PANAGORA

Teaching Machines to Understand Chinese Cyber-Slang

Chen, Lee, Mussalli

Introduction and Summary

Over the past several decades, the Chinese economy and market have both undergone dramatic changes. With the recent A-Shares index inclusions,[1] the Chinese equity market is increasingly accessible and important for foreign investors to investigate. This paper presents one of PanAgora's recent Chinese A-shares research findings, using the latest machine-learning (ML) technique. We find that:

- Despite its large size, the Chinese equity market is still quite inefficient and exhibits some interesting characteristics.
- Investors can get large amounts of data on Chinese investors if they know where to look. Having an understanding of local culture, language, and markets helps.
- By using ML techniques, we can gain insight into retail investment decisions.
- Delivering machine-learned alpha requires not only machine-learning expertise, but also domain knowledge and the right data.

Market Background

Paralleling the breathtaking growth of the Chinese economy, the Chinese domestic stock market has become the second largest stock market in the world, and one of the most liquid.[2] Even so, it is by no means a well-established, orderly, and efficient market. A few of the interesting characteristics of the Chinese A-Shares market include:

- Very heavy retail participation. By some studies, retail investors account for around 80% of the trading volume. The heavy participation of retail investors has led to interesting market behaviors (such as concept-hyping,[3] rumor-driven trading, high turnover rates,[4] etc.) and higher market volatility and inefficiencies than in more developed equity markets.
- Due to Chinese government restrictions, there is a relatively small amount of foreign capital participating in the Chinese equity market. In the past, restrictive rules on capital control and repatriation of profits have also discouraged foreign investors' involvement in it. Reforms on opening up the Chinese market are ongoing, but there is still a long way to go to achieve the same level of access as other East Asian markets. As a result, the Chinese A-shares market exhibits much more domestic influence than do other East Asian developing economies' markets.

[1] https://www.msci.com/msci-china-a-inclusion
[2] For more details on the history and evolution of the Chinese domestic stock market, see [SL, 2018].
[3] For example, see Reuters, "Chinese stock regulator challenges corporate hype to cool economic zone fever," https://www.reuters.com/article/us-china-xiongan-hype-idUSKBN17A090.
[4] See [SL, 2018] for more details.

- As with many aspects of the Chinese economy, strong government intervention and censorship can affect the Chinese A-shares market. For example, short-selling was temporarily banned during the Global Financial Crisis of 2008, and discussions about the decline of the Chinese market were discouraged[5] in 2018.

The Chinese A-shares market's distinctive characteristics, along with vast amounts of available data,[6] make it a great place for quantitative investors to extract alpha.

Considering that 80% of A-shares market trading is initiated by retail investors, it would be useful to understand what these investors think about various stocks and the market. Therefore, this paper focuses on the problem of understanding Chinese retail investor sentiments. Its data source is Chinese online investor discussion forums,[7] which retail investors love to visit because they can make stock recommendations on a plethora of A-shares stocks and defend their selections. After collecting this data, we leverage the latest natural language processing (NLP)[8] techniques and modify them to help understand the Chinese language discussion of the stock market, to gain insights into Chinese retail investor decisions.

Challenges to Machine-Learning Understanding of Chinese Retail Investors

Natural language processing, also known as computational linguistics, has come a long way from its early days of writing out symbolic rules of grammar. In the last 15 years or so, statistical language modeling and machine-learning techniques have dramatically increased the power of machines to understand spoken and written language.[9] Despite dramatic technical advances, many of the studies were done only on English and other alphabet-based languages. There is comparatively little NLP literature on machine-learning understanding of the Chinese language.

Compared to most alphabet-based languages, the Chinese language poses several challenges. The first is that written Chinese is based on characters rather than on combinations of letters. Chinese characters evolved over 5,000 years from pictures, much like ancient Egyptian hieroglyphs. A Chinese word can be a single character or composed of several characters. Compared to 26 letters in the English language, there are over 50,000 characters in Chinese, although most Chinese speakers know only a portion of them. Even an educated Chinese person's daily repertoire includes only about 8,000 characters.[10] Therefore, the number of possible permutations in Chinese is much higher than in English. To complicate the matter,

