CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter will review historic methods and challenges involved in analyzing social media data. Secondly, it examines how new technologies, and here we are talking particularly about the use of deep learning, could advance the quality and speed of analytical processes. The key area observed in this chapter is the (LSTM) memory network, which has become one of the most broadly employed strategies for investigating several kinds of social media content. Lastly, the 45th section talks about the present situation and future framework that has to do with LSTM applications.

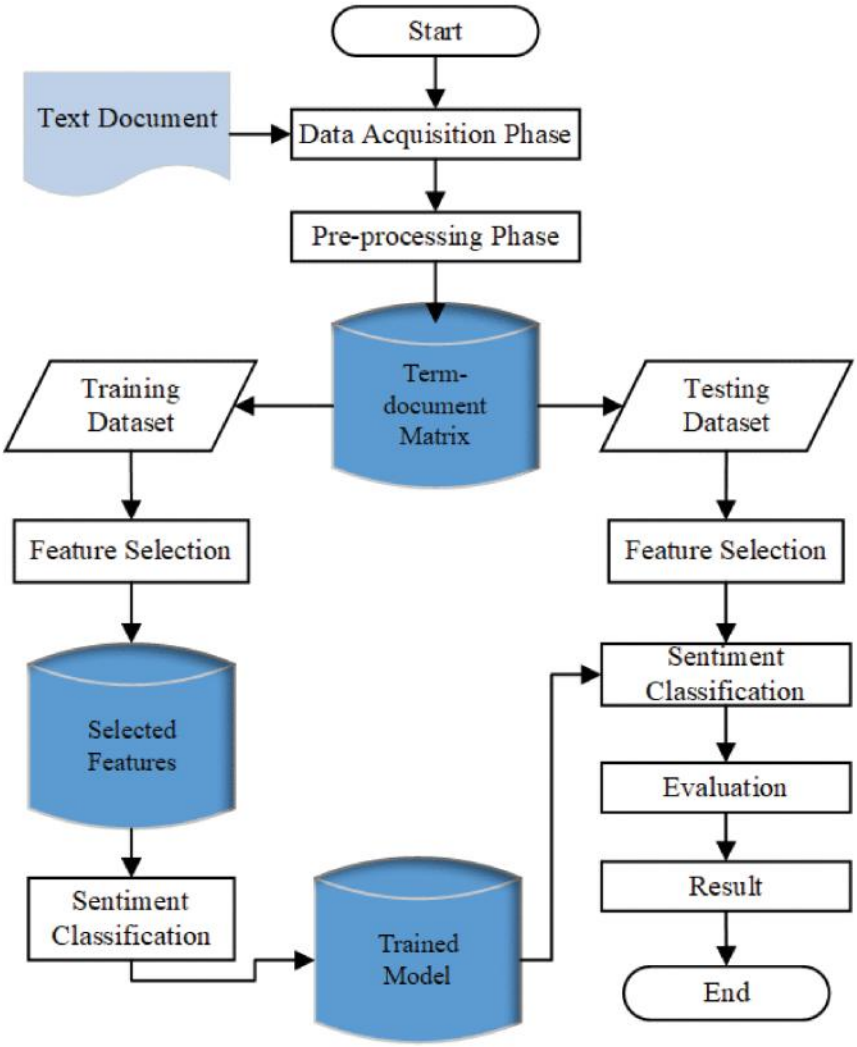
2.2 Traditional Approaches to Text and Topic Analysis

As the popularity of monitoring the contents of social media data has increased drastically in recent years, the data sources have made a leap from solely text-oriented details to a myriad of data types including text, image, media, emojis, sound, hashtag, and mixed data. In the first stage, researchers designed the perfect means to apply social network data and, consequently, extract valuable knowledge from them into a structured format. The main difficulty lied in the huge and changing scale of data sources. Some of the techniques and models developed by the researchers to solve these issues are as follows.

The most popular approaches include the Bag-of-Words (BoW) model, Latent Dirichlet Allocation (LDA), and Term Frequency-Inverse Document

Frequency (TF-IDF), which takes on more sophisticated natural language processing (NLP) algorithms.Firstly, a model is created for the BoW analysis of the text through introducing the technique called 'Bag-of-Words'. The technique is considered one of the simplest and the most frequently applied ones because other text analysis methods were later derived from it. BoW is an approach to text representation that views text as a group of just individual words, ignoring textual structure and how frequently a word occurs in the original content. In this manner, it can glean semantic and syntactic relations among the most pertinent terms within a document. However, the model has its own imperfections and faces such challenges when it comes to dealing with actual social media data.

