Evaluating Sparse and Transformer-based Representations for Chinese Weibo Sentiment Analysis Across Data Scales and Noise Conditions

Yuying Zhao

Institute of Education, University College London, London, WC1E 6BT, United Kingdom
Yuying.zhao.21@ucl.ac.uk

Abstract. This study presents a systematic comparison between traditional sparse representations and contextualised Transformer models for Chinese Weibo sentiment classification. Using a publicly available dataset of 10,500 annotated microblog posts, the analysis examines four key dimensions: text representation, data scale, noise robustness, and fine-tuning strategy. A character-level TF-IDF + Logistic Regression baseline is evaluated alongside a pretrained BERT model under controlled experimental conditions. Results show that BERT substantially outperforms TF-IDF when trained on the full dataset, achieving higher accuracy and macro-F1 through its ability to capture contextual and semantic nuances in noisy social-media text. However, in low-resource settings with only 1,000 training samples, TF-IDF remains competitive, narrowing the performance gap and demonstrating strong efficiency under data scarcity. Noise-robustness experiments further reveal that BERT maintains stable or improved performance under mild perturbations, while TF-IDF exhibits gradual degradation. Fine-tuning analysis confirms that full parameter updates are essential for BERT's effectiveness, as freezing the encoder leads to significant performance declines. Overall, the findings provide a reproducible benchmark and practical guidance for selecting sentiment-analysis models under varying resource constraints, highlighting trade-offs between expressive power, robustness, and computational cost.

Keywords: Natural Language Processing, Sentiment Analysis, Chinese Weibo, TF-IDF, BERT.

1 Introduction

In a backdrop of the swift social media evolution, social media sites like Weibo have become the important areas in the expression of feelings, opinion exchange, and information sharing. Weibo text sentiment analysis is vital in monitoring the opinion of the populace, social risk, business decision support, user profiling, and policy analysis. According to Wang and Alfred as summarised, when compared to news reports or other types of formal text (i.e., written sources), content of microblogs is often characterised by the extreme brevity of the messages, a high level of colloquialism, and a lot of noise; users often make heavy use of emoticons, internet slang, spelling mistakes, mixes of

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_99

pinyin and characters and various non-standard sentence structures [1]. This is a very unstructured lingo that poses other challenges to automated sentiment classification tasks.

The current text representation approaches can be broadly-classified as two. The former includes the classical form of sparse representations including TF-IDF, n-grams, and bag-of-words, usually combined with a linear or non-linear classification system (SVM or Logistic Regression). Those approaches have easy implementation and are highly computationally efficient and can be of value practically in cases of limited training data or computation limitations. The second group is those contextualised semantic representation techniques that are built on top of pre-trained language models, including BERT and ERNIE. These models are pre-trained on large scale unsupervised pre-training, achieving substantial performance improvements in a variety of Chinese NLP tasks, and they are commonly used in the sentiment analysis of the Chinese Weibo.

Even though the studies currently available indicate that traditional and Transformer-based models can be successfully used to reach effective results in the field of microblog sentiment recognition [2, 3], multiple gaps in research exist. First, a significant portion of the literature is dedicated to the suggestion of new deep architectures or a comparison of neural models in particular fields but few studies offering systematic comparative analysis between classical sparse representations and pre-trained Transformers in the same conditions with task or dataset constraints which is implicitly recognized in earlier literature [2, 3]. Second, although a lot of literature has been focused on determining the best model, there has been comparatively limited focus on the effect of variation in training data scale on relative performance of the various methods despite the fact that model performance is also often exceedingly sensitive to the scale of training data [4, 5]. Third, there is limited systematic empirical evidence in the comparison of the strength of various representations in the noisy conditions or the analysis of the impact of various fine-tuning strategies under12 on the stability of models and downstream performance; the gaps that have been briefly accepted in the research on noisy user-generated content [6, 7].

