International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n1.19
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Sentiment Analysis Based on Transformer

-- The Sentiment Analysis about the boy and the heron

Boyuan Yang

Jinan No.1 High School, Jinan, Shandong Province, 250100, China
2055005795fegnkuang10@gmail.com

ABSTRACT

This article aims to explore the sentiment analysis of Transformer comments on Weibo to understand the attitudes and emotional tendencies of social media users regarding the movie The boy and the heron. This paper uses natural language processing technology and sentiment analysis model Transformer to crawl the Weibo comment data set, analyzes the anti-crawling mechanism of Weibo, adds HEADERS and REFERER directly in the code to bypass the inspection, and preprocesses it. Sentiment analysis was conducted on a large number of Weibo comments related to the boy and the heron. The results of the study showed that the vast majority of the comments expressed positive emotional attitudes and held positive and appreciative attitudes towards the boy and the heron. This study reveals that social media users have positive attitudes and emotional identification with the boy and the heron, and the audience really likes the film, which is of great help for the filmmakers to further understand the advantages and disadvantages of the film. Performing Weibo sentiment analysis based on advanced natural language processing technologies such as transformers can not only improve the accuracy and efficiency of sentiment analysis, but also help promote the progress and application of artificial intelligence technology. This is important for the development of more intelligent social media analysis tools and systems.

KEYWORDS

Sentiment analysis; Transformer; Weibo; The Boy and the Heron

1. INTRODUCTION

In the field of artificial intelligence, the term "Transformer" specifically refers to a neural network architecture that has gained significant attention in recent years. Introduced by Vaswani et al. in their seminal paper "Attention is All You Need", transformers have revolutionized sequence transduction or neural machine translation, which involves transforming an input sequence into an output sequence for tasks such as speech recognition and text-to-speech conversion [1]. The fundamental operation of Transformers revolves around the concept of "attention". Essentially, the attention mechanism enables the model to focus on specific elements within the input sequence, thereby enhancing its ability to comprehend context and dependencies between words or features [2]. This is particularly crucial in language translation tasks where word meaning often depends on context or relationships with other words within a sentence. Over the past decade, convolutional neural networks have been the dominant deep neural architectures used in computer vision. However, transformers, which are based on self-attention mechanisms and can capture relationships between different features, have gained popularity in natural language GPT-3. This success has led to increased research into using transformers for vision tasks. Researchers have explored various methods of representing sequence

information from different data sources to leverage transformer architecture for these tasks. For example, Wang et al. investigated nonlocal networks with self-attention mechanisms to capture long-range dependencies in video and image recognition while Carion et al.'s DETR treats object detection as a direct set prediction problem solved by a transformer encoder-decoder architecture. Chen et al.'s iGPT is another pioneering work that applies pure transformer models to self-supervised pretraining for image recognition. Weibo, owing to their vast user base and diverse topical coverage, wield significant influence over the daily lives of individuals. Consequently, sentiment analysis of Weibo assumes paramount importance, as it offers invaluable insights into the collective sentiments, opinions, and reactions permeating through online discourse. By scrutinizing Weibo content, sentiment analysis enables researchers, businesses, and policymakers alike to comprehend public sentiment, gauge reactions to events or products, and inform decision-making processes. Thus, leveraging sentiment analysis in the context of Weibo holds immense potential for unraveling societal trends, understanding public sentiment dynamics, and enhancing decision-making frameworks.

2. LITERATURE REVIEW

Currently, sentiment analysis primarily relies on traditional machine learning algorithms such as SVM, information entropy, CRF, etc. These methods can be categorized into supervised learning, unsupervised learning, and semi-supervised learning approaches. While supervised learning has shown promising results due to its reliance on large amounts of manually labeled data, it incurs high labeling costs for the system. On the other hand, unsupervised learning eliminates the need for manual labeling but often falls short in meeting practical requirements due to its dependence solely on algorithmic outcomes. Semi-supervised learning strikes a balance by leveraging both a small number of labeled samples and a larger pool of unlabeled samples to enhance overall performance while considering manual labeling costs. In this paper, we propose a novel deep learning algorithm that utilizes semi-supervised learning methodology along with an autoencoder algorithm for analyzing sentiment tendencies in Chinese Weibo. These Weibo with emotional information is very valuable resources. Sentiment analysis can obtain the mood of netizens at this time and the view of a certain event or thing, which can excavate its potential commercial value and make a certain contribution to social stability.

