A Comprehensive Evaluation of Machine Learning and Transformer-Based Techniques for Sentiment Analysis

Haoran Xi

Leaden College, Nanjing University of Information Science and Technology, Nanjing, Jiangsu Province, China

202283710021@nuist.edu.cn

Abstract. Emotions, as an essential part of text, play a significant role in opinion mining, enabling the judgment of users' attitudes and the monitoring of public opinion. This study compares statistical learning methods, deep neural networks, and pre-trained models for sentiment polarity classification tasks using the Internet Movie Database (IMDB) movie review dataset, the ChnSentiCorp Chinese sentiment dataset, and financial data. Empirical evidence shows that traditional approaches, such as support vector machines—which go beyond classic algorithms—still demonstrate strong performance in classifying data. However, deep learning models based on bidirectional long short-term memory networks achieve considerable improvements in accuracy due to their ability to capture contextual information. More impressively, the Internet Movie Database (BERT) model achieves 100% on all evaluation metrics after knowledge distillation, further confirming the advantages of pre-training techniques for these datasets. These results provide insights into how different models evolve and highlight the need to balance model accuracy, computational cost, and data requirements in practical applications.

Keywords: sentiment analysis, deep learning, pretrained language models, traditional machine learning, natural language processing.

1. Introduction

Sentiment Analysis is also known as opinion mining and it can be considered as subset of NLP which extracts and understand the subjective information from the text. It's mostly about looking for feelings in writing and sticking them in different places, like good ones, bad ones, or not important ones. It is important branch in text classification and it got much attention both academically and industrially recently.

Sentiment analysis can only be judged based on what it can do, it's used by social media networks to ascertain if anyone has attitude-groups, to see how new fads arise. e-commerce business looks at what customers say about the things so they know if people like them and then they make the things even better. Financial analysts get their feelings from the news articles and people talking about the market and decides when to buy and sell. And governmental bodies have also used the same type of strategy as to what the public would do in regards to a major change in a policy or a major event.

User generate content is a lot of, so sentiment analysis is one that we can use the unstructured data to get more use full information, machine learning and deep learning also improve by the development of machine learning and deep learning technology, because there's many ways of the metho and in different fields so we have to have a great evaluation of the present one and the effectivity and possible development of new one.

There have been some changes with sentiment analysis in terms of methodology. In the beginning it was primarily machine learning algorithms such as the naive bayes classifier, support vector machines (SVMs), decision trees models etc. that used linguistic features such as n-grams that human beings made themselves. But it was tough for them to be really proficient with what exactly they meant by the nearly those words, and being good at following what happened all over that long a bit of the text.

To combat the aforementioned flaws, Zhang et al. created a brand-new network, i.e. the union or the combination of CNN and LMST. The idea behind this network, is that since CNN can distinguish local text pattern by LSTMs, LSTMs are good at sequence dependency modeling, so we have much

better syntactic and semantic representation, hence we achieve much better sentiment classification [1]. Architecture, like the transformer BERT, then came around - breaking from how things were done before was the introduction of a pre-training step that enabled bi-directional context understanding, followed by fine-tuning with regards to a certain task to set the new normal [2]. They are further upgraded and we get ROBERTA, ERNIE, MacBERT for predicting sentiments better [3-5].

The present research is going multilateral, Liu et al. combined textual, visual, and audio datasets. in these cases, it shows than a holistic emotion recognition system is better when using a combination of different kind of info [6]. This path is a bit more complex than just seeking for simple pattern matches and moving forward with statistical learning, to much more complicated neural representations and combining all different kinds of senses and how usable and technical it is keeps improving.

The sentiment prediction after the method review is based mainly on three categories: traditional ML, deep learning, pretrained language models. In explaining its elementary construction as well as for the sake of comparing the relative merits and practical limitations for its use.

And list many frequently data sets and scores of frequently used sentiment analysis to evaluate the effectiveness. Looking at the latest experiment. Apart from these new problems such as cross-domain and detecting the sentiment of the sarcastic expressions, there comes another one which is explaining the model.

