Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 14, No. 1, 2025

Application of Social Media Medical Data in Public Opinion Monitoring and Psychological Counseling During Public Health Emergencies

Tianyu Yang, Sining Chai

Northeastern University, Shenyang, Liaoning, China

Abstract: The massive amount of medical data generated by social media provides a new perspective for responding to public health emergencies. Focusing on the field of big data and data analysis, this paper deeply explores the application of social media medical data in public opinion monitoring and psychological counseling during public health emergencies. Guided by technical implementation as the core, it constructs a full-process technical framework of "data collection - preprocessing - public opinion monitoring - psychological counseling", and thoroughly analyzes the principles and optimization paths of key technologies such as distributed crawlers, medical field word segmentation, and multi-feature fusion. By verifying the performance differences of different technical schemes, it illustrates the technical implementation effects combined with practical cases, and comprehensively demonstrates its specific applications in real-time public opinion monitoring, trend prediction, insight into public psychological states, and formulation of psychological counseling strategies. The aim is to provide technical support and practical guidance for improving the response capacity to public health emergencies and realizing scientific and efficient public opinion management and psychological intervention.

Keywords: Social Media Medical Data; Public Health Emergencies; Public Opinion Monitoring; Psychological Counseling; Big Data Analysis.

1. Introduction

In recent years, public health emergencies such as the Ebola epidemic and Zika virus epidemic have occurred frequently, posing a serious threat to global public health security. In the response to these incidents, the spread of public opinion and the psychological state of the public profoundly affect the prevention and control effects and social stability. With a huge user base of over 4.8 billion, social media has become an important channel for the dissemination of medical and health information. However, traditional public opinion monitoring and psychological counseling methods such as questionnaires and on-site interviews have problems such as information lag, coverage limitations, and simplistic analysis. Social media medical data has the advantages of real-time performance, extensiveness, and diversity. Combined with big data analysis technology, it can accurately capture public opinion dynamics and insight into public psychology, providing an innovative path for the scientific response to public health emergencies[1]. It is of great significance for improving emergency management theories and safeguarding public mental health.

2. Technical Processing System of Social Media Medical Data

The technical processing of social media medical data is the foundation for subsequent applications. It is necessary to realize the transformation of data from an "unordered state" to "standardized application" through "data collection - preprocessing - public opinion monitoring - psychological counseling". The core technologies focus on the efficient acquisition of multi-modal data and its specific application in the medical field.

2.1. Multi-modal Data Collection Technology

2.1.1. Distributed Crawler and Load Balancing Design

According to the degree of interface openness of different social media platforms, a hybrid collection scheme of "API interface + distributed crawler" is adopted. For platforms with open APIs (such as Weibo and WeChat), text, release time, geographical location, and interaction data published by users are obtained through API keys. The collection efficiency can reach 500-1000 pieces per second, and the legality of the data is guaranteed (in line with platform user agreements).

For platforms with unopened APIs (such as some medical forums and short video comment sections), the Scrapy-Redis distributed crawler architecture is used to capture unstructured data. To solve the problems of node task overload and data duplication, a consistent hashing algorithm is introduced to construct a task allocation mechanism: URL tasks and crawler nodes are mapped to the hash ring respectively, and the node weight is dynamically adjusted according to the node CPU utilization rate (threshold set to 70%). When the load of a node is too high, some tasks are automatically migrated to low-load nodes to achieve uniform task allocation[2]. At the same time, a Bloom filter is used to correct the crawled URLs. The URL is mapped to a binary vector through 3 hash functions, and the memory occupation is reduced by more than 80% compared with the traditional hash table. It supports the correction of millions of URLs per second, effectively avoiding data duplication caused by excessive crawling.

2.1.2. Synchronous Collection and Association of Multi-modal Data

For "text + image + voice" multi-modal data, it is necessary to solve the problems of temporal alignment and content association. A time point synchronization mechanism is adopted to add a time point accurate to milliseconds to each

information release behavior of users (text input, image upload, voice message). A matching threshold of ±500ms is set to classify the multi-modal data released by the same user within this time window as a single information behavior, realizing temporal alignment[3].

For video content and comment data on short video platforms, the inter-frame difference method is used to extract video key frames. The key frames are preprocessed (grayscale and edge detection) through the OpenCV library, and then the text information in the frames is recognized by the Tesseract OCR technology. The Jaccard similarity calculation is performed between the recognition results and the comment area text (threshold set to 0.6). When the similarity meets the standard, the association between video content and user comments is established, realizing multi-dimensional data fusion of "video images - user feedback". For example, during the influenza epidemic in a certain region, this technology successfully associated 300,000 short videos with their corresponding comment data, providing complete multi-modal information for subsequent public opinion analysis[4,5].

