Topic-Based Analysis of Social Media Posts Using RNN and LSTM

ZHU QIAN

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews historical methodologies and challenges associated with analyzing social media data. It then explores the role of new technologies, particularly deep learning techniques, in improving analytical capabilities and efficiency. In particular, this chapter focuses on Long Short-Term Memory (LSTM) networks, which have become one of the most widely used methods for analyzing different types of social media content. Finally, this chapter focuses on current issues and future directions related to the application of LSTM.

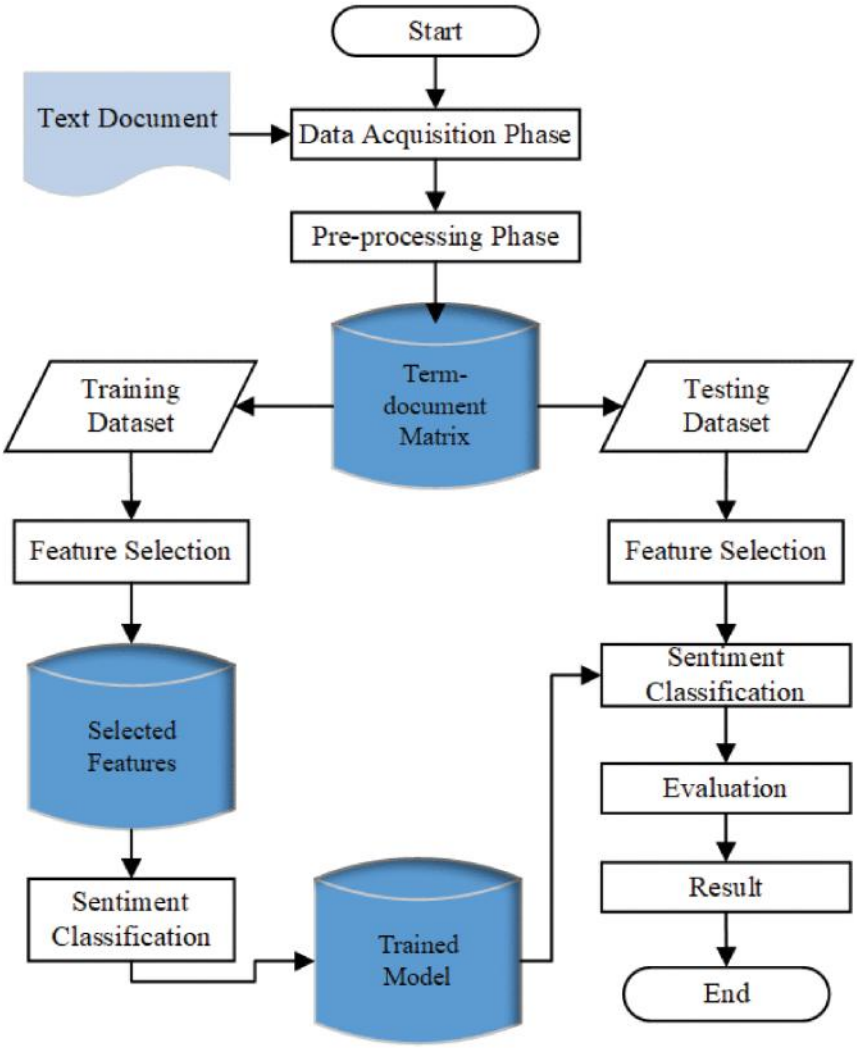
2.2 Traditional Approaches to Text and Topic Analysis

With the rapid growth of social media platforms, the content of social media data has evolved from purely textual information to a wide variety of content types, including text, images, media, emojis, sound, hastag, and mixed data. In the early stages, researchers developed powerful methods to extract meaningful insights from the original social media data and translate them into structured datasets. The primary challenge was the vast and dynamic nature of these data sources. To address these challenges, researchers developed several techniques and models. The most popular models include the Bag-of-Words (BoW) model, Latent Dirichlet Allocation (LDA), and Term Frequency-Inverse Document Frequency (TF-IDF), all of which are based on advanced natural language processing (NLP) algorithm.