In response to these research gaps, the current study provides a systematic comparison of the character-level TF-IDF + Logistic Regression to that of pre-trained BERT models on a publicly obtained dataset on Chinese Weibo binary sentiment classification. The analysis concentrates on the three dimensions of text representation methods, data scale and how effective it is to noise. The two are compared under the same task and data partitioning factors to help develop a benchmark of the relative advantage of the two. At the same time, this paper looks at how different size of training sets (between 1k and entire dataset) affect the performance of the models. It generates synthetic noise in order to evaluate model behaviour in a setting of inadequate supervision and poor quality inputs and hence shows under which conditions Transformer models continue to benefit their performance. Lastly, comparative studies of fine-tuning strategies such as full-parameter fine-tuning and frozen encoder training, measure the role of task-specific gradient updates on model performance to provide usable advice in trading off between performance and resource cost in resource-constrained setting.

As it was analyzed above, the current study offers a data-focused and systematic analysis of the comparative performance of Transformer models and TF-IDF methods in terms of the Chinese Weibo sentiment classification with data of different size and noise levels. It sets a standard and methodological precedent in future studies.

2 Related Works

Text representation is an important part of natural language processing; its goal is to transform raw textual data into numerical vectors. Various forms of representation reflect different amounts of linguistic information, including frequency-based sparse features, distributed semantic embeddings, and more recently contextualised representations based on large pretrained models.

Early research primarily relied on traditional sparse representations such as bag-of-words, n-grams, and TF-IDF. These methods construct feature vectors by counting the occurrence frequency of characters or words, offering strong interpretability and low computational cost. However, they are limited in their ability to capture word order, contextual dependencies, and the non-standard expressions commonly found in Weibo posts, such as irony, metaphor and informal phrasing; consequently, their effectiveness is constrained in more complex sentiment-classification tasks.

With advances in deep learning, distributed word embeddings and more sophisticated neural architectures were introduced into Weibo sentiment analysis. Ling et al. proposed a hybrid neural network that integrates ELMo-based contextual embeddings with corpus-level co-occurrence statistics, using multi-channel CNNs to extract local features and a Bi-LSTM to model sentence-level semantics [8]. Their experiments demonstrated consistent improvements in precision, recall and F1-score over earlier CNN- and LSTM-based baselines, indicating that contextual deep representations were advantageous even prior to the emergence of Transformer models.

Survey work further highlights that sentiment analysis on Weibo requires extensive preprocessing due to its highly colloquial and noisy linguistic style [1]. Lexicon-based approaches can handle negation and degree adverbs but exhibit limited generalisation. Enhanced machine-learning classifiers, such as AdaBoost-SVM combinations, have shown promising results in some contexts, yet often lack systematic validation on genuine Weibo corpora. Continued developments in deep learning improved attention mechanisms and convolutional architectures, enabling models to better capture semantic nuances in short, highly variable texts.

More recently, pretrained language models have produced substantial performance gains on Weibo sentiment tasks. Li et al. proposed a hybrid model that uses BERT to generate contextualised embeddings, augmented with sentiment-lexicon features, BiLSTM layers for long-range dependencies, attention mechanisms for weighting salient emotional cues, and CNN layers for local feature extraction. Their experiments on COVID-19-related Weibo data show that the BERT-based model markedly outperforms traditional machine-learning and earlier deep-learning baselines [2]. These studies collectively demonstrate the strong advantages of Transformer-based models in

handling noisy short text, providing clear motivation for comparing TF-IDF with BERT in this work.

Overall, text representation methods have evolved from simple statistical features to deeply contextualised semantic models, with each approach exhibiting distinct differences in computational cost, data requirements and robustness. These differences form the conceptual foundation for this study's systematic comparison between TF-IDF and BERT [9,10].

3 Methodology and Experimental Setup

3.1 Dataset

The corpus utilised in this study originates from the Chinese Weibo Sentiment Classification Dataset, publicly released by GitCode user bond007 on 11 October 2024 under the MIT open-source licence. This data is randomly gathered via multi-thematical corpora on the open domain of the Chinese social network Weibo. The dataset contains 10500 sentiment-annotated things on Weibo, split into a training (10000 of those) and a test set (500 things on Weibo). The entries will be represented as MID (unique identifier of a piece of the Weibo text), the text on Weibo, and a binary sentiment indicator (1 - positive, 0 - negative). The positive sentiment training set is the largest (54.97 percent) and the negative sentiment training set is the smaller (45.03 percent).