2.2. Data Preprocessing and Feature Engineering Technology

2.2.1. Adaptive Word Segmentation Optimization in the Medical Field

Traditional word segmentation tools (such as jieba) have low accuracy in processing medical professional terms. To solve this problem, a medical field word segmentation dictionary is constructed, which includes more than 120,000 professional terms from the *Medical Subject Headings (MeSH)* and *Chinese Subject Headings for Traditional Chinese Medicine*. A dictionary priority matching algorithm is adopted: during the word segmentation process, terms in the dictionary are matched first. If the matching is successful, they are retained as independent words; otherwise, the maximum matching method is used for word segmentation. To further improve the word segmentation accuracy, a Conditional Random Field (CRF) model is introduced to optimize the word segmentation results. Using 5 million medical documents from the PubMed Central database as training corpus, a CRF feature template including the part-of-speech of adjacent words is constructed. Through 10-fold cross-validation, the accuracy of medical term segmentation of this optimization scheme increased from 78.3% of the traditional jieba to 95.6%, laying a foundation for subsequent text feature extraction.

2.2.2. Attention Mechanism for Multi-modal Feature Fusion

The feature fusion of multi-modal data (text + image) is the key to improving the accuracy of subsequent analysis. For text data, the Word2Vec model is used to convert the segmented words into 300-dimensional word vectors, and then the Long Short-Term Memory (LSTM) network is used to extract text semantic features; for image data (such as CT images), the Convolutional Neural Network (CNN, ResNet50) model is used to extract visual features.

To realize the deep fusion of the two types of features, cross-modal attention weight calculation is introduced: assuming the text feature sequence is $T = [t_1, t_2, \ldots, t_n]$ and the image feature sequence is $I = [i_1, i_2, \ldots, i_m]$, the association weight between each text feature $t_j t_j$ and image feature i_k is calculated as $w_{jk} = \frac{\exp(\cos(t_j, i_k))}{\sum_{k=1}^{m} \exp(t_j, i_k)} w_{jk} = \frac{\exp(\cos(t_j, i_k))}{\sum_{k=1}^{m} \exp(t_j, i_k)}$ (where cos( ) is the cosine similarity). After normalizing the weights through the Softmax function, the high-weight image features are weighted and fused with the corresponding text features to obtain the multi-modal feature vector $F = \sum_{j=1}^{n} \sum_{k=1}^{m} w_{jk} (t_j \oplus t_k)$ ($\oplus$ is feature concatenation). Experiments show that compared with simple feature concatenation, this fusion method increases the Silhouette Coefficient of subsequent public opinion topic clustering from 0.52 to 0.78 and the clustering purity by 23.5%, effectively solving the problem of incomplete information in single-modal data[6,7].

3. Application of Social Media Medical Data in Public Opinion Monitoring During Public Health Emergencies

3.1. Real-time Monitoring of Public Opinion Dynamics

By real-time collecting and analyzing social media medical data, we can quickly capture the public's discussion hotspots, viewpoints, and emotional changes regarding public health emergencies. In the data collection work on Facebook and Twitter related to the Ebola epidemic, facing the challenge of upgraded platform anti-crawling mechanisms, the technical team carried out multiple rounds of targeted optimizations and achieved remarkable results. In terms of anti-crawling optimization of the crawler cluster, the account ban rate on Twitter was as high as 40% when using a single-threaded crawler in the initial stage. Subsequently, the technical team adjusted the architecture to "15 slave nodes + 2 threads/node" and combined with an IP pool of over 1000 for dynamic switching (changing 1 IP every 30 seconds), successfully reducing the account ban rate to below 5%. When using Playwright to process dynamic pages, the technical team found that Facebook's "scroll-loaded comments" function required simulating user stay time, so a random interval of 1-2 seconds was set, which effectively solved the problem of only being able to crawl the first 20 comments previously. After optimization, the single crawl volume per node increased from 20 to more than 200. In the data preprocessing link, aiming at a large number of invalid links in Ebola-related texts, the technical team accurately eliminated third-party links by optimizing expressions, significantly improving the filtering accuracy. At the same time, in the fine-tuning of medical word vectors, based on 300-dimensional GloVe pre-trained vectors, 50,000 Ebola epidemic medical texts were used for fine-tuning, and semantic matching was optimized through the Contrastive Loss function. Finally, the vector cosine similarity between "Ebola transmission" and "influenza transmission" was greatly reduced, the filtering rate of irrelevant texts was significantly improved, and the data signal-to-noise ratio was stabilized within a reasonable range[8,9,10]. At the same time, through sentiment analysis, the public's emotional fluctuations were grasped in real time. The proportion of public negative emotions quickly rose from 30% to 55%, and the relevant departments promptly released authoritative information based on this to stabilize public sentiment.