To analyze text content from social media, researchers developed an efficient method known as the Bag-of-Words (BoW) model. This is one of the simplest and most widely used techniques for text analysis, and many subsequent analysis methods are based on it. The Bag-of-Words (BoW) model represents text as a collection of individual words, ignoring grammatical structures and the frequency of words in the original content. As a result, it extracts semantic and syntactic relationships from the most meaningful and relevant words. However, this model has limitations when applied to real-world social media data. Social media content is not limited to individual sentences but often consists of a combination of relevant and irrelevant replies from various users. Therefore, using Bag-of-

Words (BoW) to capture contextual information is often challenging, and in many cases, nearly impossible.

To address the new challenges arising from real-world use cases, researchers developed a new technique for text analysis: Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a foundational method that measures the frequency of terms within a document while accounting for their relative importance across the entire corpus. The key difference between TF-IDF and the Bag-of-Words (BoW) model is that TF-IDF tracks the frequency of meaningful words and incorporates their distribution within the corpus, providing a more nuanced representation of the text. This enhancement overcomes some of the limitations of the BoW model, enabling researchers to analyze not only individual text segments but also the relationships between multiple documents, thereby extracting more meaningful and contextually relevant information. TF-IDF is particularly effective in identifying key terms within social media content, facilitating the tracking of trending topics and relevant keywords. As a result, it is widely applied in social media analysis platforms.



However, similar to the BoW model, TF-IDF has limitations when applied to real-world business challenges. It struggles to fully capture the context of the content and often misses critical information that could be highly significant. While TF-IDF is more powerful than BoW in certain respects, it tends to be less accurate, which can negatively impact the prediction of trends.

2.3 Topic-Based Analysis Using Deep Learning

Researchers have recognized that text analysis alone is insufficient to meet the evolving demands of social media platforms, which are developing at a rapid pace and generating more complex forms of content. Thanks to advancements in techniques, particularly deep learning, new methods have been swiftly applied to contemporary analysis systems. As social media metadata becomes increasingly complex—encompassing not only text but also images, emojis, hashtags, and multimedia—this data can be classified as unstructured. This shift implies that traditional analysis techniques are no longer effective for such data. As discussed in earlier sections of this chapter, models like the Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) struggle to extract useful information from these diverse data types. While they can still process original text, they are often inefficient, requiring substantial time and resources to do so.

Despite the limitations of traditional techniques, the Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) models still provide valuable insights and serve as foundational tools for research. New techniques, particularly deep learning methods, focus on addressing the challenges of filtering noise and identifying informal content from social media. These methods aim to transform less meaningful text into high-quality, actionable information. Structured and formal data are of higher quality and are better suited for quantitative analysis in research platforms. This approach reduces the loss of meaning and enhances the classification of relevant topics. For understanding complex subjects and nuanced themes, deep learning models outperform traditional methods.

As technology matures, addressing the complex and unstructured nature of modern social media metadata has become increasingly important. Among the most widely used deep learning techniques for this task are Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. As discussed earlier in this chapter, identifying meaningful relationships within diverse contexts remains a significant challenge for any analytical method. The system must have sufficient memory capacity to store temporary data, process input data in real-time, and produce analysis results with continuous updates. Thanks to advancements in Graphics Processing Units (GPUs), hardware limitations are no

longer a major constraint. Researchers can now focus on optimizing GPU usage to create efficient and powerful platforms based on deep learning techniques.

Deep learning methods can handle more complex sentences and entire documents, allowing the model to identify meaningful relationships within the data. One of the key strengths of deep learning is its ability to effectively reduce the impact of informal and fragmented data commonly found in modern social media. Looking to the future, deep learning techniques have the potential to automatically learn hierarchical features from raw data. This implies that the more data the system processes, the more accurate the output results will become. This represents a significant revolution in the way topic data is analyzed.