The dataset is quite compatible with the goals of this study: First, the given binary sentiment labels could provide definite and consistent supervised indicators of model performance, allowing to compare the different text representation strategies in sentiment classification problems systematically. Secondly, the dataset maintains all the inherent noises of Weibo text, such as emoticons, internet jargon, non-standard spelling, mixed pinyin, hyperlink, hashtags, hyperlinks, media placeholders. These highly unstructured linguistic forms authentically reflect the complexity of social media text, thereby providing ideal conditions for testing the robustness of sparse representations and Transformer-based representations in noisy environments.

3.2 Feature Representations

Two representative categories of text representation are examined in this study.

The first category is traditional sparse representations, operationalised through character-level TF-IDF (Term Frequency – Inverse Document Frequency). Given the highly colloquial and noisy nature of Weibo posts, word segmentation tools tend to introduce additional errors; therefore, character-level n-grams (1 – 2 characters) are used. To control the dimensionality of the feature space, min_df is set to 2 and max_features is capped at 50,000. Logistic Regression serves as the classifier on top of TF-IDF vectors, providing an interpretable and computationally efficient baseline.

The second category is contextual semantic representations based on Transformer architectures. This study adopts bert-base-chinese as the primary pretrained model. Owing to its large-scale unsupervised pretraining, BERT captures semantic dependencies and contextual cues that are essential for understanding short and noisy

Chinese social-media text. All experiments use the same tokenizer, with the maximum sequence length fixed at 128. A linear classification layer is applied on top of BERT to perform binary sentiment prediction.

3.3 Model Training

Full fine-training is the default training strategy to use in Transformer-based models. The training process involves updating all the parameters to suit the sentiment patterns that occur in Web set Weibo posts. The training is performed within a gpu. The important hyperparameters are a learning rate value of 2e-5, random seed, 42, a batch size of 8 and a training epoch of 32 that is used to evaluate. Loss curve falls continuously with level adjustments of 0.588 to almost 0.291 as a sign of consistent convergence.

To further measure the contribution of the task-specific adaptation, one more frozen-encoder setup will be used, where the Transformer layers are frozen, and only the classification head is learned. Such a design enables the assessment of the dependence on the addition of performance on the update of the pretrained representations. In the case of the traditional model, TF-IDF vectors are taught by means of Logistic Regression by using the liblinear solver on CPU. The data splits are controlled because all models are trained and assessed on the same slices illustrated experimental controlled conditions.

3.4 Experiment design

This study employs a systematic set of controlled experiments across four dimensions: representation method, data scale, noise perturbation and fine-tuning strategy.

First, a full-data experiment is conducted. Both the TF-IDF baseline and the fully fine-tuned BERT model are trained on the entire 10,000-sample training set to compare their performance under ample supervision.

Second, a small-sample experiment is included. A subset of 1,000 training samples is randomly drawn to examine model effectiveness under limited labelled resources and to compare the data efficiency of sparse vs. contextual representations.

Third, a noise robustness experiment is developed. In order to mimic text noise on social-media, the 10% of the training data is contaminated using lightweight synthetic noise in the form of emojis, repeated characters and conversational fillers. This experiment evaluates the sensitivity of each of the models to distortion at the surface.

Last, the comparison of a fine-tuning strategy is done in the BERT framework. The results of full fine-tuning, frozen-encoder settings are contrasted to assess the contribution of task-specific gradient updates as well as disentangle the nature between pretrained representations and task adaptation.

4 Results and Discussions

4.1 Baseline Performance Comparison

The TF-IDF + Logistic Regression model as a traditional baseline attains its development accuracy of 0.68 and the test accuracy of 0.716. The classification report demonstrates that there is an evident unequal recall between classes: the model has a sufficiently high recall on the negative posts (dev recall = 0.8489; test recall = 0.8581), however, the recall on positive samples is significantly smaller (dev recall = 0.5418). This is an indication of the fragility of linear models to non-homogeneous linguistic structures and their inability to utilize subtle sentiment clues.