3.2. Predicting the Development Trend of Public Opinion

Based on historical and real-time social media medical data,

machine learning and deep learning algorithms are used to construct a public opinion prediction model. In the prediction of the rumor that "Zika virus causes microcephaly in newborns", the prediction model faced the problem of "redundant transmission features", which was effectively solved by the technical team through a series of optimizations: among the initially constructed 25-dimensional features (including "user followers", "post likes", "comment sentiment tendency", etc.), Pearson correlation analysis found that the correlation between "user followers" and "post forwarding rate" reached 0.85, with serious redundancy. Therefore, "user followers" was eliminated, and 18-dimensional core features were retained. After Z-score standardization of these 18-dimensional features, the number of principal components was determined to be 12 with the help of a scree plot (the cumulative variance contribution rate of the first 12 principal components reached 92%). PCA dimensionality reduction effectively reduced the model complexity, reducing the training time of the LSTM model from 2 hours to 45 minutes. At the same time, aiming at the intraday fluctuations in the "7-day topic discussion volume" (such as a sharp drop in discussion volume at night), a 72-hour sliding window was used for exponentially weighted smoothing to eliminate fluctuation interference. Finally, the smoothness of the discussion volume time series curve was improved by 40%, and the prediction error of the model for the rumor transmission peak was reduced by 15%. Based on this, the relevant departments, in conjunction with social media platforms, marked and refuted the rumors, effectively controlling the spread of rumors. In addition, by analyzing the historical discussion data of the public on prevention and control policies, the model can predict the public's acceptance and reaction after the release of new policies, helping policymakers optimize policy content and release strategies to avoid negative public opinion[10,11].

3.3. Identifying Key Transmission Nodes

In the social media communication network, key transmission nodes (such as well-known bloggers, opinion leaders, and industry experts) have a significant impact on public opinion. Through social network analysis technology, indicators such as degree centrality, betweenness centrality, and closeness centrality of nodes can be calculated to identify key transmission nodes. In the identification of health bloggers during a certain epidemic, the initial use of only the "number of followers" indicator led to misjudgment. Later, the accuracy of identification was improved by optimizing the three-dimensional evaluation system: in terms of structural indicator calculation, when calculating degree centrality, it was found that the "number of followers" included a large number of zombie fans (forwarding rate < 0.1%), so the "proportion of active fans" (defined as the proportion of fans with interactive behaviors in the past 7 days) was introduced. After optimization, the degree centrality score of a certain health blogger increased from 0.7 to 0.9. When calculating betweenness centrality using the Brandes algorithm, the time consumption exceeded 1 hour due to the number of nodes exceeding 100,000. Later, the calculation time was shortened to 10 minutes with no loss of accuracy through the GraphX distributed graph computing framework. In terms of content indicator quantification, when extracting keywords using TextRank, aiming at the problem of low weight of medical terms, medical field word frequency weighting was introduced (medical term weight × 1.5), which increased the weight of keywords such as "epidemic prevention" and "vaccination" in the posts of a certain health blogger by 30%, and the content similarity score increased significantly. At the same time, key nodes spreading rumors are monitored and intervened to cut off the rumor transmission chain and effectively control the direction of public opinion development.

