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 will examine relevant issues and review previous studies on the topic. The initial section of the chapter will discuss the subject matter and the current state of research on social media. Subsequently, we will explore advanced techniques, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, and analyze their application in the analysis of social media content.

2.2 Traditional Approaches to Text and Topic Analysis

The rapid growth of social media has generated a vast amount of textual data, prompting researchers to develop methods for extracting meaningful insights from these large and dynamic datasets. Early approaches to text mining and topic modeling were primarily based on traditional statistical techniques, which established the foundation for more advanced methods in natural language processing (NLP). These traditional methods, including the Bag-of-Words (BoW) model, Latent Dirichlet Allocation (LDA), and Term Frequency-Inverse Document Frequency (TF-IDF), were originally designed for structured and well-formed text. They were employed in the analysis of social media content before the development of more sophisticated machine learning algorithms.

The Bag-of-Words (BoW) model is one of the simplest and most widely used techniques in text analysis. It represents text as a collection of individual words, disregarding grammar and word order while preserving the frequency of each word within the document. The BoW model is particularly effective for extracting basic features from short and concise social media posts, where the primary concern is the occurrence and frequency of words, rather than their syntactic or semantic relationships. However, its simplicity also imposes significant limitations. The model's inability to capture contextual information, such as word order or dependencies between terms, restricts its performance, particularly when dealing with more complex and nuanced social media data.

Term Frequency-Inverse Document Frequency (TF-IDF) is another foundational technique in text analysis. TF-IDF quantifies the importance of a word within a document

relative to a corpus, based on its frequency within the document (TF) and its rarity across the entire corpus (IDF). Words that appear frequently in a document but are rare across other documents are considered highly significant. This approach is particularly effective for identifying key terms within a document or set of documents and has been widely applied in social media analysis to detect trending topics or relevant keywords. However, similar to the Bag-of-Words (BoW) model, TF-IDF does not capture the semantic relationships between words, which limits its ability to fully understand the context in which the terms appear. Moreover, it faces challenges in addressing the noisy and informal nature of social media language, where abbreviations, slang, and hashtags often distort conventional word usage.

2.3 Topic-Based Analysis Using Deep Learning

While traditional methods have laid the foundation for text and topic analysis, they exhibit several inherent limitations, particularly when applied to social media data. One major challenge is the noise and informality that characterize social media language, which includes the use of emojis, hashtags, and colloquial expressions. Traditional models often struggle to process these features effectively, potentially leading to a loss of meaning or misclassification of topics. Furthermore, methods such as the Bag-of-Words (BoW) model and Term Frequency-Inverse Document Frequency (TF-IDF) are ill-suited to capturing semantic meaning, word dependencies, or contextual variations within social media posts. As a result, these techniques may fail to accurately identify the underlying topics or themes in more complex or nuanced social media content.

Despite these limitations, traditional approaches like BoW, Latent Dirichlet Allocation (LDA), and TF-IDF remain valuable tools for basic text analysis and topic extraction, particularly when the dataset is relatively clean and well-structured. However, as social media content becomes increasingly diverse and unstructured, there is a growing need for more advanced methods that can effectively address the dynamic nature of social media language. This demand has driven the development of sophisticated machine learning techniques, particularly deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which are better equipped to handle the complexity of modern social media analysis.

Deep learning methods have revolutionized the way topics are analyzed in social media text. By leveraging neural networks to automatically learn hierarchical features from raw text, deep learning techniques are able to model complex relationships and dependencies between words, sentences, and even entire documents. In the context of social media, where posts are often brief, fragmented, and informal, deep learning models

are particularly useful in capturing the semantic meaning behind the text, rather than relying solely on superficial keyword frequency.

Among the most popular deep learning approaches for topic-based analysis are Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers, each of which has distinct strengths. CNNs have been widely used for feature extraction from text, particularly for identifying patterns or phrases that are indicative of specific topics. RNNs, particularly Long Short-Term Memory (LSTM) networks, are designed to capture sequential dependencies within text, which is essential for understanding context in social media posts. In recent years, transformer-based models like BERT and GPT have gained traction due to their ability to capture long-range dependencies and contextual meaning with more efficiency and accuracy than previous models.

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

The Long Short-Term Memory (LSTM) networks are a relatively recent technique in the field of deep machine learning. According to recent studies, LSTMs have proven to be one of the most efficient methods for analyzing social media topics and subjects, surpassing traditional techniques in performance. The development of LSTMs began in 1997, but due to limitations in hardware technology, they gained widespread popularity only after 2015, following advancements in artificial intelligence and the development of Graphics Processing Unit (GPU) technology.

In 2012, one of the most significant milestones in artificial intelligence (AI) was Google's breakthrough in deep learning, particularly through its research on deep neural networks (DNNs). A key development during this period was the Google Brain project, led by Jeff Dean and other prominent researchers, which achieved remarkable results by leveraging Graphics Processing Units (GPUs) to accelerate deep learning tasks. Notably, this was the year in which Google utilized deep learning techniques to substantially enhance image recognition, exemplified by a model capable of learning to recognize cats in YouTube videos, as part of the broader efforts of the Google Brain initiative. This achievement built upon the architecture of GPUs, marking a pivotal moment in the evolution of AI research and applications.