Comparatively, BERT produces notably gooder results when made to run under the same conditions of data and attains a test accuracy of 0.916 and the macro-F1 score of 0.901. In contrast to linear decision bounds, BERT has the advantage of representing contextualisation, which features multi-level semantics and implicit sentiment signals. The confusion matrix indicates that BERT makes high consistency and high consistency between negative (133/155) and positive (325/345) cases, which means closer generalisation.

This comparison at the base depicts the underlying disparity between sparse representations of n-grams and contextual Transformer encoders: the first tend to be based on surface patterns, and the latter have the potential to represent more substantial semantic structures of posts in Weibo.

4.2 Effect of Data Size

Both the models were retrained with 1,000 labelled samples only to evaluate the efficiency of the data. On this low-resource environment, TF-IDF reliably achieves a precision of 0.844 and a macro-F1 of 0.816, whereas BERT scores 0.864 accuracy and a macro-F1 of 0.844. The performance difference is also much smaller compared to the full data condition.

With small data TF-IDF is competitive due to its low parameter space and the simplicity of its frequency signals. BERT maintains a lead although at a diminished margin which implies that Transformer models need moderate supervision in order to realize their potential. As the size of the training increases, BERT will show an evident upward trend whereas TF-IDF will reach the limits of representation. These findings are consistent with the current literature: pretrained Transformers can effectively be scaled by more data, but sparse models do not scale due to the limited expressive capacity.

4.3 Effect of Noise

In order to estimate the noisy language conditions of social media, 10 percent of the training data were perturbed by minor perturbations, e.g., the added emojis, repeat words and filler phrases. In this case TF-IDF model shows a minimal deterioration in its performance: the accuracy drops to 0.842 and the macro-F1 of 0.8123. This can be

expected, given the fact that when there are sparse representations of n-grams, such representations of frequency are very sensitive to surface frequencies.

On the contrary, BERT showcases a small, yet, steady growth with the accuracy increasing to 0.870, and the macro-F1 also growing to 0.852. It is worth noting that the recall of the negative ones is raised considerably (0.8194 to 0.8452). This strength is explained by the fact that Transformer encoders are contextual since they rely less on single characters and more on bigger semantic patterns. Moreover, the light perturbations can also serve as a regularising system in the low-resource context, where overfitting is minimised, and the pretraining corpus of BERT itself contains vast volumes of informal and noisy text-related information, which makes it more resistant. Combining these two results we can state that Transformers will not degrade over time - and can even become better - when there is relatively mild noise, but TF-IDF is progressively becoming worse.

4.4 Effect of Fine-tuning Strategy

Comparison of full fine-tuning and frozen-encoder setting can give ideas of the significance of task-specific adaptation to Transformer-based models. In the frozen encoder, overall performance has a significant decrease in values as accuracy drops to 0.684 and macro-F1 to 0.556. The negative samples are of particular concern with a recall of 0.2194, although the model still has an intermediate level of the recall on positive samples (308/345). These findings indicate that pretrained representations are inadequate with no additional modifications to the representations to learn informal, vague, and context-specific sentiments that are mainly common in Weibo text.

However, making the gradients go through every Transformer layer as in full fine-tuning brings significant benefits in all metrics. This shows that to be able to classify sentiments effectively in short, noisy text facilitated by social-media, you need to adapt beyond the head of the classifier because much of the emotional signal is contained in subtle contextual patterns that must be reflected in the encoder internal representations. Practically speaking, the results suggest that the encoder freeze can be only appropriate due to strict computational constraints, and even in this case, the decrease of accuracy must be considered.

4.5 Discussion of results

The experimental findings indicate that there are a number of uniform patterns about behaviour of the sparse and contextualised representations in Weibo sentiment classification. In all the environments, opting to use BERT with full fine-tuning has the best results with a top accuracy of 0.916 and a macro-F1 of 0.9015 on the entire training set. This establishes that contextualised representations provide a clear upper limit in case ample data and computing power is present. By contrast, TF-IDF baseline shows significantly lower accuracy when used with the full dataset (accuracy 0.716) indicating that surface-level n-gram features are biased towards less informal sentiment expression and less semantic variation as required by a social-media context. The behaviours of all models in each of the experimental conditions have been summarised

in table 1 with the differences in the behaviours of sparse and contextualised models illustrated.