Lastly, we will show some directions for future work on how to resolve the issues and make better and more practical usage of sentiment prediction systems. It is an empiricism that does practical fine-tuning experiments based on the conclusions reached from papers.

2. Main Techniques Overview

2.1. Application of Traditional Machine Learning Methods in Sentiment Analysis

Statistical machine learning algorithms were the technical foundation of sentiment classification before the emergence of deep learning. and the work has given some very important criteria for people that will follow this to study the topic [7]. The most typical feature of this kind of technical system is an emotive semantic clue selected by a manual feature description method with appropriate explanations for the text. Among the commonly used feature extraction methods, the most obvious feature is that they are based on statistical methods, using continuous word-sequence modeling methods and feature representation based on inverse-document-frequency weighting.

Looking at how complicated the model was compared to the old method; it is no comparison to the modern neural network. but however, a lot of empirical studies gave us that a lot of use case scenarios we have a lot of data or even less training data with a small to medium sized data but the traditional tech solutions are good for it it's worth trying it, in the algorithms they almost all use 2 types classic models. Firstly, Naive Bayes classifier uses the Bayesian theory for constructing a prediction model in which each and every attribute is conditionally independent. 2. SVM algorithms used the maximum margin concept for classification and create an optimal Decision Hyperplane in higher dimensions. Three, using logistic regression model which has a good tradeoff between both the prediction accuracy of the model and how interpretable the model is even though these are old but great studies on algorithms which does help me understand sentiments.

Especially what it needs to be mentioned is that we must convert the original text to feature vector number before applying it traditionally. Most often bag-of-words representation model is implemented with the feature building strategy that employs the TF – IDF weightage approach. Although type of this kind words' representation can show enough distribution characters of this certain word, but it would cause huge mistake to maintain original native's meaning connection as well as word's syntax. Regarding Traditional ML technology, there are definitely some advantages and low computing cost, which can easily be used in real-world engineering. transparent algorithms can be debugged easily,and detailed arguments can be made,and the performance is fine with a small

number of data. But it is still not very convenient. Modeling complex language phenomena does well, not scalable to process a large amount of unstructured text, requires too many professional domain experts for necessary feature engineering work.

Although in the studies academically, people are just talking about the deep neural network, but for the machine learning models, it also acts as a reference point for the researchers.

2.2. The Development of Deep Neural Networks in Sentiment Analysis

Recently, due to the availability of larger amounts of annotated data over the past few years and steady improvement on computer hardware, deep learning has seen the most success with various kinds of prediction problems other than other methods, and it's fundamental changing the sentiment analysis space [7]. New method is the main difference from previous works, it learns suitable text features by itself on raw data, without the tedious manual feature design which was the essence of previous methods. And if you get one that's really big and really strange, it does even better.

Current academic studies mainly concentrate on three types of typical neural networks: To start with,the CNN was firstly used in the computer vision task. Adding sliding window to try to get shorter parts of the local patterns from word sequence and trying to find the significant phrases containing emotions faster in some sorts of category identification tasks like in short texts appearing on social networking platforms and trying to include Recurrent Neural Network model (RNN) and its derivative, LSTM: By introducing a memory unit, they managed to keep long context information, semantics, flowing well and intact from one step to the next so as to overcome this shortcoming of the standard RNN structures as last one up here; and finally is bi-directionals LSTMS which takes into account the forward and backward pass happening at the same time w.r.t. each other offering a whole deeper awareness of every single word unit as 9: Also, it should also be noted that GRU, a type of reduction in LSTM, reduced the number of parameters of the model, preserving the performance of the LSTM model. The reduced parameters model can be employed in situations where fewer resources or a practical implementation of a very fast-iterating model is required [8-10].

out of above deep learning methods, the most popular way for pre-processing is to utilize distributed word vector representation likeWord2Vec, GloVeand Fast Text to convert these discrete words into a large continuous vector space where similar words become close to each other compared to using the previous way of bag-of-word sores presentation or TF - IDF representation which can catch more correct word semantics.