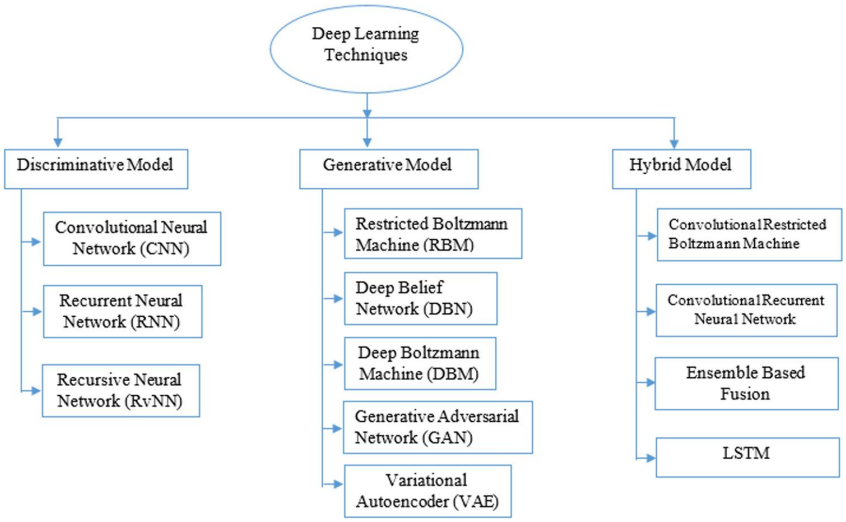
GPUs, initially designed for parallel processing in graphics rendering, proved to be highly efficient for the computationally intensive tasks required by machine learning algorithms. By harnessing the parallel processing capabilities of GPUs, Google was able to significantly accelerate the training of deep neural networks, enabling the development of more complex models and producing results at an unprecedented speed. This shift in

computational infrastructure facilitated a new era of deep learning research, characterized by increased model complexity and efficiency.

Simultaneously, NVIDIA introduced the CUDA (Compute Unified Device Architecture) software framework, which further advanced the integration of GPUs into AI and scientific computing. CUDA allowed developers to write software capable of fully utilizing the parallel processing power of GPUs, thereby drastically improving the speed and efficiency of calculations required for AI tasks, such as the training of large-scale neural networks and the processing of vast datasets. The combination of Google's deep learning advancements and NVIDIA's CUDA software revolutionized the efficiency of machine learning workflows, empowering researchers and developers to push the boundaries of AI technology. The impact of these innovations was profound, laying the foundation for the rapid advancements in AI that followed in the subsequent years.

LSTMs address the vanishing gradient problem by incorporating a more sophisticated architecture that includes gates to regulate the flow of information. The key components of an LSTM unit are the Cell State, Forget Gate, Input Gate, and Output Gate. These gates enable the LSTM to preserve and manipulate information over multiple time steps, making it particularly well-suited for tasks that require capturing long-range dependencies, such as understanding the context of multi-turn conversations on social media or identifying topics that evolve over time.

LSTMs have been widely applied to various tasks in social media analysis, particularly those related to topic detection, sentiment analysis, and misinformation identification. One key advantage of LSTMs is their ability to track the evolution of topics over time. By retaining information about previous discussions, LSTMs can use this context to determine the topics of new posts. This feature is particularly valuable on platforms like Twitter, where topics can change rapidly, and the model must capture both short-term trends and long-term thematic shifts.



In sentiment analysis, LSTMs are frequently used to determine the sentiment of social media posts, even when these posts are brief and informal. By modeling the sequential dependencies between words and phrases, LSTMs can classify posts as positive, negative, or neutral, which often serves as a precursor to identifying whether misinformation or misleading content is being shared.

Additionally, LSTMs can be leveraged to detect misinformation by identifying anomalous patterns in the topics being discussed. By modeling the typical flow of conversations and comparing it to new posts, LSTMs can flag posts that deviate significantly from the expected topic or tone, which may indicate potential misinformation.

2.5 Challenges in Topic-Based Social Media Analysis

Social media platforms have become a rich source of textual data, providing valuable insights into public opinion, trends, and user behavior. However, analyzing this data presents unique challenges due to the inherent characteristics of social media language and its context. Unlike traditional forms of written communication, social media content is often informal, noisy, and highly dynamic. This section examines the specific challenges posed by social media data in topic-based analysis and explores how deep learning techniques, particularly Long Short-Term Memory (LSTM) networks, address issues such as informal and evolving language, the use of emojis, hashtags, and mentions, and the presence of multimodal content.

Despite their advantages, LSTMs are not without limitations. One significant issue is their computational complexity: LSTMs require considerable computational resources, particularly when dealing with large datasets typical of social media platforms. Another challenge is model interpretability, as the complexity of LSTMs makes it difficult to

understand how the model arrives at a particular decision. This lack of transparency can be problematic, especially in critical applications such as misinformation detection.

Future research on LSTMs is likely to focus on improving training efficiency, reducing the computational burden, and enhancing model interpretability through techniques like attention mechanisms and explainable AI (XAI). Additionally, integrating LSTMs with other architectures, such as transformers, may lead to even more powerful models for topic detection and misinformation analysis on social media.

2.6 Summary

The evolution of text and topic analysis in social media is marked by the transition from traditional methods to modern deep learning techniques. Traditional approaches, such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and Latent Dirichlet Allocation (LDA), have played a foundational role in the processing of textual data. However, these methods exhibit significant limitations in addressing the informal and dynamic nature of social media content, including the prevalence of abbreviations, slang, and hashtags. Additionally, they struggle to effectively capture semantic relationships or contextual meaning.

To overcome these challenges, this review emphasizes the increasing prominence of deep learning methods, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. These advanced techniques are particularly effective in modeling complex dependencies and extracting semantic insights from the unstructured and noisy data typical of social media platforms. In particular, LSTMs are recognized for their ability to manage long-range dependencies, making them well-suited for tasks such as topic detection, sentiment analysis, and misinformation detection. Nevertheless, these methods face challenges related to their high computational requirements and limited interpretability.

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.