Table 1. Summary of Model Performance Across All Experiments

Experiment Category	Experiment/Condition	Model	Accuracy	Macro-F1
Baseline	Full training set	TF-IDF + LR	0.716	-
	Full training set	BERT (full fine-tuning)	0.916	0,9015
Small-sample (1,000 instances)	No noise	TF-IDF + LR	0.844	0.8157
	No noise	BERT (full fine-tuning)	0.864	0.8443
Noise robustness (10% perturbation)	Small-sample + noise	TF-IDF + LR	0.842	0.8123
	Small-sample + noise	BERT (full FT)	0.870	0.8523
Fine-tuning strategy	Frozen encoder	BERT (frozen)	0.684	0.5561
	Full fine-tuning	BERT (full FT)	0.916	0.9015
Cross-validation (5-fold)	Full dataset	TF-IDF + LR	0.8313 (±0.0060)	0.8290 (±0.0063)
	Full dataset	BERT (full FT)	0.8678 (±0.0117)	0.8665 (±0.0118)
Data-scale experiments	Train size = 1,000	TF-IDF + LR	0.8460	0.8069
	Train size = 2,000	TF-IDF + LR	0.8560	0.8266
	Train size = 5,000	TF-IDF + LR	0.8720	0.8470
	Train size = 9,000	TF-IDF + LR	0.8800	0.8566
	Full BERT training	BERT (full FT)	0.9200	0.9051

The discrepancy between performance is reduced significantly in low-resource situations. TF-IDF has an accuracy of 0.844 and macro-F1 of 0.8157 when using only 1,000 training instances, which is competitive against Berts 0.864/0.8443. It implies that the linear models are useful in the cases of data scarcity and can be seen as the useful baselines in small samples, where the representational density of Transformers is not capable of being fully capitalized on. The experiments of data-scale further

demonstrate that TF-IDF gains more and more supervision gradually and achieves 0.880 accuracy when trained with 9,000 samples, which is two-thirds of BERT. These findings suggest that the difference between sparse and contextual models is not homogenous but conditional on the quantity of labelled data.

Experiments on noise robustness place a stronger emphasis on the visible distinctions between the two methods. Introducing 10 percent mild perturbations led to a minor decrease in accuracy and macro-F1 on TF-IDF, which agrees with the fact that it relies on the distribution of surface tokens. BERT, however, had a moderate but statistically significant growth with macro-F1 changing by 0.8443 to 0.8523. This phenomenon can probably be explained by the fact that Transformer encoders are contextual and the text in the pretraining corpora is often informal or noisy. The findings indicate that almost simple perturbations could be a normalising signal in deep models, and sparse representations are prone to superficial distortion.

The experiments of fine-tuning prove that the work of BERT heavily requires task-specific adaptation. Freezing encoder causes the significant decrease in accuracy (0.684) and macro-F1 (0.5561) and the model performs especially poorly on negative sentiment. Gradient-based adaptation of all layers, which full fine-tuning effectively restores to its limits, is needed to model the subtle and situation-specific sentiment cues of Weibo discourse, which remains true. In practical sense, this result means that encoder freezing can only be used when the ability to calculate is more important than accuracy.

Lastly, the findings of cross-validation support the general tendency: BERT demonstrates a consistent high level of performance (macro-F1 0.8665 ± 0.0118,) whereas TF-IDV is rather stable yet with the smaller maximum (0.8290 ± 0.0063). Together, these findings indicate that while Transformer models provide superior robustness and representational capacity across realistic social-media conditions, traditional sparse methods retain practical value under limited-resource constraints, especially when training data are scarce or fine-tuning budget is restricted.

5 Conclusion

This study set out to address several gaps identified in the existing literature on Weibo sentiment classification, namely the lack of controlled comparisons between sparse and contextualised representations, the limited understanding of how data scale shapes relative model performance, the absence of systematic noise-robustness evaluation, and the insufficient examination of fine-tuning strategies for pretrained Transformers. Through a unified experimental framework, this work provides empirical answers to each of these issues and offers a clearer characterisation of when and why different modelling approaches succeed.

