IOP Conference Series: Earth and
Environmental Science

PAPER • **OPEN ACCESS**

A Big Data Analysis of Job Position Status Based on Natural Language Processing

To cite this article: Qi Liu 2021 *IOP Conf. Ser.: Earth Environ. Sci.* **693** 012026

A Big Data Analysis of Job Position Status Based on Natural Language Processing

Qi Liu*

University of Illinois at Urbana-Champaign, Champaign, IL, USA.
Email: liuqibaobao1994@126.com

Abstract. With the development of the Internet and computer technology, news on the Internet has gradually been able to influence the development of the economy. In some special periods, with the development of major events, the economy and finance have undergone similar development and changes. This paper proposes a simplified natural language processing method for multi-grained text. First, the news text data is classified according to sentiment, and the correlation between corporate loans and job retention ability is further analysed. The experimental results show that under certain conditions, the Loans issued to SMEs are more conducive to retaining the number of jobs than large enterprises.

1. Introduction

Data from customers and markets has high value to Fintech companies. With large datasets, information about consumer preferences, consumption habits, and investment behaviour can be extracted and used to develop predictive analytics. ATM and credit cards can be regarded as the earliest financial services. Online and mobile banking services appeared very early, followed by Block Chain and other B2C payments.

Fintech is a term used to describe financial technology, an industry encompassing any kind of technology in financial services from businesses to consumers. Fintech describes any company that provides financial services through software or other technology and includes anything from mobile payment apps to cryptocurrency. Predictive analysis refers to the use of past information and mathematical algorithms to predict consumer behaviour. The data collected can also help formulate marketing strategies and fraud detection algorithms [1, 2, 3].

The common smart investment advisor application in Fintech is a typical application, mainly for customers to develop financial planning. By using artificial intelligence technology and intelligent applications to collect and complete basic asset condition evaluations and give suggestions online, it greatly reduces the waste of human resources. The latest Fintech has been completely changed by artificial intelligence. From financial operations to customer interaction, most of the work has been taken over by artificial intelligence. Natural language processing is an important branch of artificial intelligence. Since most financial activities require the use of human language to communicate, the use of NLP technology is very beneficial to the performance of Fintech. From data processing and analysis to regulatory compliance to instant customer service, the goal is to enhance rapid business decision-making and further increase productivity.

NLP solutions have the ability to mine and find key information and classify them from unstructured data. By using NLP technology, the document search capabilities could be improved quickly. The processing time is greatly reduced and the accessibility of data is further improved.

The technological development of NLP ranges from rule methods, statistical methods to deep learning. Deep learning methods can handle NLP tasks with fully annotated corpus, model the input and output, and output the results on the test data. However, once the distribution of the test data is not in the distribution of the training data, errors will occur. The reason is that NLP models do not have a large number of general knowledge models that humans have.

The difficulty of natural language processing is to eliminate the ambiguity in morphology/syntax/semantics. This elimination requires a lot of knowledge, including linguistic knowledge and general knowledge, which leads to the main difficulty of NLP: natural language is very complex, but vocabulary and syntax are very complex. Limited, resulting in many ambiguities in the morphology, syntax and semantics of human languages. In addition, the knowledge needed to eliminate lexical/syntactic/semantic ambiguities is difficult to obtain, and how to process and design NLP models is a very difficult task.

With the progress of globalization and possible epidemics, it is necessary to consider capital flows and disease-related costs when designing economic models. For example, consumer demand may drop significantly, and people's confidence in the economy may weaken. This paper proposes an NLP method based on self-media information mining, the purpose is to analyse the current situation of enterprises during the epidemic.

The structure of this paper is as follows: first introduce the development of machine learning and deep learning, then give a BERT-based text sentiment classification method, and finally experiments and conclusion are given.

2. Machine Learning and Deep Learning

Some common applications of machine learning include face recognition, target detection, audio recognition, etc. The machine learning algorithms in Fintech technology mainly exist in the form of chat bots, search engines and other applications. The financial system is the main area of AI implementation, and the number of financial companies using machine learning continues to grow.