There are distinct advantages to the two most popular approaches for topic-based analysis: Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. CNNs are primarily used for feature extraction from text and have been widely applied in this area. They are particularly effective in identifying indicators of specific topics. In contrast, Recurrent Neural Networks (RNNs), particularly LSTM networks, are more commonly employed to capture sequential dependencies within text, which is crucial for understanding context in social media posts. Researchers can use these methods to analyze various types of content. Depending on the specific requirements of the task, the appropriate method can be chosen based on the nature of the data.

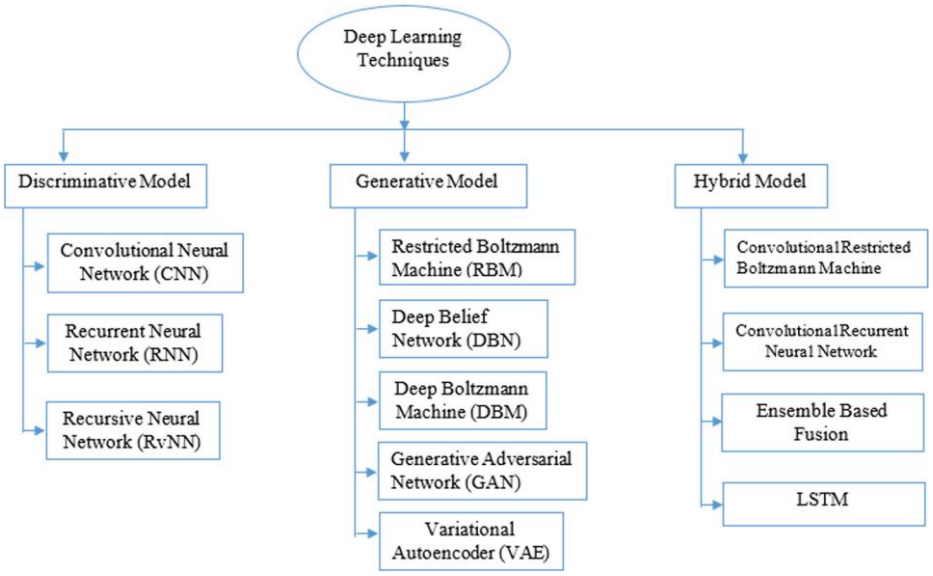
2.4 Introduction to Long Short-Term Memory Networks (LSTMs)

In this section, we will focus on introducing Long Short-Term Memory (LSTM) networks, which have become among the most widely used methods for topic-based analysis of social media in recent years. LSTMs are efficient and powerful tools for developing robust analytical frameworks.

The study of Long Short-Term Memory (LSTM) networks began in 1997. However, due to the limitations of hardware technologies at that time, LSTMs were primarily a theoretical concept with little opportunity for real-world application. Their widespread popularity emerged after 2015, when advancements in hardware, particularly Graphics Processing Units (GPUs), made artificial intelligence increasingly feasible. As a part of AI, deep learning finally gained the computational power necessary to tackle complex projects. This development enabled Long Short-Term Memory (LSTM) networks to be effectively applied in real-world scenarios.

During the same period, Google developed a new Artificial Intelligence (AI) platform that addressed the challenges of parallel computing at the software level. Researchers at

Google made significant advancements through their work on deep neural networks (DNNs), which played a key role in improving Long Short-Term Memory (LSTM) networks. Led by Jeff Dean, this project resulted in numerous achievements. Google combined Graphics Processing Units (GPUs) with deep learning techniques to accelerate computation for complex tasks. Unlike traditional central processing unit (CPU) cluster platforms, this new system significantly outperforms conventional systems in terms of speed. With these advancements, Google made a pivotal contribution to the evolution of Artificial Intelligence applications.