Here we are not discussing a social media content composed of separate sentences, but which in turn is made up of both useful and senseless responses from different users/persons. In the same manner, transforming paragraphs into bags of words is also an insurmountable task, and in many situations, even impossible. To adjust such novel issues as a result of real life challenges, an innovative technique was developed by the researchers, which is represented by the novel text analysis: Term Frequency-Inverse Document Frequency (TF-IDF). A basic method known as TF-IDF is undertaken where the frequency of terms in a document is calculated that also takes into account of their relative importance in the whole corpus.



To conclude, there is a difference between BOG and TF-IDF models. In the BOG model, the frequency of meaningful words is not considered, whereas in TF-IDF model, the frequency of meaningful words is considered as well as the way they are distributed throughout the corpus, which makes the representation of the text more precise. Such an increase overcomes the restrictions which the bag-of-words model represents, hence the analyst can do more than just the document by document correlation of the text, bring up a more nuanced and better-internalized understanding of the issues under consideration.

2.3 Topic-Based Analysis Using Deep Learning

The analysis of TF-IDF keywords is a powerful tool for the extraction of dominant words in the text based on social media. It helps not only for tracking the most popular topics but also the significant keywords. Therefore, it is widely used in social networks monitoring systems. Similarly to BoW, the other approach, TF-IDF has its own limitations. Indeed, the BoW model is merely superficial, and it involves no deeper insights in context, so at times, it does not manage to appreciate data which could be very important. In spite of the fact that TF-IDF can be more effective than BoW in other terms, on the contrary, it shows lower accuracy, which is an obstacle for forecasting similar trends.

Social media have raised the necessity for studying text analysis, and this is a need that has been recognized by researchers. However, they do realize that currently text analysis will not be sufficient to keep up with rapidly changing social media that is developing both in complexity and structure. These developments in techniques, key of which are the deep learning, have been quickly employed in up-to-date identification platforms. Meanwhile, social media metadata is becoming more and more intricate, providing a pattern of not only texts but also images, hashtags, emojis, and other multimedia, and so this one can be regarded as a classed data. This transition means that standard methods of data analysis are now obsolete, and we have to come up with unique ways to process such data. It was pointed out at the previous part of this chapter that a more standard model, such as the Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), is unsuited for a variety of data types. Despite this, original texts are still analyzing, but the process is time-consuming, leading to resource use.

The constraints of traditional methods remain a problem, but new analytics methods like BoW (Bag-of-Words) and TF-IDF (Term Frequency-Inverse Document Frequency) contribute much to research and represent the primary tools of basic research. Current approaches, particularly the emergence of deep learning methods, revolve around the issue of handling noise and filtering out the error and detecting UN-informal content from social media. Their goal is to help guide humans to

consume small passages rather than big chunks of meaningless text. Therefore, using structured and formal data will be of better quality, which are prone to the principles of quantitative analysis in the research platform. This system prevents the loss of meaning to a certain level and makes the allocation of related topics much easier.

The information extraction models of deep learning perform better on understanding deep subjects and estranged cases than plain methods of learning. Nowadays, the most critical issue is the design of tools to resolve intricate and untraveled tasks in the area of social media metadata. The two techniques that have turned out as the most vital approaches in the area of Deep Learning for this task are Recurrent Neural Networks (RNNs) and Long-Short Term Memory (LSTM) networks. Science, as it was introduced in this chapter, REVEALING of meaningful relationships within diverse contexts remains a significant challenge for any analytical method. Memory storage, computing function, and the ability to write analyses on the fly all are the essential elements of this system. Improved performing skills in terms of memory, processing, and editing of data are several of the pros of the Graphics Processing Units (GPUs) that have been developed.

With these Optimizations now Graphics Processors (using Deep Learning techniques as their core technology) became highly efficient and versatile platforms. Deep learning functions deal with words partnerships, full sentences, complete documents, bonds that reach to deep levels in the data. The most important trainable feature of deep learning is the ability the irrelevant and non-meaningful data that are made usually in this type of media. Looking to the future, the deep learning strategies will be able to learn complex hierarchical features from theses raw files, Bringing sequence A more precise requirement for enterprises employing online social media.