3. MATERIALS AND METHODS

3.1. Dataset Preparation and Preprocessing

Collect a dataset of Weibo comments, including the comment text and the corresponding sentiment labels (positive, negative, neutral). The collected data were cleaned to remove noisy data and duplicate comments. Word segmentation is used to segment the review text into a sequence of words or subwords. Preprocessing operations such as removal of stop words and stemming are performed as needed.

3.1.1. Anti-crawler mechanism of Weibo

Currently, most websites check the HEADERS 'USER-AGENT, and some even check the REFERER. If we encounter this kind of anti-crawler mechanism, we can simply add HEADERS and REFERER to the code to bypass the check. For these sites, adding or modifying HEADERS and REFER in the code is a good way around this.

3.1.2. Crawlers based on user behavior

The behavior of users accessing the website is also a common detection method for mainstream websites. For example, the same IP has visited the same page many times in a short period of time, and some things the same account has performed the same operation many times in a short period of

time [3]. For this case, we can use IP proxy to solve. There are now paid and free IP proxies on the web, and we can crawl to them and store them, and then change the IP every few requests.

3.1.3. Start crawling reviews for The Boy and the Heron

The reviews for The Boy and the Heron are crawled using Python, stored in data.csv, and preprocessed. As a social platform, Weibo not only possesses the characteristics of rapid dissemination but also serves as a key release platform utilized by businesses for product promotion. In Weibo, the presence of numerous advertising and marketing accounts poses significant challenges to sentiment analysis. Therefore, text preprocessing for Weibo has become exceedingly crucial.

3.1.4. Crawlers based on user behavior

There are also sites where data is generated via AJAX or JS requests. We can use the browser to analyze the requests that are made to a website. If the AJAX request is found, it can be analyzed and solved using either of these methods to retrieve the corresponding data.

If you can't get the AJAX request, you can call the selenium+phantomjs framework, calling its browser kernel, to simulate human actions and trigger the JS script of the page.

3.1.5. Anti-crawler for Weibo

The anti-crawler in Weibo uses the above three mechanisms to verify the HEADERS of the client, and at the same time to prohibit access to the consent IP with large amount of access, and uses AJAX for data transmission. To break this kind of crawler, you must use IP proxy, access the same account at different times, add HEADERS, etc.

3.2. Building a Sentiment Analysis Model

Using a Transformer model and fine-tuning it for the task of sentiment analysis of Weibo comments involves adjusting hyperparameters such as the learning rate and batch size to optimize performance. The architecture of the Transformer model encompasses several key components. At its core lies an Encoder, which is composed of multiple encoder layers responsible for processing the input sequence. Each encoder layer consists of a Multi-head self-attention mechanism, adept at capturing global dependencies within the input [4]. Additionally, the Encoder incorporates Feedforward neural network layers to further process information gleaned from the attention mechanism. To aid in training stability and facilitate smoother flow of gradients, Layer normalization and residual connections are employed throughout the model. This comprehensive architecture ensures that the Transformer model can effectively analyze the sentiment of Weibo comments with precision and accuracy.

3.3. Train Model

After fine-tuning the model, the next step involves utilizing it to train on preprocessed Weibo comment data. Throughout the training process, meticulous attention is paid to the loss function and performance metrics, ensuring the model converges effectively and achieves the desired level of performance. The Transformer model is trained on the prepared dataset, with appropriate loss functions defined based on the sentiment analysis task. To optimize training, advanced optimization techniques such as the Adam optimizer with learning rate scheduling are employed. Constant monitoring of training progress is conducted using validation metrics, allowing for adjustments and fine-tuning as necessary. Additionally, early stopping mechanisms are implemented to prevent overfitting and ensure optimal model performance. This comprehensive approach ensures that the Transformer model is effectively trained to analyze sentiment in Weibo comments with accuracy and reliability.