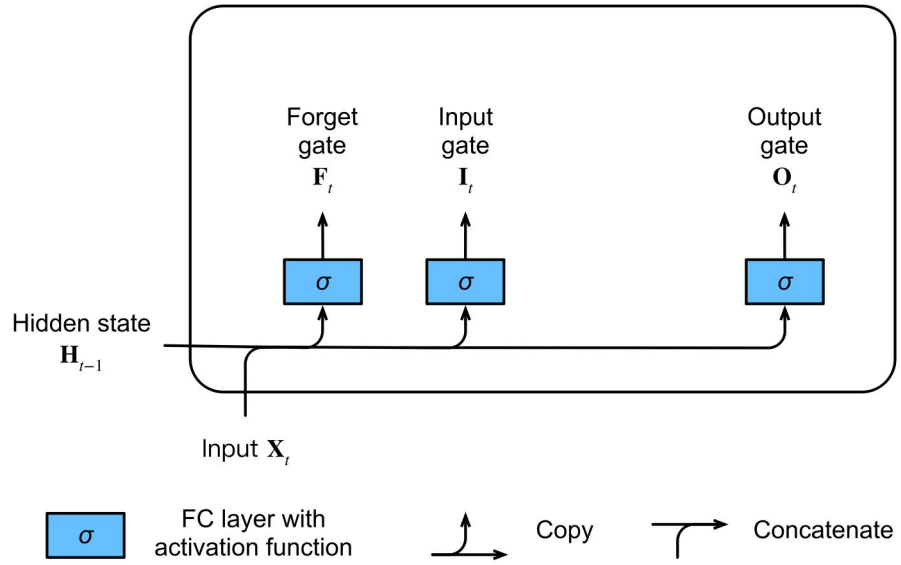


Building on this previous excellent work, Long Short-Term Memory (LSTM) networks can now effectively address the challenges posed by the following business case requirements. The basic structure of an LSTM consists of four foundational components: the Cell State, Forget Gate, Input Gate, and Output Gate. These components provide the core functionality required to achieve the goals of deep learning.

The Cell State plays a crucial role in capturing and retaining long-term dependencies in the input sequence. It functions as a memory unit that enables the network to retain information over extended periods, allowing the system to store and analyze relevant relationships across multiple contexts. Additionally, it facilitates the flow of relevant data through the network while discarding irrelevant information. This capability enables the model to handle complex sequential patterns in natural language.

Although social media is not composed solely of natural language, much of its content is still based on text, and the data can be translated into a form that can be understood in natural language. The Cell State enhances the model's ability to better understand and analyze human language with greater accuracy.

The other components of the gates are relatively easier to understand. Three gates—Forget, Input, and Output—work together to process the current time step, while also incorporating the hidden state from the previous time step. These fully connected layers, with advanced functions, compute the values for the Forget, Input, and Output gates. The Input gate determines how much information should be added to the current memory cell, while the Forget gate decides whether to retain or discard the value in the memory cell. Finally, the Output gate determines whether the memory cell should influence the output at the current time step.



The memory cell performs computationally intensive tasks, requiring high bandwidth and fast data transfer. The gates themselves do not perform specific actions; rather, the input node receives data from the previous state, and the cell computes values based on the provided functions within a defined range. These gates provide the model with flexibility, allowing it to learn more information while retaining the current results. This feature makes the model easier to train, even when faced with datasets containing long sequence lengths.

Long Short-Term Memory (LSTM) networks were introduced in 1997, featuring an innovative design and advanced framework. However, it was not until 2017, with the support of high-performance GPUs, that LSTMs were successfully applied in real-world research and experienced rapid growth. Due to their unique features, LSTMs are highly effective for analyzing complex social media content.

