WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n2.02
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Public Opinion Monitoring of Sports Stars Based on Text Sentiment Analysis

Lingfeng Yu

Aberdeen School of Data Science and Artificial Intelligence, South China Normal University, Foshan, 528299, China
yu1.22@abdn.ac.uk

ABSTRACT

Traditional methods of monitoring public opinion often rely on questionnaires or ballots, which are not only time-consuming and labour-intensive, but also difficult to reflect public sentiment changes and trends in real time. With the development of machine learning and natural language processing technologies, it has become possible to automatically analyse large-scale text data by algorithms and identify the emotional tendencies therein. In this paper, the analysis and statistics of public opinion are achieved through relevant algorithmic models.

KEYWORDS

Natural language processing, Opinion monitoring, Sports stars

1. INTRODUCTION

1.1. Status of Public Opinion on the Internet

With the rapid development of the Internet and social media, online public opinion has become an important part of social opinion. Internet public opinion is characterised by fast dissemination, wide influence and large number of participants. By monitoring online public opinion, we can understand the public's emotions and attitudes in time and predict the trend of public opinion so as to provide support for decision-making. The monitoring of online public opinion is not only limited to the field of sports, but also plays an important role in many other fields such as politics, economy and culture. For example, in the political field, monitoring online public opinion can help government departments adjust their policy direction by understanding the public's attitude and view on policies. In the economic field, enterprises can monitor online public opinion to understand consumers' needs and feedbacks, so as to adjust their product and service strategies.

1.2. The Need for Public Opinion Monitoring

In recent years, a variety of online media has flourished,making the breadth of information dissemination greatly increased. And as people pay more and more attention to related sports, a number of well-known sports stars have been born. However, what comes along with them is the PR crisis of sports stars. In the process of sports stars' daily training, competition, life and participation in social activities, crisis events occur from time to time, such as Liu Xiang's withdrawal from the competition, Sun Yang's taking of banned substances, and Lin Dan's derailment. [1] By analysing the popularity and public image of sports stars on social media, provides sports clubs and brand sponsors with accurate marketing strategy support. Knowing which sports stars can

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

attract public attention can help plan advertising and branding campaigns more effectively. Assessing the popularity of sports stars can help clubs and organisations to understand the preferences of fans and then design more attractive fan interaction activities to enhance fans' engagement and sense of belonging.

Therefore, if public opinion is not monitored, it may cause damage to the interests of the endorsers related to sports stars in a minor case, or bring adverse demonstration effects to the society in a serious case. Therefore, how to realise the monitoring of public opinion is very necessary.

1.3. Characteristics of Public Opinion of Sports Stars

The public opinion characteristics of some popular sports stars nowadays are characterized by "rice circle". The main characteristic is that the competition results are no longer the only criterion to get the public's love and support. The public has begun to take into account the appearance of the athlete, personality traits and other criteria. In the early years, fans chased the famous swimmer Ning Zetao, and even a lot of middle-aged women called him their "national husband" in the same way that showbiz celebrities do. In addition, since the opening ceremony of the Tokyo Olympics, netizens on some social media platforms have been dedicated to digging out the so-called "divine face" athletes. [2] Nowadays, there are quite a number of athletes who do not have excellent strength and have not won the relevant competition honors, but through the operation of social media, selling appearance, gained a lot of fans, causing a lot of controversy. Some people think that the fundamental thing for athletes is to have excellent strength and honor, while others think that they can like athletes also like her appearance and character. Therefore, this paper mainly analyzes the controversial example of the change of public opinion before and after the track and field athlete Yanny Wu's snatch at the competition.

1.4. Case Study: Yanny Wu's Robocall Incident

Yanny Wu is a highly regarded track and field athlete who was fined for snatching a run in an important race on 1 October 2023, an event that has been widely discussed on social media. In order to study the impact of this event on public sentiment, this paper collects data on comments from two time periods before and after Yanny Wu's snatch run.

After the snatch-run incident, Yanny Wu's public image changed significantly. By analysing the comment data before and after the snatch, we found that before the incident, most of Yanny Wu's public comments were positive, and the common adjectives used in the comments included "excellent" and "potential", etc. However, after the incident, the negative comments were mostly positive. However, after the incident, there was an upsurge in negative comments, with negative terms such as "arrogant" and "irresponsible" appearing in the comments.

This case shows that major negative events have a profound impact on the public image of sports stars. By monitoring and analysing public opinion in a timely manner, it can help the relevant parties to take timely countermeasures to reduce the damage of negative events on the image of sports stars.

