



Semnan University



The Role of Social Networks in Predicting the Coronavirus Infection Rate

Fahimeh Hafezi¹, Maryam Khodabakhsh^{1,*}

¹ Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran.

*Corresponding author: m_khodabakhsh@shahroodut.ac.ir

Received: 2024-07-08 Revised: 2025-06-09 Accepted: 2025-09-23

Abstract

The coronavirus pandemic highlighted the potential of social media as a valuable source of information for monitoring public health trends. User-generated content can provide insights into public reactions to the pandemic and help estimate changes in infection rates over time. This study proposes a model for estimating coronavirus incidence using coronavirus-related social media posts. Specifically, it investigates whether online content can serve as an indicator of infection rates and whether incorporating emotional features related to fear and hope can improve estimation performance. Semantic similarity is calculated using three conventional word-embedding techniques, Word2Vec, GloVe, and FastText, as well as Sentence-BERT (SBERT), which generates sentence-level semantic representations. Experimental results indicate that combining SBERT-based semantic similarity with fear achieved the best performance, with an R^2 value of 0.52. This outperformed the corresponding Word2Vec-based combination ($R^2 = 0.47$) and the other evaluated feature combinations. Overall, the findings highlight the potential of combining transformer-based semantic representations with emotion-related features to estimate coronavirus incidence from social media data.

Keywords: Coronavirus, Social media, Incidence rate estimation, Synergy, sentiment analysis.

1. Introduction

Social media platforms have become valuable sources of user-generated information that can complement conventional approaches to monitoring infectious diseases. During the coronavirus pandemic, users shared information, personal experiences, concerns, and emotional reactions through online platforms. Analyzing this content may help identify changes in public attention and provide useful indicators for estimating the progression of an epidemic. Previous studies have investigated the relationship between social media activity, online search behavior, and reported COVID-19 cases [1,2]. However, approaches based primarily on predefined keywords may retrieve irrelevant posts while overlooking relevant content that does not contain the selected terms.

Semantic similarity provides an alternative means of identifying relevant content by measuring the semantic relationship between a post and a representation of the target concept. In addition to textual relevance, emotional expressions may provide complementary information about how users respond to an epidemic. Fear and hope are particularly relevant in this context because they can reflect different emotional responses to uncertainty, perceived risk, and expectations about the future. Previous research has explored semantic similarity and emotional features for estimating coronavirus incidence from social media data [3].

This study investigates whether combining semantic similarity with fear- and hope-related features improves the estimation of daily new coronavirus cases in Iran. Three feature configurations are evaluated: semantic similarity combined with fear, hope, or both emotions. Semantic similarity is calculated using Word2Vec, GloVe, and FastText embeddings, as well as the transformer-based Sentence-BERT (SBERT) model. The results are compared with baseline approaches to assess whether SBERT improves estimation performance and whether emotional features complement semantic information. The study provides a comparative evaluation of these configurations to identify the combination that achieves the strongest estimation performance.

2. Methodology

2.1. Dataset and preprocessing

The experiments use Persian-language Instagram posts collected during the first wave of the coronavirus pandemic in Iran. The first dataset contains 4,919,839 Persian-language posts collected between January 21 and April 21, 2020. A manually annotated dataset containing 2,067 coronavirus-related posts labeled as fear or no fear was used to support the identification of fear-related content. A separate labeled Persian-language dataset containing hopeful and negative

examples was also used to develop the hope classifier. The number of examples in this second dataset is not reported here. For the incidence-estimation experiments, coronavirus-related posts containing the keyword “corona” were collected from February 20 to June 20, 2020. The posts were preprocessed before feature extraction. Daily official counts of newly reported coronavirus cases in Iran were obtained from Iran Open Data and used as the target values for the estimation model. The study therefore examines whether daily indicators extracted from social media are associated with the observed daily case counts.

2.2 Semantic similarity and emotional features

The proposed approach represents coronavirus-related content through semantic similarity and emotion-related information. Semantic similarity is calculated using word representations obtained from Word2Vec, GloVe, and FastText [4–6]. These embedding techniques represent words in a vector space, enabling the semantic relationship between a post and the coronavirus concept to be measured using cosine similarity. In addition to these conventional embedding techniques, Sentence-BERT (SBERT) is used to generate sentence-level semantic representations and calculate semantic similarity between posts and the coronavirus concept. This enables a comparison between conventional word embeddings and a transformer-based approach to semantic representation.