Deep Neural Network is not doing so bad either – they're showing some pretty good performance numbers, but they do have some tough problems as well. They need a lot of labeled data to feed them for training purposes and it consumes a lot more compute time than we would like sometimes, which consumes a lot of power and then you've got this whole inner part of how that decision was made and it's really just hiding itself in what we call a 'black box' model, which makes it quite difficult to explain. Deep learning is still, as of now, the most popular model that's used for Sentiment Analysis. The latest experiments show that even though they made an improvement on the context model of LSTMs, bi-LSTMs, those models using Trans form as a pre-trained base surpass all models even those models that had a lesser quantity of training data [2, 3, 10].

2.3. Pretrained Linguistic Models

Recent years have seen rapid progress of pretrained language models (PLMs), which brought great changes to sentiment prediction methods. Unlike traditional machine learning or early neural networks training, PLMs adopt an unusual twophase training process, which starts with unsupervised pretraining and then moves on to the task-specific finetuning phase. This framework helps us get advanced language features on auto pilot which helps to make it super helpful in understanding emotions from texts that are hard to read.

Among various kinds of PLMs implementation, BERT (Bidirectional Encoder Representations from Transformers) stands out. Devlin et al. put forth this design in 2019 alongside bidirectional transformer networks as well as masked language modeling targets while the model was being trained

initially. From experiments, we see that fine-tuned BERT always trumps both traditional machine learning models as well as earlier versions of neural networks in sentiment analysis. They are more accurate and trustworthy.

On going research is creating lots of boosted BERTs. RoBERTa doesn't have next sentence prediction but training data is larger and batch is big, RoBERTa gets better on sentiment etc. For Chinese Language Applications, the ERNIE model and MacBERT models demonstrate very good results. Baidu's ERNIE takes advantage of knowledge-informed masking for training purposes to allow more easily grasping of emotion-imbuing expressions with names, phrases, etc. MacBERT replaces regular MASK token with semantically equivalent terms during training to remedy the pretrain/fine-tune mismatch problem.

Practicing PLM demonstrates the sentiment predicting benefits of PLM. First, remove manual feature engineering but maintain good performance. Second, they give good results even when fine-tuning from few data. But they can show very malleable sentiment in most parts of the area. And PLMs have another issue, as well; they need a lot of power from our computing machines while they learn at first, and again while they teach at the end. It also doesn't give much insight with regards to how they made the decision to predict something. According to my personal observation, it is really tough to pick out the sarcasm, irony and other hard to tell emotions and expressions that does not exists in the pretraining data.

Still, regardless of this, pretrained lang model became the go-to for sentimental analysis. And they are being hugely utilized in academia as well as industries. Sentiment Prediction: Each iteration of a model as well as the development of better fine-tuning methods the role of PLMs in sentiment prediction of the future will only become stronger. Parameter Efficient Fine-tuning technologies such as Adapter and Prompt-tuning, try to reduce the resources required for PLMs. Maybe it could make the practical applications of PLMs faster for various applications.

3. Applications & Comparative Analysis

3.1. Benchmark Datasets for Sentiment Analysis

What kinds of corpus to use for evaluation is an elementary factor in building nice SA/systems. In reality, good quality datasets are great for training and validating the models but have big effects on real world generalizations. The newest research created quite a number of standard data sets, some of them with different places, languages, modes, etc. The above data has some advantages over checking the model.

IMDB Movie Reviews Dataset is regarded as amongst the most well-established benchmarks with a total of 50K reviews of movies, which were labeled by experts and they have a 50-50 ratio of positive and negative reviews in the experiment. There was a special spot on testing a model of how it'd handle a long storyline in terms of context. Architects with sequence parts do well. It includes a lot of different types of LSTMs as well as a transformer-based model

The Stanford sentiment treebank (SST-2) is like another kind of test where we have to guess whether a sentence is good or bad. And it's also phrase level annotations by themselves don't let us evaluate sentence level sentiment compositionality—note that this is a dataset that appears in the GLUE benchmark set. This is the test model as BERT and such as similar model form.