2. RESEARCH INTRODUCTION

2.1. Model Introduction

2.1.1. Commonly Used Text Sentiment Analysis Methods

Convolutional Neural Networks (CNNs) initially achieved great success in the field of computer vision and were later introduced to the field of natural language processing.CNNs extract local features by performing convolutional operations on text and reduce the dimensionality of the features through pooling operations to improve computational efficiency. The convolution operation extracts

n-gram features of the text through convolution kernels of different sizes. The convolution kernel slides over the text to capture local word sequence features, such as word phrases. Pooling operations usually use Max-Pooling or Average-Pooling to downscale the output of the convolutional layer, retaining the most important features and improving the generalisation ability of the model. CNNs perform well in text classification and sentiment analysis tasks, capturing local patterns and important information in text.

Long Short-Term Memory Network (LSTM) is a special kind of Recurrent Neural Network (RNN) that excels in processing and predicting time series data. LSTM effectively solves the problems of gradient vanishing and gradient exploding in traditional RNNs by introducing forgetting gates, input gates and output gates. The forgetting gate determines how much of the input information at the current moment is retained; the input gate determines how much of the input information at the current moment is updated to the cell state; and the output gate determines how much of the information in the cell state is outputted to the next moment. LSTM is able to memorise long-distance dependencies, which makes it very suitable for processing sequential information in textual data, and it also performs well in sentiment analysis.

BERT model structure introduction

2.1.2. Transformer structure

The Transformer model, proposed by Vaswani et al in 2017, is a deep learning model based on a self-attention mechanism and widely used in the field of natural language processing (NLP). The basic units of the Transformer model are the encoder and decoder, with each encoder and decoder layer consisting of multi-head self-attention mechanisms and feedforward neural networks.

Multi-head self-attention mechanism: Self-attention mechanism captures global context information by calculating the relative weight of each word in the input sequence with other words. By introducing multiple parallel self-attention layers, the multi-head self-attention mechanism enables the model to pay attention to a variety of information at different positions in the input sequence, thus improving the expressive power of the model.

Feedforward neural network: Each self-attention layer is followed by a feedforward neural network, which is used to nonlinear transform the output of the self-attention layer. Feedforward neural networks usually consist of two linear transformation layers and a ReLU activation function.

BERT model uses Transformer encoder part, stacked 12-layer encoder (Bert-base) or 24-layer encoder (Bert-large), The Mask Language Model (MLM) and Next Sentence Prediction (NSP) tasks were used in the pre-training phase.

2.2. Bidirectional Characteristics of BERT

An important innovation of the BERT model is its bidirectional nature. Traditional language models, such as GPT (Generative Pre-trained Transformer), can only generate word representations from left to right or from right to left, and cannot simultaneously consider left and right context information. BERT, through the MLM task, randomly masks part of the words in the input sequence, forcing the model to predict the masked words simultaneously from left to right and from right to left during the training process, thus learning a more comprehensive word representation.

In addition, the NSP task allows BERT to understand the sentence-level relationship by having the model predict whether two sentences occur consecutively, and thus perform better on the sentence-pair task.

The paper "Distilling the Knowledge in a Neural Network" by Geoffrey Hinton, Oriol Vinyals, and Jeff Dean discusses a way to improve the performance of machine learning algorithms by Training multiple models on the same data and then averaging their predictions. However, using a set of models can be computationally expensive and difficult to deploy. The authors propose a technique called

"distillation" that compresses the knowledge in a set into a more easily deployable model. They show that this approach can significantly improve model performance on tasks such as MNIST and speech recognition. [3] The Distilled BERT model used in this paper uses exactly this approach. Bert is a 12-layer transformer encode while Distilled BERT is a 6-layer transformer encode. The Distilled BERT used in this paper is pre-trained by loading some of the parameters of Bert directly into the Distilled BERT structure as initialization, which can then be fine-tuned for good performance in a wide range of tasks. Related research has shown it is possible to reduce the size of a BERT model by 40% while retaining 97% of its language comprehension and increasing its speed by 60% [4].

2.3. The Need for BERT Distillation

Bert-based models are very popular in NLP as they were originally introduced in [5]. As performance improved, many, many parameters emerged. To be precise, BERT has more than 110 million parameters, and BERT-large has not been discussed here. If you have ever trained a large NLP model such as BERT or RoBERTa, you know that the process is very long. Due to the sheer size of such models, training can take days. When it comes time to run them on small devices, you may find that you are paying a huge cost in memory and time for today's ever-increasing performance. From this, the need for model distillation is obvious, as BERT is very versatile and performs well, as well as the fact that later models are built in essentially the same way, similar to RoBERTa [6]. So being able to correctly extract and use the content contained within BERT allows us to achieve two things at once. The concept of distillation is very intuitive: it is the process of training small student models to mimic large teacher models as closely as possible.