This leads to the inference that the better the functioning of the system, the more accurate its results will be. This is a remarkable extension of routine data processing methods. For the topic-based analysis, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) work better in most cases. The primary role of CNNs is the extraction for the purpose of features out of text. These

systems have been used in this area and have widely applied. They stand out in the sense of identifying factors distinctive to a certain topic. In contrast, RNNs, especially LSTMs, are mainly used to discern meaning out of running text as they focus on capturing sequential connections useful for maintaining context in social media postings. The method used shall be designed so that researchers can look into different types of information in order to study them.

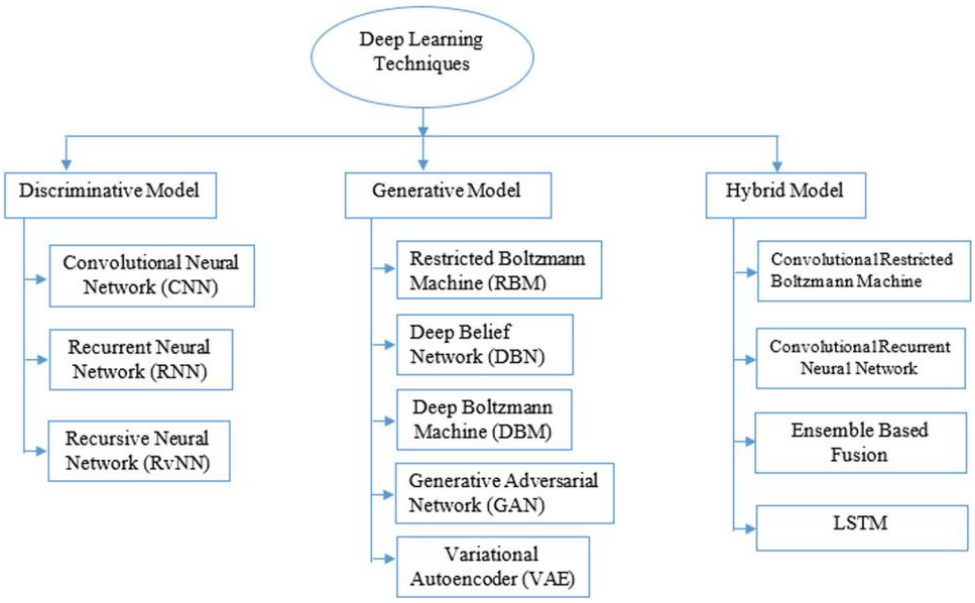
2.4 Introduction to Long Short-Term Memory Networks (LSTMs)

There is a variety of methods of categorization that permit choosing the appropriate one depending on the specifics of the data. Here's the section that's going to deliver a presentation of Long Short-Term Memory (LSTM) networks that are being among the most recent applications for topic-based social media recognition systems.

LSTMs present an advantage in terms that they are prudent and all important tools in creating very dependable modulation frameworks. In 1997, “The analysis of Long Short-Term Memory (LSTM) networks” came to be. Despite the obstacles which the hardware technology presented at that time, LSTMs were rather regarded as a theoretical concept but in a very little way as LSTMs commenced to advance quickly and were being applied in the field to effect real-world application. The period extended in years 2015 to 2016, and I can witness that AI or artificial intelligence found its way to the fore most and paved the way to its wide acceptance when, advantages in hardware, particularly Graphics Processing units (GPUs), enabled the bursting into the scene of AI, specifically deep learning.

Thus, in the course of developing AI, deep learning eventually achieved the computational ability necessary for it to be used with complex systems. This approach is now applied in real-world areas. Concurrent to this, Google made its own unique Artificial Intelligence (AI) platform which aimed at handling challenges like

the parallel computing in terms of software. Google researchers showed significant improvements through their work on deep neural networks (DNNs), which were crucial for progress in improving LSTM networks. This undertaking was a joint effort with Jeff Dean and, as a result, many goals were achieved. Combining graphics technology processing units - GPUs with deep learning technology, Google sped up procedure related to these tasks. This result is encouraging as the new method is cosmodrome very far behind conventional systems in their speed. These technique improvements enabled Google to derive effective algorithms that could learn highly sophisticated representations of input data sets and thus it made significant contributions to the development of Artificial Intelligence applications.