For commercial sentiment analysis apps: amazon product review is awesome, because it is huge and has millions of reviews and different sections. Practically speaking, it still has some value for studying domain adaptations and large models. Detailed star ratings can enter into the sentiment analyses outside only two classifications.

Chinese language studies mostly use two kinds of benchmark corpuses: ChnSentiCorp and Weibo Sentiment Corpus, by comparing these two datasets, ChnSentiCorp is easier to classify, there are professional hotel, product reviews compiled. The opposite is Weibo; they contain lots and lots of analysis difficulties compared to other datasets as well. The dataset contains more colloquial words

and more emotions. Too much internet lingo, like you know nothing about any of it, and a model like erine, MacBert.

Emergence of multimodal datasets (CMU-MOSI, CMU-MOSEI, etc.) with text, vision, audio. This is an active area of study. and not directly appear in this paper, but be used for cross-modal affective computing.

It's hard to use some datasets like annotations with all kinds of bias, different labelling, don't work well when transferred to another domain. This points to making sure you have a good dataset for what you want to do with your research.

3.2. Performance Assessment Methodologies

When evaluating a sentiment predict system, we need to be careful when picking our measure so we can actually test the models well. From the above discussion, we see that the attributes of performance are dependent on the type of measure so it is necessary to use different sets of measures depending on the situation.

So, the paper could easily say that classification accuracy is the easiest to say - total amount of correct predictions. But now when we look at real experiments, we have the problem of having imbalanced classes and it is very common. And we could already start seeing that here; Alternative metrics try to solve it by using trickier measuring methods.

precision remember analysis supplies the other story on how great our model is. Predictability is measured by how accurate the predictions are and coverage is determined by recall: F1-score is everything combined together. And that's something helpful for us as we're in an operational environment were having a false positive or having a false negative that's bad for us. So, a commercial sentiment system would need a number like this if they wanted to do business.

Probabilistic classifiers, the AUC - ROC gives you this really cool kind of threshold-agnostic way to evaluate, it's not something people really use as a common benchmark but it's really cool because for lots of different problems it really does matter if we're really sure about our predictions.

There are some multi case classes that add a bit of complexity but at the end the difference between macro-micro is how the case will look like. The former treats all the classes equally with same probability, but the latter puts more weight on the frequent classes (another point to be noted when compared the model on domain where class distribution is different)

The vast majority of the current research reports on F1 – score and accuracy and while they both indicate that something is interpretable and informative, the right metric for your app may be to find that if you really need the highest recall of negative sentiments in your CS app, or if you want something that is super accurate like content moderation systems, but this is a big piece of what I'm trying to build something that's useful for sentiment.

3.3. Comparative Analysis of Model Performance Experiments

In fact, in case 1, on the basis of comparing it in the experiment, the paper can compare to analyze the sentiment analysis technology of the past. According to the results of the two benchmark dataset, IMDB, ChnSentiCorp,as the model becomes more complex, both precision and f1 score will increase sharply.

A few calcu's left over to get some reasonable baselines on the datasets with classical SVM, NB In terms of this particular angle where ordinary words' characteristics were used when SVM achieved 78%, when people want to find out all the hidden meanings that a word could give, they don't really do well, they are really better for times where you have to know very well what you're talking about but not really how you achieve it.

The neural network model, clearly, wins. The bidirectional LSTM, for the IMDB English movie review dataset, got up to 85, +7 compared to what was normally used. The main reason for the improvement is that the neural networks can model sequences. And also, it comes with extra computation.

