Chapter 7

Summary and Future Research

Summary of Research

The importance of data in every avenue that mankind walks on has become so evident that it needs no further introduction. Everything around us is gradually getting driven by data – be it the services we consume, the products we buy and even the government we choose. The vast potential and value which lies latent in data is forcing organizations and individuals across all domains to acknowledge and adapt approaches and techniques that can unlock that potential to the fullest. What initially started out with traditional statistical techniques being applied on extremely structured data, has now moved on to sophisticated machine learning, deep learning, Artificial Intelligence (AI) approaches which primarily focus on unstructured data.

The rate at which data is generated today is beyond imagination. Business processes, social media, e-commerce, wearables, gadgets are only some of the sources from which data spawns. In addition to the huge volume and speed, the nature of the data brings forth its own set of problems. More than 3/4th of the share of data generated has no structure and all the existing traditional approaches fail when applied on them. Hence, there is a need for the techniques and approaches which help extract value from unstructured data which we have discussed in this thesis.

Chapter 1 covers all the issues and setbacks that come with big data and unstructured data and are discussed and stated in the introductory chapter. It contains descriptions of different sources and kinds of unstructured data such as text, audio, video etc. A high-level discussion is provided on the techniques and approaches that can be employed, with specific focus on textual data.

In Chapter 2, the concept of document models is discussed with respect to the Bernoulli approach, that is, basis is the presence or absence of primary blocks of the documents, namely tokens. The chapter primarily deals with how an unstructured dataset consisting of text documents is converted to structured content with mathematical and statistical foundation. Examples and sample code snippets in R and Python to execute the same have been included for Bernoulli document model.
Chapter 3 is basically the continuation of Chapter 2. The focus in this chapter is on Multinomial document model. It is similar to the Bernoulli model, but the presence flag in the former is now replaced with the frequentist method which takes into account the number of times the tokens occur in the text. In this chapter, the idea of Naïve Bayes approach is discussed and how it is applied coupled with the document models. In addition to these, estimation of parameters in a Naïve Bayes and Multinomial setup using Maximum Likelihood Estimation (MLE) technique has been discussed. Similar to the Bernoulli approach, this chapter too contains an illustrative example and the code snippets for implementation. A brief comparison of two document models has also been provided.

In Chapter 4, from the supervised document model approaches discussed in Chapters 2 and 3, we move onto discussion of unsupervised topic modeling techniques to obtain latent topical structure across text documents and further fine-tuning with help of machine learning. First, the inception and need of having an approach that captures semantic relationships have been discussed, inspired by the work of Deerwester (1990) on the topic of Latent Semantic Indexing (LSI). The underlying mathematical concepts governing LSI are explicitly discussed in detail. Additionally, illustrative examples and code snippets are provided. The first part of this chapter throws lights on the pros and cons of LSI approach and this paves the way into the latter half of the chapter focussing on fine tuning of the textual topic. It also contains detailed explanation of the concepts of subnets and hypernyms and how these can be employed to bring forth added benefits on existing topic modeling approaches. The example demonstrates how construction of subnets in the dataset adds more variability explanation on top of the topics provided by existing approaches. The example emphasizes how subnets and hypernyms ultimately help in achieving a fine-tuned topical representation.

Chapter 5 deals with obtaining of key conversational drivers for textual data coupled with the attached sentiment and mood states. Such an approach helps in detecting the key drivers of conversation that is, whether it is having a positive impact or not or whether the text content has potential to become viral and much more. One very useful application of the same involves tracking the positive and negative drivers of reviews and ratings posted by users on e-commerce platforms, and as a result, adopt strategies for the products well in advance. A detailed discussion has been added on the process flow, with complete explanation of each of the working components.
There are added code snippets too using R and Python which illustrate the manner in which the implementation can be done. The topic of Latent Dirichlet Allocation (LDA) inspired by the work of Blei, Ng and Jordan (2003) has also been discussed. Detailed discussions on the mathematical basis, the notations and terminologies, comparison with other generative models are provided for LDA. An end-to-end implementation using R has been discussed with code snippets and sample data.

In Chapter 6, we focus on Mood State and Behavior Prediction model in Social Media through Unstructured Data Analysis. Behavior Dirichlet Probability Model (BDPM), which can capture the Behavior and Mood of user on Social Media is proposed using Dirichlet distribution. Additionally, real-life illustration and R code snippets are provided.

**Future Research**

Over the last few years, our primary focus was on the domain of text analytics. We have attempted to extract maximum value from unstructured text data and leverage it to the stakeholder’s benefit. Going by the present trends in the industry and the huge influx of different kinds of unstructured data in audio, image and video format, we are next planning to focus on audio data.

The idea is to capture utility of audio data in the quick service (QS) industry scenario. With more and more brands looking to revamp their image and turning their attention to smart customer services, our objective is to integrate a voice based feedback system for these brands. Present trends in digital applications indicate that customers prefer spending lesser time on the apps but want to achieve more benefits and value. Instead of having a text-based feedback platform, a voice-based analytics platform driven under the setup of mathematics and statistics foundations would reduce the consumer’s effort and ease the mechanism of providing feedback and reviews for the services received. Our research aims to utilize this audio input from the consumers and ultimately drive conversational intelligence models on it. Since we are already doing extensive research with text data, in our future endeavor, we shall be focusing on effectively converting the data in audio format to text format through statistical driven algorithms.

The second task which we are planning to address in the QSR (Quick Service Restaurants) space is automated Point of Sales (POS) systems based on audio inputs. This can be one more extension of the smart services provided at a physical level interaction with consumers. The objective is to
make a real-time intelligent POS system which accepts speech-based ordering from consumers
and at the same time, invokes the historical transactions for the user and make smart
recommendations based on the user’s purchase behavior.