3.4. Evaluating the Effect of Public Opinion Response

After the relevant departments take public opinion response measures, continuous analysis of social media medical data can evaluate the effect of the measures. When responding to an infectious disease epidemic in a certain region, the measure of "regular daily release of epidemic data" was implemented. The technical team verified its effect through technical means from multiple dimensions: in terms of search volume statistics, an "official keyword index database" containing 10 core words such as "confirmed cases" and "cure data" was constructed based on Elasticsearch. A 1-hour collection window was set, and aggregate queries were used to count the search volume. It was found that the search volume increased by 40% on the 3rd day after the implementation of the measure. Further, the paired-samples t-test was used to exclude the impact of random fluctuations, confirming that the increase in search volume was directly related to the measure. In terms of sentiment tendency verification, 100,000 texts were selected using the stratified sampling method by region and user activity. After classification by the DistilBERT model, it was found that the proportion of keywords such as "transparent data" and "timely and accurate" in positive comments reached 65%. With the help of the Chi-square test, the significance of the increase in the proportion of positive comments from 45% to 70% was verified. At the technical level of feedback clustering, initially, when using DBSCAN clustering, the clustering was fragmented due to Eps=0.2. Later, the Silhouette Coefficient was calculated through the elbow method, and the parameters of Eps=0.3 and MinPts=5 were determined. Finally, the feedback was clustered into 5 core types, among which the "questions about the details of prevention and control measures" accounted for 22%, providing a clear direction for the optimization of subsequent work. Based on this, the relevant departments further optimized the content of information release, added a measure interpretation link, and further improved the effect of public opinion management.

4. Application of Social Media Medical Data in Psychological Counseling During Public Health Emergencies

4.1. Insight into Public Psychological States

With the help of text analysis and sentiment analysis technologies, we can deeply insight into the psychological state of the public during public health emergencies. In the insight into the psychological state during the Ebola epidemic, the technical team achieved a breakthrough through multi-technical integration to solve the problems of "difficulty in subdividing negative emotions and inaccurate identification of psychological differences among groups": the initial general sentiment model (such as VADER) could only distinguish "positive/negative/neutral". Later, a "BiLSTM-CRF + medical psychology dictionary" model was

constructed, with 30,000 annotated texts (including labels such as "panic - virus mortality rate") as the training set, and a medical psychology dictionary containing more than 200 exclusive words was embedded. After optimization through the contrastive loss function, the F1-scores for identifying "panic", "anxiety", and "worry" reached 90%, 88%, and 89% respectively, an increase of 35% compared with the general model. Dependency parsing was used to extract the "emotion - cause" pairing relationship, and the weight of cause phrases was calculated by TF-IDF. It was found that the weights of "high virus mortality rate" and "limited treatment methods" reached 0.82 and 0.78, with an attribution accuracy of 91%. In the identification of special groups, medical staff occupation keywords were extracted through BERT-NER and combined with behavioral characteristics to construct a mapping model[12,13]. It was found that the frequency of keywords such as "fatigue" and "fear of infection" among medical staff was 3.2 times that of the general public, and the proportion of occupational burnout was 45%, with an identification accuracy of 87%. These analysis results provide an accurate basis for formulating personalized psychological counseling plans.

4.2. Formulating Targeted Psychological Counseling Strategies

Based on the analysis results of the public's psychological state, combined with the characteristics and needs of different groups, psychological counseling strategies are formulated. For the general public with severe panic, social media platforms, in conjunction with psychological experts, have launched a series of popular science short videos to explain the principle of virus transmission, prevention measures, and scientific treatment methods, alleviating the public's unknown fear of the virus. An online relaxation training mini-program has been developed to provide training courses such as deep breathing, meditation, and progressive muscle relaxation to help the public relieve anxiety. For medical staff, an exclusive psychological support community has been established, inviting psychologists to regularly carry out online lectures and one-on-one consulting services. A "Medical Staff Voices" sharing section has been set up to encourage medical staff to communicate with each other and release pressure. During the Zika virus epidemic, for pregnant women, obstetrics and gynecology experts and psychological experts were organized to carry out online Q&A activities to answer questions such as pregnancy protection and fetal health checks, effectively alleviating their anxiety.

4.3. Evaluating the Effect of Psychological Counseling

After the implementation of psychological counseling measures, the effect is evaluated by continuously monitoring and analyzing social media medical data. In the evaluation of the effect of psychological counseling during a certain public health emergency, the technical team constructed a technical system of "data collection - statistical verification - trend analysis": real-time collection of social media texts based on the Flink stream processing engine, sampling 100,000 texts daily, and using the optimized DistilBERT model with INT8 quantization for emotion classification. Comparing before and after the implementation of counseling measures, the proportion of negative emotions decreased from 60% to 30%, and the proportion of positive emotions increased from 20% to 45%. Subsequently, the paired-samples t-test (sample size 5000) verified that the mean difference of negative emotions was 30%, t-value = 5.82 ($P < 0.01$), confirming that the emotional improvement was statistically significant. At the same time, linear regression analysis was used to verify the slope of emotional changes (positive emotion slope 0.08, negative emotion slope -0.05), confirming that the counseling measures were continuously effective. In the qualitative analysis of user feedback, the TextRank algorithm with a window size of 5 and a damping coefficient of 0.85 was used to extract keywords such as "practical", "anxiety relief", and "insufficient personalization", with a coverage rate of 90%. Then, the DBSCAN algorithm with Eps=0.3 and MinPts=5 was used to cluster the feedback into categories such as "satisfied with training effect" and "hoping for personalized counseling", among which "personalized needs" accounted for 22%, providing a clear direction for subsequent strategy adjustments. At the same time, cluster analysis is used to classify public feedback. In response to the demand of some users for "increasing personalized psychological counseling", the relevant departments promptly adjusted their strategies and launched personalized counseling plans based on the evaluation of users' psychological states, further improving the quality and effect of psychological counseling and effectively safeguarding public mental health[14,15].