Machine Learning technically is to search for useful representations of some input data, within a predefined space of possibilities, using guidance from a feedback signal [4]. The research and construction of machine learning is a special algorithm or method that allows the computer to learn from the data to make predictions. Machine learning is the application and science that make data meaningful, which is the most exciting field in computer science. In the era of abundant data, computers can use self-learning to convert data into knowledge. Mitchell formally defines the algorithm researched in the field of machine learning: if there is a learning task T, and experience E is provided, the performance of the computer program will also be evaluated, if its performance on task T P follows the experience E Provide and improve, then it can be said that the program learns task T from experience E [5].

Machine learning can be divided into supervised learning and unsupervised learning according to whether it is supervised or not. In the former, the computer will get training data and target results. The purpose of learning is to learn the rules that can map input to output according to feedback, when not every input has a corresponding When the target result or the form of the target result is limited, supervised learning can be divided into semi-supervised learning, active learning, and reinforcement learning. In unsupervised learning, the training set has no artificially labelled results, and the computer needs to discover the structural rules in the input data by itself [6, 7, 8].

Deep Learning is a specific subfield of machine learning, which is a new take on learning representations from data that puts an emphasis on learning successive layers of increasingly meaningful representations. The deep network is a multistage information-distillation operation, where information goes through successive filters and comes out increasingly purified. Where Learning means finding a set of values for the weights of all layers in a network, such that the network will correctly map example inputs to their associated targets. Loss/Objective Function takes the predictions of the network and the true target and computes a distance score, capturing how well the network has done on this specific example. Optimizer implements what’s called the Backpropagation algorithm, to adjust the value of the weights a little, in a direction that will lower the loss score for the current example [9, 10]. A typical deep learning neural network framework is shown in Fig.1.

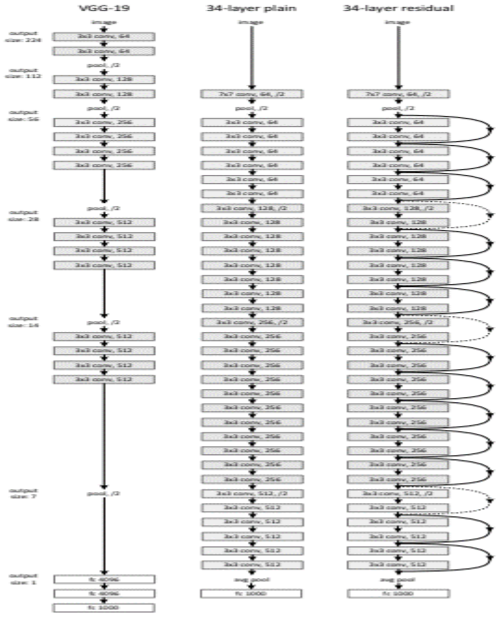

Figure 1. An illustration of the deep learning architecture [8].

3. A Text Sentiment Classification Method

Human language is an abstract information symbol, and computers cannot directly understand and process human language. Natural Language Processing is an important direction in the field of artificial intelligence. NLP spans multiple fields from artificial intelligence to computational linguistics, with the goal of accurately and quickly processing a large number of natural language corpora.

Embedding vectors extracted from text data can be used for keywords and semantic search. For example, when searching for document records in Q&A applications, even if there are no keywords, embedding vectors can achieve contextual meaning matching result retrieval. Second, these embedding vectors are used as input to downstream models. Generally speaking, the NLP model needs to be input in the form of a vector, that is, the word is either expressed as a unique index value or expressed as a word embedding, where the vocabulary word matches a fixed-length feature embedding. And Bert's processing is that no matter what context the word appears in, it will be dynamically represented by the context word. Embedding in NLP is a method used to represent discrete variables as continuous vectors, which could be used in a computation process. There is a simple method for embedding called one-hot encoding, which takes discrete entities and maps each observation to a vector of 0s and a single 1 signaling the specific category.

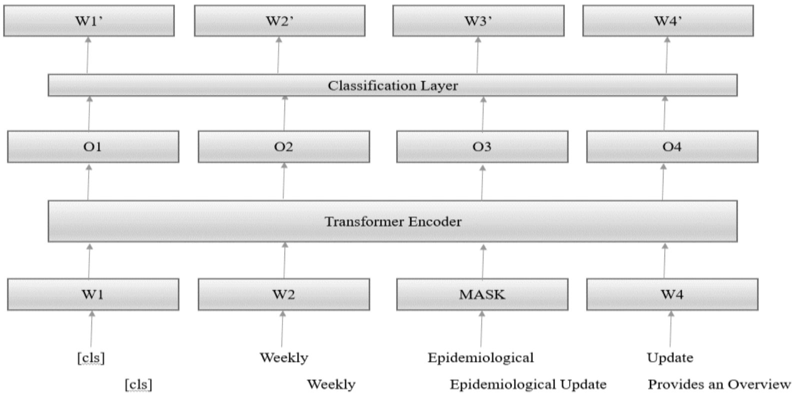

Figure 2. The simplified multi-granularity Bert structure.

