GitHub repo of Feature Based Opinion Mining Research Project (2017)

[7] Canadian Undergraduate Computer Science Conference Talk Slides (2017)

[8] Undergraduate Research Symposium Talk Slides (2017)

[9] Undergraduate Research Project Summary (2017)