Fear and hope are incorporated as additional features. For each emotion, a binary classifier based on Bidirectional Encoder Representations from Transformers (BERT) [7] is trained to identify whether the corresponding emotion is expressed in a post. The annotated examples provide the supervision needed to train these classifiers. The resulting emotion labels are then used alongside semantic similarity to characterize the daily social media content. Three feature combinations are investigated: (1) semantic similarity and fear, (2) semantic similarity and hope, and (3) semantic similarity, fear, and hope. These combinations are evaluated using different semantic representation methods to assess both the contribution of emotional features and the effect of the semantic similarity model on estimation performance. An overview of the proposed framework is shown in **Figure 1**.

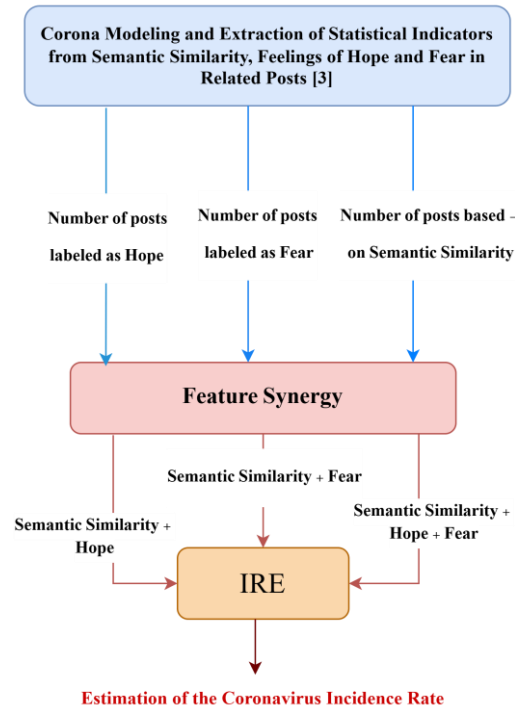


Figure 1. Overview of the proposed framework, including data preprocessing, semantic similarity calculation, fear and hope classification, daily feature aggregation, and incidence estimation

2.3 Incidence estimation and evaluation

For each day, the extracted post-level features are aggregated into daily statistical indicators. These indicators include the number of posts identified as semantically related to coronavirus, the number expressing fear, and the number expressing hope. The resulting daily features are used as input to a Support Vector Regression (SVR) model, which estimates the number of new coronavirus cases. SVR is a supervised learning method suitable for regression tasks and has been applied to problems involving nonlinear relationships [8]. The model is evaluated using three-fold cross-validation. Performance is measured using the coefficient of determination R^2 , which indicates the proportion of variance

in the target values explained by the model. Higher R^2 values indicate better agreement between estimated and observed case counts within the evaluated data. The proposed combinations are compared with one another and with baseline approaches reported in previous work.

3. Results and Discussion

3.1. Effect of semantic similarity and fear

The first experiment examines the combination of semantic similarity and the fear feature. As shown in Table 1, Word2Vec achieves the highest R^2 value (0.47), followed by FastText (0.44) and GloVe (0.30). When SBERT is used to calculate semantic similarity, the combination with fear achieves an R^2 value of 0.52, representing the best performance among the evaluated configurations. These findings suggest that fear-related information may complement semantic similarity in identifying posts relevant to the outbreak. The higher R^2 obtained with SBERT compared with Word2Vec (0.52 versus 0.47) also suggests that the choice of semantic representation may affect estimation performance. However, these results demonstrate an empirical association within the evaluated dataset and do not establish a causal relationship between online fear and infection rates.

3.2. Effect of semantic similarity and hope

The second experiment combines semantic similarity with the hope feature. Word2Vec achieves an R^2 value of 0.42, while FastText and GloVe obtain 0.36 and 0.26, respectively. With SBERT, the semantic-similarity-and-hope combination achieves an R^2 value of 0.43. Although this value is slightly higher than the Word2Vec result of 0.42, it remains below the corresponding SBERT-and-fear result of 0.52. Within the evaluated configurations, these results suggest that fear provides a more useful complementary feature than hope for estimating daily new cases. This finding should be interpreted in the context of the selected dataset, time period, features, and evaluation procedure.

3.3. Combining semantic similarity, fear, and hope

The third experiment integrates all three features. The resulting R^2 values are 0.42 for Word2Vec, 0.24 for GloVe, and 0.37 for FastText. For SBERT, the combination of semantic similarity, fear, and hope achieves an R^2 value of 0.50. Although this result is higher than the corresponding Word2Vec result of 0.42, it does not exceed the SBERT-and-fear configuration, which achieves 0.52. These findings indicate that adding more features does not necessarily improve estimation performance. In this study, incorporating hope alongside fear does not provide an additional improvement over the SBERT-and-fear combination. Across the evaluated SBERT configurations, the combination with fear alone yields the highest R^2 value.