2.5 Challenges in Topic-Based Social Media Analysis

There is no perfect solution that can address all the challenges encountered in real-world cases. Although deep learning has seen significant development over the past 20 years, the challenges continue to grow alongside the expansion of social media platforms. As datasets increase in size, the demand for computational resources becomes more urgent, and the requirements grow exponentially. When the scale of data increases tenfold, building a platform based on LSTM models becomes highly complex. The capital expenditure required becomes unacceptable, and with such large-scale data, the results also become more difficult to interpret.

In the future, hardware engineers will continue to work on developing more powerful GPUs, and advanced hardware will enable faster computations. As a result, researchers are increasingly focusing on the software layer. Building on advancements in hardware, they aim to develop new frameworks that will improve training efficiency, reduce computational burdens, and enhance model interpretability.

2.6 Summary

Since the advent of social media, there has been a continuous effort to extract meaningful information from vast, complex, and high-speed data streams running across the internet. Techniques like Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) have proven effective for analyzing text content, particularly on traditional social media platforms. These techniques have evolved alongside the development of business needs, and high-performance hardware support is crucial for the best software designs. Long Short-Term Memory (LSTM) networks have demonstrated their value, particularly with the advancement of GPU technology.

Today, with powerful deep learning systems, social media content can be analyzed with high accuracy. However, challenges remain, and Recurrent Neural Networks (RNNs) and LSTM networks continue to offer viable solutions to address these issues. Looking ahead, new techniques may emerge to eventually replace them.

REFERENCES

Elman, Jeffrey L., "Finding Structure in Time," Cognitive Science, vol. 14, no. 2, pp. 179-211, 1990.

Chee-Hong Chan, Aixin Sun, and Ee-Peng Lim, "Automated Online News Classification With Personalization," 4th International Conference of Asian Digital Library (ICADL), pp. 320-329, December 2001.

P. Duygulu, K. Barnard, J. de Freitas, and D. A. Forsyth, "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," Proceedings of the European Conference on Computer Vision, pp. 97-112, 2002.

G. Carneiro and N. Vasconcelos, "Formulating Semantic Image Annotation as a Supervised Learning Problem," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 163-168, 2005.

L. Cao and L. Fei-Fei, "Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes," Proceedings of the IEEE International Conference on Computer Vision, pp. 1-8, 2007.

Motaz K. Saad, "The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification," 2010.

Tian Xia and Yanmei Chai, "An Improvement to TF-IDF: Term Distribution Based Term Weight Algorithm," Journal of Software, p. 413, 2011.

Y. Li, T. Li, and H. Liu, "Recent Advances in Feature Selection and Its Applications," Knowledge and Information Systems, vol. 53, pp. 551-577, 2017.

N.M. Ali, S.W. Jun, M.S. Karis, M.M. Ghazaly, and M.S.M. Aras, "Object Classification and Recognition Using Bag-of-Words (BoW) Model," IEEE 12th International Colloquium on Signal Processing & Its Applications (CSPA), pp. 216-220, March 2016.

J. Cao, T. Chen, and J. Fan, "Landmark Recognition with Compact BoW Histogram and Ensemble ELM," Multimedia Tools and Applications, vol. 75, pp. 2839-2857, 2016.

N. Passalis and A. Tefas, "Learning Bag-of-Features Pooling for Deep Convolutional Neural Networks," Proceedings of the IEEE International Conference on Computer Vision, pp. 5755-5763, 2017.

A.A.A. Karim and R.A. Sameer, "Image Classification Using Bag of Visual Words (BoVW)," Journal of Al-Nahrain University-Science, vol. 21, pp. 76-82, 2018.

N. Martinel, G.L. Foresti, and C. Micheloni, "Wide-Slice Residual Networks for Food Recognition," IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 567-576, March 2018.

Ganitkevitch, Juri, Benjamin Van Durme, and Chris Callison-Burch, "PPDB: The Paraphrase Database," Proceedings of HLT-NAACL, pp. 758-764, 2013.