Pre-training language model sets a new benchmark. BERT Distill from Knowledge get 93% accuracy test of IMDB and over 90% F1 Chinese test. The above results can entirely demonstrate the advantages of transfer learning: and the large-scale pretrained language knowledge is very good at transferring to these tasks and also quite notable was that on the test set of the Professional Finance sentiment dataset - 200 pieces of data with 5 different sentiments; every score on all evaluations returned a 1.00 as can be seen in table 1. It is as a test on some more modern pre-training methods having better generalization performance in some slightly less data fields.

From our experiment result we can observe that the performance of the model has positive correlation with the computing requirement. From the application point of view, the conditions of the hardware, the characteristics of the data and the purposes of the company all have to be considered. as well as in the other direction, the paper thinks for some future research it might be able to do a better domain-specific adaption and an interpretable model for the even more critical cases.

Table 1. Performance of the DistilBERT model on the stock_emotions_subset.csv dataset.

Model	Accuracy	Macro F1	Dataset
DistilBERT (ours)	1.00	1.00	stock_emotions_subset.csv (200 samples)

4. Challenges and Future Perspectives

And although there is a lot of progress in many forms of different machine learning in regards to sentiment analyses, still a lot is needed and many of these problems in sentiment analyses affect the ways people use the sentiment predicting system for their individual uses. Firstly, let us look at the problems then possible research directions with solutions to problems.

4.1. Challenges

4.1.1 Data Reliability and Annotation Challenges

The quality and representativeness of the training sample is the root of good sentiment analysis. There are better standard evaluation corpora like IMDB or SST-2, but real-world data sources such as Twitter or Weibo include informal writing, irony and sarcasm and make many cases ambiguous, making accurate classification difficult.

Sentiment interpreteation is subjective. There would also be some labels not aligned up. Experiment shows that even good models still struggle when it comes to emotions, because there are not very many examples of complicated feelings in the training data. It’s easy to see if we think about less obvious ones like irony, or mixed feeling.

4.1.2 Cross-Domain Generalization and Multilingual Limitations

The system nowadays is not that flexible per subject. When models from one topic are used on another topic (e.g. trying to use an entertainment model for financial news) the performance will be worse, it is mostly because the terms are different as are the way people talk.

And as far as them being multilanguage-wise, there’s another whole bunch of problems. English-based models also get decent results, but does not work on low resource language and code-mixed stuff. Models like ERNIE can gain some gains on Chinese task without even one multilingual benchmark, hard to judge.

4.1.3 Transparency and Model Explainability

As sentiments have become more complex even transformers, that’s like as a form of architecture have become more complicated even have interpretability problem. Complex models can bring about improved accuracy, but they behave like a black box which makes it difficult to use them for places where seeing how people made decisions matter.

In practical life we want predictions and good reasons for our predictions. Approaches conventionally have understood paths via interpretable paths, where as we already have models with neural do not have that, so in those high stakes situations that is important.

4.1.4 Resource Requirements and Practical Deployment

Advanced sentiment analysis models encounter problems as well. Training and fine-tuning big neural architectures need a lot of hardware resources so firms with little infrastructure won't be able to participate.

Even the best still needs to put in effort. Experiments show and this is further confirmation that the need for more resource efficient solutions that are better suited for real time and edge computing is still required.

4.2. Prospects for Cutting-edge Research Directions

Considering that there are already many technical bottlenecks at present, the paper has found certain routes which might cause a breakthrough. Primary direction is multimodal emotion, these are types of techs that work by adding many things like letters, pictures, and sounds. Though there do exist some datasets with basic structure, one should still put efforts on exploring the cross-modal feature fusion more deeply. low-resource learning paradigms has to be created. Few-shot learning, if Prompt tuning can provide a new solution in a world with limited labeled data. In particular, for some field or for some language group

On the unsupervised side of things, developers are testing out new kinds of self-supervised tasks so that they could be able to just find some sentiment features inside a massive volume of text all by its lonesome without ever having to take the time to label it themselves. Also, at this time for this network to be used by them by using this technology that would be reducing the size of the model and distilling the knowledge from the network, creating a lightweight, which is important, and also it has some explainable like attention visualization of what's important so we can also see why and how they predict works for us.

especially need us to notice is that every step forward must be accompanied by a reasonable proportion of innovation and practice. In the future, sentiment analysis systems need to not only improve precision, accuracy, effectiveness, but also make many improvements such as stability, interpretability, efficiency, etc., to reach the requirements of various applications. Multilevel optimization makes sentiment computing get out of the lab and serve in reality for the service of the big society.