Table 1. Performance of the proposed feature combinations using different word-embedding techniques

Feature combination	Word2Vec R^2	GloVe R^2	FastText R^2	SBERT R^2
Semantic similarity + fear	0.47	0.30	0.44	0.52
Semantic similarity + hope	0.42	0.26	0.36	0.43
Semantic similarity + fear + hope	0.42	0.24	0.37	0.50

3.4. Comparison with baseline approaches

The proposed approach is compared with a baseline based on Twitter data and an approach using semantic similarity alone. The Twitter-based reference method [1] reports an R^2 value of 0.27, while the semantic-similarity-only approach reported in [3] achieves an R^2 value of 0.32. Building on the semantic-similarity approach in [3], this study incorporates emotion-related features to evaluate whether emotional information can improve estimation performance. The proposed combination of Word2Vec-based semantic similarity and fear achieves an R^2 value of 0.47, while the corresponding SBERT-based combination achieves 0.52, the highest value among the approaches compared. Compared with semantic similarity alone, these results suggest that incorporating fear-related information may provide additional predictive value. Furthermore, the higher R^2 obtained with SBERT than with Word2Vec suggests that sentence-level semantic representations may improve estimation performance within the evaluated experimental setup. These findings should be interpreted within the scope of the available data and evaluation procedure. The results are reported in Table 2.

Table 2. Comparison of the proposed approach with baseline methods

	Method	R^2
baseline	Twitter-based [1]	0.27
	Semantic similarity alone baseline [3]	0.32
Ours	Semantic similarity (W2v) + fear	0.47
	Semantic similarity (SBERT) + fear	0.52

4. Conclusions

This study investigated whether combining semantic similarity with fear- and hope-related features could improve the estimation of daily new coronavirus cases using Persian-language Instagram posts from Iran. Three feature combinations were evaluated using Word2Vec, GloVe, and FastText embeddings, with support vector regression (SVR) as the estimation model. In addition, Sentence-BERT (SBERT) was evaluated as a transformer-based method for calculating semantic similarity. The results showed that combining semantic similarity with the fear feature yielded the best performance, achieving an R^2 value of 0.52 with SBERT, compared with 0.47 using Word2Vec. The SBERT-based combinations with hope and with both fear and hope achieved R^2 values of 0.43 and 0.50, respectively. These results suggest that fear-related information may complement semantic representations when extracting signals relevant to epidemic monitoring from social media. They also indicate that incorporating additional features does not necessarily improve estimation performance, as the combination of fear and hope did not outperform the fear-based combination. The results further suggest that the choice of semantic representation is relevant to the performance of the estimation model. The proposed approach may serve as a supplementary tool for monitoring epidemic trends using online data. However, social media-based estimates should not be considered a substitute for official epidemiological data. Future research should evaluate the approach across additional platforms, languages, geographical settings, and longer observation periods. Further investigation of alternative feature aggregation strategies and more extensive validation procedures is also needed to assess the robustness, reliability, and generalizability of the proposed approach.

5. References

- [1] Gharavi, Ehsan, Nazemi, N., and Dadgostari, F. "Early Outbreak Detection for Proactive Crisis Management Using Twitter Data: COVID-19 a Case Study in the US." arXiv preprint arXiv:2005.00475 (2020).
- [2] Ayyoubzadeh, Seyed Mohammad, Seyed Mehdi Ayyoubzadeh, Hoda Zahedi, Mahnaz Ahmadi, and Sharareh R. Niakan Kalhori. "Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study." JMIR Public Health and Surveillance 6, no. 2 (2020): e18828.
- [3] Hafezi, Fahimeh, and Maryam Khodabakhsh. "Coronavirus incidence rate estimation from social media data in Iran." Journal of AI and Data Mining 11, no. 2 (2023): 315-329.
- [4] Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. "Efficient Estimation of Word Representations in Vector Space." Arxiv Preprint Arxiv:1301.3781 (2013).
- [5] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543. 2014.
- [6] Bojanowski, Piotr, Grave, Edouard, Joulin, Armand, and Mikolov, Tomas. "Enriching Word Vectors with Subword Information." Transactions of the Association for Computational Linguistics 5 (2017): 135-146.
- [7] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." In Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers), pp. 4171-4186. 2019.
- [8] Zhang, Fan, and Lauren J. O'Donnell. "Support vector regression." In Machine learning, pp. 123-140. Academic Press, 2020.