2.4. Use of Models

This paper uses the Distilled BERT model that has been pre- trained and debugged. Pre-training a model avoids the need to train the model from scratch, which typically requires significant computational resources and time. Using pre-trained models significantly reduces the time spent on model development and training, enabling researchers and developers to deploy applications faster. Pre-trained models are typically trained on large-scale datasets, which allows them to learn rich linguistic features and generalized textual representations. The breadth and depth of this learning typically provides better performance in downstream tasks, especially with less data. In contrast, this paper investigates the domain of sports and athletics, where there is less relevant comment data, making the use of pre-trained models more appropriate. Flexibility and Adaptability: pre-trained models can be fine-tuned to target specific tasks. The flexibility of this approach allows a single pre-trained model to be adapted to many different tasks and needs. Reduced risk of overfitting: by using a pre-trained model as a starting point, the model has already learned a large number of linguistic features, thus reducing the risk of overfitting even on smaller datasets, which is difficult to do when training completely from scratch. Moreover, the Distilled BERT model used in this paper is parented to BERT. a large number of studies have demonstrated the broad applicability of this model, enabling successful migration to multiple domains and tasks.

Model Basics: Distilled BERT is a "lightweight" version of the BERT model, designed to maintain similar performance while reducing model complexity and runtime. It is distilled from

BERT through a process called knowledge distillation, which reduces model size and increases speed while maintaining model performance. It initializes a sentiment analysis pipeline using the Hugging Face transformers library. This sentiment analysis model uses distilbert-base-uncased-finetuned-sst-2- english, a model based on the Distilled BERT architecture and fine-tuned for sentiment analysis tasks.

Training dataset: SST-2 (Stanford Sentiment Treebank): This model is fine-tuned using the SST-2 dataset, a standard dataset widely used for sentiment analysis that contains sentences from movie reviews, each labeled with its sentiment tendency (positive or negative).

Performance and Usage: Since the model is optimized for sentiment analysis tasks, it performs well in understanding sentiment tendencies in text and is suitable for application scenarios that require fast and accurate sentiment judgments, such as social media analytics, customer feedback analytics, and so on.

Risks, Limitations, and Bias: Based on some experiments, we observe that the model may produce biased predictions for underrepresented populations. For example, when nothing in the input indicates such a strong semantic change, this binary classification model will give very different probabilities based on the country/region being positively labeled (0.89 if the country is France, but 0.08 if the country/region is Afghanistan).

3. SENTIMENT ANALYSIS MECHANISM

The model performs sentiment analysis for each comment and updates the counters for positive and negative comments based on the analysis results. At the same time, a composite sentiment score is calculated, with positive sentiment increasing the score based on its confidence level and negative sentiment decreasing the score. The output sentiment score (or confidence score) reflects how confident the model is in its predictions (positive or negative). The closer the score is to 1, the more confident the model is in its predictions. When the model predicts a POSITIVE outcome, the corresponding score indicates the level of confidence that the model believes the comment is positive. The higher the score, the more confident the model is that the comment is positive. When the model predicts a NEGATIVE result, the corresponding score indicates the level of confidence that the model considers the comment to be negative. The higher the score, the more confident the model is that the review is negative. This paper also adds a function of word frequency statistics. Segmentation and lexical labeling are performed on each comment, adjectives are extracted, and the five adjectives with the highest frequency of occurrence are counted. It is convenient to make relevant solutions after specific analysis of opinion monitoring.

3.1. Analyzing Object Datasets

The dataset that is the subject of analysis in this paper is mainly from the social media platforms Facebook and Twitter. To ensure the timeliness and relevance of the data, the dataset is divided by the time segmentation point of Yanny Wu's snatch-and-run event on 1 October 2023, and the comment data before and after the snatch-and-run event are collected separately. Each time segment of the dataset contains 200 comments, which are all within two months before and after the event to ensure the timeliness and relevance of the comments.

The data collection process was carried out in strict accordance with the following steps:

Data collection: social media API interfaces were used to capture comment data related to Yanny Wu. In order to ensure the representativeness of the data, we chose comments with a high degree of activity on social media platforms.

Data cleaning: the crawled data were pre-processed to remove noisy data, duplicate data, and comments irrelevant to the analysis. The retained comments are all users' real reactions to Yanny Wu's grabbing incident.

Data labelling: In order to facilitate the subsequent sentiment analysis, the comments are labelled according to two categories: positive and negative. The labelling was done by professionals to ensure the accuracy and consistency of the labelling results.

3.2. Data Entry and Analysis of Results

In order to analyse the impact of Yanny Wu's snatch-and-run incident on public sentiment, we input the comment data before and after the snatch-and-run incident into the Distilled BERT model for sentiment analysis respectively. The analysis results are as follows:

Sentiment analysis results of the comment data before the snatch-and-run:

Total comments: 200

Positive comments: 171

Negative comments: 29

Average sentiment score: 0.72

Top 5 adjectives and their frequencies: [('first', 30), ('new', 20), ('personal', 20), ('great', 20), ('young', 19)]

From the above results, it can be seen that before the snatch-and-run incident, Yanny Wu's public image was more positive. Most of the comments were positive towards her, with an average sentiment score of 0.72, indicating a high level of public recognition. The most frequent adjectives such as "first" and "great" also reflect the public's positive impression of her.