3.4. Model Evaluation

In classification tasks such as sentiment analysis, where understanding the accuracy of predictions is crucial, a range of metrics is employed to assess the performance of Transformer models. These include accuracy, which measures the proportion of correctly classified instances among all instances. Precision, on the other hand, reflects the proportion of true positive predictions out of all positive predictions, offering insights into the model's ability to avoid false positives. Recall, also known as sensitivity, gauges the proportion of true positive predictions out of all actual positive instances, highlighting the model's capacity to capture all relevant instances. F1-score, a harmonic mean of precision and recall, provides a balanced measure of a model's performance in terms of both precision and recall. Additionally, the area under the Receiver Operating Characteristic (ROC) curve, known as AUC-ROC, evaluates the model's ability to distinguish between positive and negative instances across various threshold settings, offering a comprehensive assessment of its discriminatory power. These metrics collectively offer valuable insights into the performance of Transformer models in classification tasks, facilitating informed decision-making and model optimization [5].

3.5. Results Analysis

Analyze the performance of the model on different emotion categories, identify the problems of the model and the room for improvement. The model structure is adjusted according to the analysis results. According to the evaluation results, the model is tuned, such as adjusting the hyperparameters, improving the model structure, etc. to improve the performance and generalization ability of the model. There were more than 10,00 reviews, of which 847 were positive, 200 were neutral, and only 35 were negative. It is evident that "The Boy and the Heron" is indeed a commendable film, with predominantly positive feedback from the public. The emotional depth of the narrative captivates viewers, akin to witnessing an elderly man meticulously introspecting his innermost sentiments before an audience. His journey towards reconciliation with his "family," "world," and "self" evokes profound emotions through his genuine sincerity and unwavering courage. From my perspective, his life truly mirrors that of the film's protagonist, making his act of baring his soul all the more remarkable. Moreover, amidst such arduous circumstances, he manages to wield his pen to impart invaluable wisdom and reverence for life. As I reached the story's conclusion, I beheld the extraordinary magnanimity concealed within an ordinary old man; tears welled up in my eyes as a testament to this kind-hearted and courageous individual who stands as a magnificent creator.

4. CONCLUSION

The primary focus of this study is the sentiment analysis of movie reviews for "The Boy and the Heron". Firstly, it retrieves comments about this movie from Weibo for preprocessing purposes. By employing the Transformer model to analyze the sentiment of these reviews, out of a dataset comprising more than 10,00 reviews, 847 are classified as positive, 200 as neutral, and only 35 as negative. Given the continuous growth of Weibo platforms, there remain numerous aspects that warrant further investigation. Within the context of sentiment analysis, it becomes possible to analyze users' emotional orientations towards certain Weibo comments such as trending events and ascertain prevailing sentiments among netizens regarding these events. Additionally, by considering temporal factors along with location and individuals involved in these trending events, deeper insights can be extracted through data mining techniques including predictive analysis. Such endeavors contribute significantly towards ensuring social stability. Additionally, the sentiment classification strategy can be modified to enhance the accuracy of analyzing linguistic phenomena exhibited by users, such as examining their usage of degree adverbs like "very" and "super", as well as incorporating punctuation and repeated words in the text for a comprehensive overall modeling. Furthermore, apart from capturing trending events, an individual's entire collection of Weibo can also be acquired for analysis

purposes. By scrutinizing all Weibo posted by an individual, we can derive insights into their emotional inclinations towards specific events or preferences for certain brands.

REFERENCE

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press, Cambridge.

[2] Gamb ń, Á.F., Yazidi, A., Vasilakos, A., et al. (2024) Deepfakes: current and future trends. Artif Intell Rev 57, 64.

[3] Wang, Y. & Yu, H. (2022) Research on Internet Corpus Collection Method. In: *2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, Dalian, pp. 61-67.

[4] Voita, E., Talbot, D., Moiseev, F., et al. (2019) Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. arXiv preprint arXiv:1905.09418.

[5] McAvaney, B.J., Covey, C., Joussaume, S., et al. (2001) Model evaluation. In Climate Change 2001: The scientific basis. Contribution of WG1 to the Third Assessment Report of the IPCC (TAR). Cambridge University Press, pp. 471-523.