5. Conclusion

With the characteristics of real-time performance and extensiveness, social media medical data provides an innovative solution for public opinion monitoring and psychological counseling during public health emergencies, and shows significant value in insight into public dynamics and optimizing emergency strategies. However, problems such as uneven data quality, privacy and security risks, and technical bottlenecks still restrict its in-depth application. In the future, it is necessary to break through application obstacles by improving the data governance technology system, strengthening technological innovation, and promoting multi-stakeholder collaborative cooperation, so as to fully release the potential of social media medical data, help build a more efficient and intelligent public health emergency management system, and better safeguard public health and social stability.

References

[1] Kim, E. S., James, P., Zevon, E. S., Trudel-Fitzgerald, C., Kubzansky, L. D., & Grodstein, F. (2020). Social media as an emerging data resource for epidemiologic research: Characteristics of regular and nonregular social media users in Nurses' Health Study II. American Journal of Epidemiology, 189(2), 156 - 161. doi:10.1093/aje/kwz224

[2] Garg, A., & Gupta, S. (2024). Leveraging Social Media Data for Healthcare Analytics: Opportunities and Risks. MoldStud, 1(1), 1 - 10.

[3] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M.,... van Ginneken, B. (2017). Deep learning for medical image analysis: A review. Medical Image Analysis, 42, 60 - 88.

[4] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2019). Machine learning in medicine. Nature Medicine, 25, 24 - 36.

[5] Aljarah, I., Yaseen, Z. M., & Elaziz, M. A. (2020). A review of data mining techniques for healthcare data. Journal of Healthcare Engineering, 2020, Article ID 8828590.

[6] Perner, P. (2021). Medical data mining and knowledge discovery: A review. Pattern Recognition Letters, 145, 2 - 10.

[7] Alqahtani, A., & Maglogiannis, I. (2023). Leveraging social media data for COVID-19 pandemic monitoring and prediction. Journal of Healthcare Engineering, 2023, Article ID 7684701.

[8] Khan, S. A., & Alharthi, A. A. (2022). Social media analytics for public health emergency response: A review. Journal of Medical Systems, 46(8), 106.

[9] Smith, J., & Johnson, A. (2023). Social Media - Based Interventions for Mental Health During Public Health Emergencies: A Review. Journal of Mental Health and Crisis Intervention, 8(2), 112 - 125.

[10] Wang, Y., & Li, Z. (2022). Social Media as a Tool for Mental Health Promotion and Crisis Intervention in Public Health Emergencies. Asia - Pacific Journal of Public Health, 34(5), 456 - 465.

[11] Crea, Thomas M., et al. "Social distancing, community stigma, and implications for psychological distress in the aftermath of Ebola virus disease." Plos one 17.11 (2022): e0276790.

[12] Cénat, Jude Mary, et al. "Psychological distress among adults from the urban and rural areas affected by the Ebola virus disease in the Democratic Republic of the Congo." Social psychiatry and psychiatric epidemiology 56.1 (2021): 57-62.

[13] James, Peter Bai, et al. "Health-related quality of life among Ebola survivors in Sierra Leone: the role of socio-demographic, health-related and psycho-social factors." Health and Quality of Life Outcomes 20.1 (2022): 10.

[14] Arafat, Md Yeasin, Sanjana Zaman, and Mohammad Delwer Hossain Hawlader. "Telemedicine improves mental health in COVID-19 pandemic." Journal of global health 11 (2021): 03004.

[15] Vivalya, Bives Mutume Nzanzu, et al. "Develo** mental health services during and in the aftermath of the Ebola virus disease outbreak in armed conflict settings: a sco** review." Globalization and Health 18.1 (2022): 71.