There are two main limitations with respect to this method. One is that the size of the network would be too large, and the other is similarity between entities would not be represented by this one-hot encoding. We can greatly improve embedding by learning them using a neural network on a supervised task. The embedding form the parameters of the network which are adjusted to minimize loss on the task. The resulting embedded vectors are representations of categories where similar categories, relative to the task, are closer to one another [11, 12, 13].

BERT is a pre-trained language representation method developed by Google. In SQuAD1.1, the top level test of machine reading comprehension, the GLUE benchmark was pushed to 80.4%, and the accuracy of MultiNLI reached 86.7% [14]. Pre-training models such as BERT have migrated the knowledge in the corpus into the Embedding of the pre-training model through unsupervised learning of a large number of corpora. Fine-tuning by adding structure to specific tasks can adapt to current tasks. ELMo could adjust the word embedding according to the given context. Its network architecture uses a two-layer bidirectional LSTM on top of the word embedding layer learned by word2vec.

After ELMo, each token would be corresponding to several embedding vectors, on which we perform a weighted sum. This weighted sum is then input into the downstream task. are learned together with parameters in the downstream task. ELMo is a feature-based language representation model.

GPT uses static word embedding as input, plus a certain number of layers of the Transformer decoder on top of it. When fine-tuning specific NLP tasks with supervised learning, GPT, unlike ELMo, which connects to other model layers as feature representations, does not need to re-build new model structures for tasks. Instead, it directly connects the last layer to softmax as the task output layer, then fine-tunes the entire model.

BERT is a two-stage model where the pre-trained for the bidirectional language model and the fine-tuned for downstream tasks were provided. Some new structures use three-stage training: language model pre-training, downstream task language model fine-tuning, and downstream classification task fine-tuning. This paper proposes a simplified multi-granularity Bert structure, which is shown in Fig.2. Generally speaking, the input of Bert contains two granular information of token and segment. Because the downstream model of this paper has a clear goal, token and segment are almost the same, so in order to reduce the computational cost, the two are combined as one input and the embedding layer is cancelled. In addition, in terms of word segmentation The method of word segmentation according to different granularities is adopted, and different language models are trained and merged on these different granularities, because the text with larger granularity is more convenient for users.

4. Experimental Results

COVID-Twitter-BERT (CT-BERT) is a transformer-based model based on BERT [15], which is pre-trained on a large collection of Twitter messages about COVID-19, mainly on January 12 to 4, 2020 A total of 22.5 million Twitter message collections collected keywords related to COVID-19 during the month of 16th. The enterprise loan data set comes from the salary protection plan loan data [16], this data set contains all the loan enterprise information provided through the Paycheck protection plan. In this paper, the threshold is set to 100k to distinguish the size of the enterprise and count the job retention of both.

This paper adjusts train batch size and learning rate, etc., to classify the sentiment of the epidemic keywords from the COVID Category data set within three months, such as epidemics, Community spread, Containment, etc., and compare them with the proportion of the number of people infected by the epidemic in the population, which is shown in Fig.3 and Fig.4. For comparison, it can be observed that the trend of the panic index obtained by Negative words is close to the trend of the number of diagnosed people. Therefore, it can be considered that the trend of negative words in the media is consistent with the development trend of the epidemic to a certain extent.

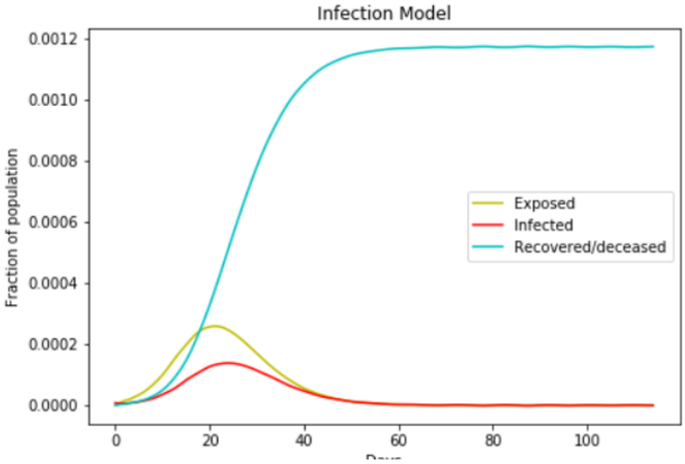

Figure 3. The infection model of media terms and the development trend.