First, the controlled comparison between character-based TF-IDF and a pretrained BERT model confirms that contextualised representations offer a clear performance advantage when moderate-to-large labelled datasets are available. BERT consistently achieves higher accuracy and macro-F1 scores across settings, demonstrating its superior ability to model informal, heterogeneous and context-dependent sentiment

cues common in social-media text. Second, the data-scale experiments reveal that the performance gap between sparse and contextual models is not fixed but shaped by the amount of supervision: while BERT benefits strongly from larger datasets, TF-IDF remains competitive in small-sample conditions and improves steadily as data increase, approaching Transformer performance at higher scales. These results explain the trade-off drawback in practice between representational richness and availability of data.

Third, the noise-robustness experiments indicate the stability-and-even-better performance of BERT when small perturbations are put in place, however, TF-IDF weakens over time, because it uses superficial token statistics. Herein lies the usefulness of contextual modelling in real-world user-generated content, where noise is natural, and there is no way to avoid it. Fourth, the experiments using fine-tuning indicate that fine-tuning of parameters is necessary in order to have good downstream performance. Encoder freezing results in significant performance drop, especially in negative sentiment, and thus it is important to note the relevance of gradient-based updates in mapping pretrained representations to task-specific semantics.

Taken together, these results provide a comprehensive, data-driven assessment of the relative strengths of sparse and contextual models for Weibo sentiment classification. Transformer-based models are preferable when robustness, semantic sensitivity and access to labelled data are priorities. However, TF-IDF remains an efficient and practically viable alternative in low-resource or computationally constrained environments, offering a strong baseline with stable performance.

Future research may extend this work to multi-class emotion classification, cross-platform generalisation, or more advanced data augmentation and adversarial training techniques. Such extensions would deepen our understanding of model robustness and extend the applicability of sentiment-analysis systems to broader real-world scenarios.

References

1. Wang, D., Alfred, R.: A review on sentiment analysis model for Chinese Weibo text. In: 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), pp. 456–463. IEEE, Shenzhen (2020)
2. Li, H., Ma, Y., Ma, Z., Zhu, H.: Weibo text sentiment analysis based on BERT and deep learning. Applied Sciences 11(22), 10774 (2021)
3. Lyu, X., Chen, Z., Wu, D., Wang, W.: Sentiment analysis on Chinese Weibo regarding COVID-19. In: Huang, D. et al. (eds.) Lecture Notes in Computer Science, vol. 12430, pp. 710–721. Springer, Cham (2020)
4. Zhang, Y., Wallace, B.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing, pp. 253–263. Asian Federation of Natural Language Processing (2017)
5. Sun, C., Huang, L., Qiu, X.: Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentences. In: Proceedings of NAACL-HLT 2019, pp. 380–385 (2019)
6. Mozafari, M., Farahbakhsh, R., Crespi, N.: A BERT-based transfer learning approach for hate speech detection in online social media. Online Social Networks and Media 21, 100114 (2020)

7. Peng, H., Cambria, E., Hussain, A.: A review of sentiment analysis research in Chinese language. Cognitive Computation 9, 423–435 (2017)
8. Ling, M., Chen, Q., Sun, Q., Jia, Y.: Hybrid neural network for Sina Weibo sentiment analysis. IEEE Transactions on Computational Social Systems 7(4), 983–990 (2020)
9. Shankar, V. & Parsana, S.: An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing', Journal of the Academy of Marketing Science, 50(6), pp. 1324–1350. (2022) doi:10.1007/s11747-022-00840-3.
10. Danish, S.M.H., Hasnain, S.M.E., Ashraf, H. & Rukaiya, R.: Comparative Analysis of BERT and TF-IDF for Textual Semantic Similarity Assessment, 2024 26th International Multi-Topic Conference (INMIC), Karachi, Pakistan, pp. 1–6. (2024) doi: 10.1109/INMIC64792.2024.11004377.