5. Conclusion

This study comprehensively reviews the development process of sentiment analysis technology, and conducts comparative analysis and compares the technological development path of sentiment analysis from traditional machine learning to the use of pre-trained language models for comparisons. Sentiment analysis technology belongs to the primary work of the part of NLP and also has irreplaceable practical significance and applications in social media monitoring, market trend prediction, financial public opinion early warning, and so on.

Early emotion recognition systems used mostly traditional methods such as probability classifiers and support vector machines. those all need handcrafted feature engineering like words occur frequency and tf-idf weight. though the computation efficiency is relatively higher, the degree of semantic understanding is relatively lower, more especially in complex conditions or vague phrases, their performance is quite poor.

With the introduction to deep learning technology, this is a revolutionary change. Convolutional neural networks are used with Recurrent neural networks to automatically extract text features. Distributed word vectors are used as an input to be integrated into the system to improve the accuracy of sentiment recognition. But there were problems like needing a lot of data and it not being easy for the model to understand also happening - these formed new tech problems when doing things from what was learned.

Based on the Transformer structure the pre-trained model is also a change in the technical paradigm. Using a DistilBERT to get a 100% fully correct classification in the financial sentiment

data and this shows that for a general model like this one it has an amazing generalization ability despite being only a small number of entries. And it is due to the fact that there is a good amount of language understanding achieved from pre-training on such a large amount of data. They can understand the sentiment of the text.

And, it explored in-depth about the evaluation system in the world of sentiment analysis. Cross-domain and multilingual benchmark datasets give big support for model making, and aimed evaluation numbers make sure comparison checks do not taste. And current technologies still need to solve many problems: how to deal with noisy data, how to do cross-domain adaptation, and how to make models more interpretable and computationally efficient.

For future works, we could consider something like developing some multimodal frameworks, developing some new low resource learning methods, getting into unsupervised representation learning, or trying out a better engineering deployment solution. Also, very important is highlighting that if we're looking for a tech breakout, we cannot forget about moral problems and practical application value. An ideal emotion analysis system needs to be precise, strong, transparent and effective. The development needs to integrate technology innovation and responsibility naturally. When it comes to a “two-track” like this, the emotion-analysis technology should be able to produce social value in a greater number of ways.

References

[1] L. Zhang, S. Wang, & B. Liu. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253. (2018).

[2] J. Devlin, M. W. Chang, K. Lee, & K. 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 (NAACL). (2019).

[3] C. Sun, L. Huang, & X. Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). (2019).

[4] H. Xu, & K. Mao. A survey of aspect-level sentiment analysis based on deep learning. IEEE Access, 9, 71857–71870. (2021).

[5] P. Liu, W. Li, & Y. Zou. Empowering BERT with knowledge graph embedding for document classification. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI). (2019).

[6] Q. Zhong, et al. Knowledge-enriched transformer for emotion detection in conversations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). (2020).

[7] Y. Zhang, & Q. Yang. A comprehensive review on cross-domain sentiment analysis. ACM Computing Surveys. (2022).

[8] Z. Lin, et al. A unified model for sentiment analysis with multi-task learning. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). (2020).

[9] P. Zhou, et al. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). (2019).

[10] X. Huang, et al. Prompt-based fine-tuning for sentiment classification in few-shot settings. In Proceedings of the 29th International Conference on Computational Linguistics (COLING). (2023).