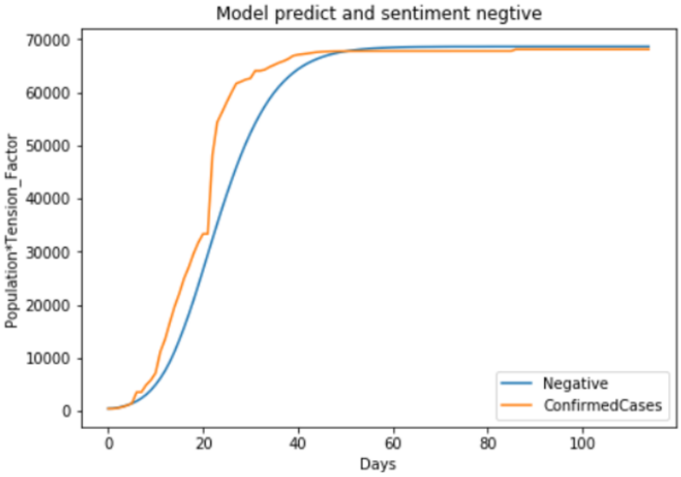

Figure 4. The model predicts and sentiment negative of media terms and the development trend.

Table 1. Comparison results based on loan data of salary protection plan.

Number of jobs/Total loans	>100k(%)	<100k(%)
February	0.021	0.018
March	0.010	0.011
April	0.012	0.014
March	0.016	0.017
April	0.021	0.015

The results are shown in Tab.1. By calculating the ratio of loans to retained jobs to show the contribution of different companies to providing jobs. It can be seen from the table that with the development of time, the retention trend of the number of jobs in large enterprises that have loans is not as good as the retention trend of small and medium enterprises that have loans. In June, the difference in the number of jobs between the two has narrowed. 100k enterprises are more capable of recovery. In short, granting more loans to SMEs will help retain the total number of jobs in the market.

5. Conclusion

This paper classifies text sentiment based on BERT epidemic news and self-media texts, and compares its changes with the contribution of small and medium-sized enterprises and large companies to job retention. Experimental results show that granting more loans to SMEs during the epidemic is better for retaining the total number of jobs. The next step will continue to improve the NLP model based on the bert structure.

6. References

[1] Demirguc-Kunt, A., Beck, T., & Honohan, P.(2007). Finance for all ?: Policies and pitfalls in expanding access. A World Bank policy research report (pp. 1–268).

[2] Lee, I., & Shin, Y. J.(2018). Fintech: Ecosystem , business models, investment decisions, and challenges. Business Horizons, 61(1), 35–46.

[3] Zetzsche, D. A., Buckley, R. P., Arner, D. W., & Barberis, J. N. (2017). From FinTech to TechFin: The Regulatory Challenges of Data-Driven Finance. Social Science Research Network.

[4] Krizhevsky, A.,Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90.

[5] Mitchell, T. (1997). Machine Learning. McGraw Hill.

[6] Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of The 32nd International Conference on Machine Learning (pp. 448–456).

[7] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778).

[9] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W. Macherey, K. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv Preprint ArXiv:1609.08144.

[10] Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R. J., & Saenko, K. (2015). Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1494–1504).

[11] Bird, Steven, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., 2009.

[12] Gardner, Matt, et al. Allennlp: A deep semantic natural language processing platform. arXiv preprint arXiv:1803.07640 (2018).

[13] Ruder, Sebastian. Neural transfer learning for natural language processing. Diss. NUI Galway, 2019.

[14] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. N. (2018). 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).

[15] Müller, M., Salathé, M., & Kummervold, P. E. (2020). COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. ArXiv Preprint ArXiv: 2005. 07503.

[16] https://www.kaggle.com/govtrades/sba-paycheck-protection-program-loan-data