Sentiment analysis results of the post-snap comment data: Total comments: 200

Positive comments: 76

Negative comments: 124

Average sentiment score: -0.28

Top 5 adjectives and their frequencies: [(‘not’, 36), (‘wrong’, 31)), (‘arrogant’, 30), (‘high’, 24), (‘refueling’, 17)]

Compare the model outputs before and after the preemption:

Table 1. Comparison of public opinion before and after Yanny Wu's snatch-and-run incident

	Before jump the gun	After jump the gun
Total comments	200	200
Positive comments	174	76
Negative comments	29	124
Average sentiment score	0.72	-0.28
Top 5 adjectives and their frequencies	(‘first’, 30), (‘new’, 20), (‘personal’, 20), (‘great’, 20), (‘young’, 19)	(‘not’, 36), (‘wrong’, 31), (‘arrogant’, 30), (‘high’, 24), (‘refueling’, 17)

In contrast, Yanny Wu's public image dropped significantly after the snatch-and-run incident. The number of negative comments increased dramatically and the average sentiment score dropped to -0.28, indicating a significant increase in negative public sentiment towards her. The highest frequency of negative adjectives such as "not", "wrong", and "arrogant" indicate the public's negative perception of her. This result suggests that major negative events have a profound impact on the public image of sports stars. Through timely public opinion monitoring and sentiment analyses, relevant parties can quickly understand the changes in public sentiment and adopt effective public relations strategies to mitigate the damage of negative events on the image of sports stars.

4. SUMMARY

This study monitored and analysed the public opinion of sports star Yanny Wu by using machine learning and natural language processing techniques, especially the Distilled BERT model. By comparing the data of comments before and after the incident, it is found that before the incident, Yanny Wu's public image is more positive, and most of the comments have a positive attitude towards her, with an average sentiment score of 0.72. The statistics of adjective frequency shows that positive adjectives often appear in the comments, such as "first" and "great". great". After the incident, there was a significant increase in negative comments, with the average sentiment score dropping to -0.28. Many negative adjectives such as "arrogant" and "wrong" appeared in the comments. This shows a significant increase in negative public sentiment towards her. This result shows that major negative events have a profound impact on the public image of sports stars, and timely monitoring of public opinion and sentiment analysis can help the relevant parties to quickly understand the changes in public sentiment and adopt effective public relations strategies to mitigate the damage of negative events on the image of sports stars.

Through this study, we not only verified the effectiveness of Distilled BERT model in sentiment analysis, but also demonstrated the practical application value of machine learning and natural language processing techniques in public opinion monitoring. The results show that these techniques can efficiently and accurately process large-scale text data and capture changes in public sentiment in real time, providing powerful data support for relevant decision-making.

5. OUTLOOK

In future research and applications, we plan to further optimise and expand the opinion monitoring system to improve its accuracy, real-time performance and applicability. Firstly, more social media platforms and news websites will be included to obtain more comprehensive public opinion information, including data in different languages and cultures, so as to improve the comprehensiveness and accuracy of the analyses. Second, we will continue to explore and introduce more advanced machine learning algorithms and models, such as Generative Adversarial Networks (GANs) and Reinforcement Learning, etc., to enhance the accuracy and efficiency of sentiment analysis. In terms of application, we will not only use it within the sports field, but also expand it to many other fields, such as politics, economy, culture, etc. We will validate and improve the versatility and adaptability of the system through multi-disciplinary application research. In addition, we plan to work with sports clubs, brand sponsors and other relevant stakeholders to provide accurate public opinion analysis reports and marketing strategy support to help them better understand and respond to changes in public opinion and develop more effective PR and marketing strategies. Finally, we will continuously improve the system functions and user experience through user feedback to ensure that the public opinion monitoring system meets the actual needs and to improve the stability and reliability of the system. Through these efforts, we hope to continuously improve the performance and application value of the Public Opinion Monitoring System, provide powerful information support and decision-making basis for all parties, and help them cope with the complex and changing public opinion environment.

REFERENCES

[1] Chen, X., Liu, H., Fan, F., et al. (2018). Research on crisis public relations of sports stars in China in the era of new media. Journal of Shenyang Sports Institute, 37(6), 30-36.

[2] Yu, Z. (2022). Curbing the spread of sports "rice circleization" chaos is imperative. Teaching Method Innovation and Practice, 5(2).

[3] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

[4] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Google AI Language.

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.