The work that has been carried out to develop on these prior achieved levels makes LSTMs much more feasible to be implemented to deal with some challenging issues that have been specified for the use case. When you come to the LSTM basis, the architecture is divided into four major parts–the Cell State, Forget Gate, Input Gate, and Output Gate. These constituents provide deep learning systems with the basic functionality necessary to meet their objectives.

The Cell State is vital in the information capture and preservation of the input data sequences long-run association. It acts as a memory module for the network, which keeps the knowledge over prolonged time by storing and analyzing the relationships between interconnections of a vast number of contexts. Apart from that, it ensures steady data flow through the system and eliminates the inaccurate. Such an ability allows the unit to handle complicated patterns, not only those shown above but also incoming streams from natural language, for instance.

Yet social media does not consist solely of natural language, but a large proportion of its volume is still based on text that may be translated into a natural language form. The crucial impact of Cell State is to make the model more descriptive, in only a single word to clarify more and have higher precision. Components of the doors except of the gates are relatively easier to interpret. Three gates: Forget, Input, and Output provide a decision at the current time step based on the hidden state at the previous time step as well as the target. The rectifying connections, with different activation functions, calculate the coefficients for the Forget, input, and output gates in this structure. At the input gate, the system makes a diagnosis of the quantity of the information that must be stock to the main memory block. What's more, the Forget gate is the initial step for the system to decide if the value in the memory cell should be saved, or imposed. The Output door identifies what adjustment the memory cell takes in relation to the real advance of time. LSTMs are widely deployed for audit social networking posts to determine the mood. Such posts aren’t usually long and, hence, could be informal.

2.5 Challenges in Topic-Based Social Media Analysis

LSTM networks, through surfacing relations amidst the seen and unseen cohesions, can classify a post as positive, negative, or neutral, which might serve as an indicator to malicious content. On the other hand, LSTMs can also be employed to deter the dissemination of fake information using novel patterns. LSTM models prove to be especially useful when they visualize and eventually compare the usual

conversation flow with the newly issued posts, in order to discover the posts that break away from the most likely context.

The latter case may indicate that the previous post contains incorrect information. Because of the provided textual data, there are many font sizes and shapes to meet public needs, exist relevant news, and analyze customer behavior. Nevertheless, uncovering these insights presents a unique challenge which is both due to the nature of the social media language and the whole setting of it. Apart from the fact that online writing is conversational and rapidly changing, traditional form of writing is forward, written with maintaining decorum, neat, and never undisciplined.

This is done largely by studying social media data, particularly the text and all the issues raised by the textual aspect of data such as informal conversations and changing language, which is still maturing. Here is a look at how deep learning methods, especially those that use Long Short-term memory (LSTM) networks, can handle perplexities of the subject. Despite its upsides, LSTMs raise some concerns. The most critical of these is that of computational complexity. Although LSTMs can be a useful tool in the analyzing of real-world social media data, they require a great deal of computational resources to perform accurately must be contained in this section.

In turn, on the other hand, the challenge point becomes the model's complexity, making it difficult to see how the model decided upon one specification. This lack of interpretability makes the machine learning process mysterious and thus is highly undesirable, mainly for applications as sensitive as misinformation detection. The future prospects are to concentrate on the way in which the calibration process may be enhanced in integrity, cut down the computations, and explain the functioning of the model through attending to attention mechanisms, as well as explainable AI methods like XAI. Collaboration of LSTMs with the other

architectures, as transformers, may be the most dominant process of building successful detection of manipulated information on social networking platforms.

2.6 Summary

Text and topic analysis in social media has shifted from traditional methods to modern deep learning techniques. Traditional methods, such as Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and Latent Dirichlet Allocation (LDA), have played a fundamental role in text data processing. However, these methods show great limitations in handling the informal and dynamic nature of social media content, including the prevalence of abbreviations, slang, and hashtags. In addition, they have difficulty in effectively capturing semantic relationships or contextual meaning.

To overcome these challenges, this review highlights the growing prominence of deep learning methods, especially LSTMs, which are well suited for tasks such as topic detection and sentiment analysis due to their ability to manage long-range dependencies. Of course, these methods face challenges with high computational requirements.