[5] *South China Morning Post*, "Chinese censors wipe out discussions about stock market rout as key index tumbles past 2015 low," https://www.scmp.com/business/money/markets-investing/article/2164862/chinese-censors-wipe-out-discussions-about-stock.
[6] See McKinsey & Company's study "Digital China: Powering the economy to global competitiveness," https://www.mckinsey.com/featured-insights/china/digital-china-powering-the-economy-to-global-competitiveness.
[7] These online investment forums are called *Guba* or *Stock-Bars*.
[8] Recent advances in NLP, including the techniques described in this paper, use methods from machine learning, and machine learning is a sub-branch of artificial intelligence.
[9] For more details, see Wikipedia: https://en.wikipedia.org/wiki/Natural_language_processing.
[10] http://www.bbc.co.uk/languages/chinese/real_chinese/mini_guides/characters/characters_howmany.shtml

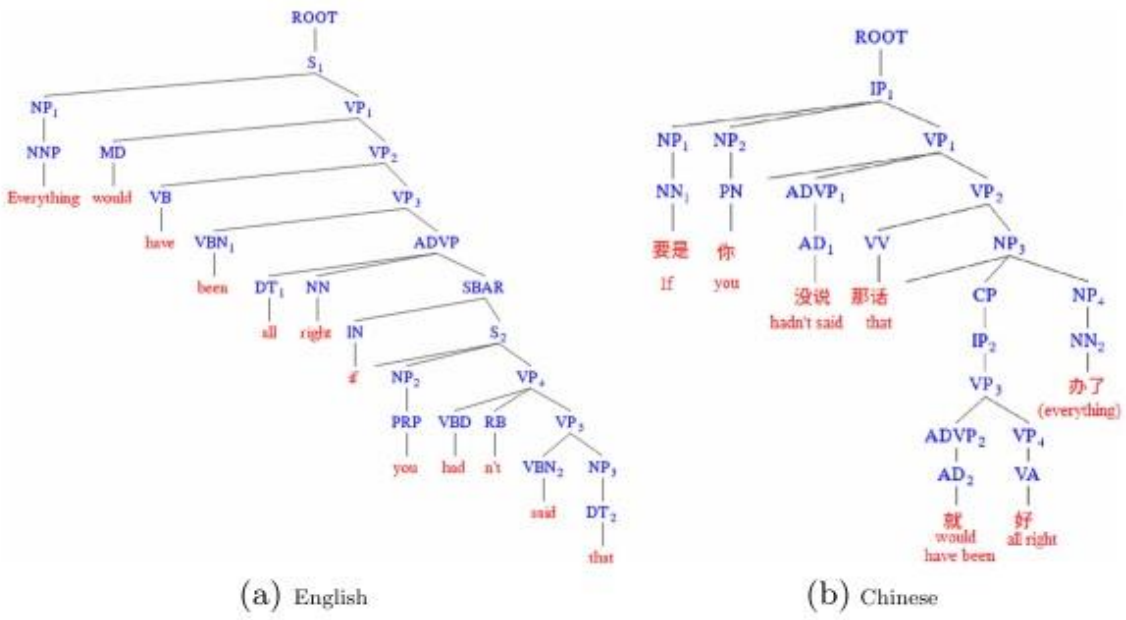
spaces are not used to separate words in written Chinese. Distinct words are understood "innately" by Chinese readers, based on the context of the sentences in which words appear.

Figure 1: Chinese characters evolved from pictures over 5,000 years.[11]

The second challenge is that the Chinese language has very different—sometimes even opposite—syntactic structure compared to most alphabet-based languages. Figure 2 shows the syntactic structure of English versus Chinese for the sentence "Everything would have been all right if you hadn't said that." Note that the syntax tree structures are very different. It's evident that most of the English syntax-based NLP algorithms—those most readily available—are ill-suited for Chinese language applications.

[11] Source: http://www.ancientscripts.com/chinese.html

Figure 2: Chinese versus English syntax tree[12]



The third challenge is that Chinese is a tonal language. That means spoken words that sound similar except for a small tonal difference can have vastly different meanings. For example, the character for *horse* (马) and the character for *mother* (妈) sound almost exactly the same when spoken in Chinese, but they obviously mean different things and are written differently. The tonal similarity of different words, coupled with strong Chinese government censorship, gave rise to a cyber-slang culture among Chinese netizens to create words that are written very differently from the way they are phonetically spoken. This allows true verbal expression while avoiding government censorship. For example, Figure 3 shows two sentences that are phonetically similar but that have very different meanings. The first sentence reads, "This restaurant is really rubbish," while the second sentence reads, "This restaurant is really spicy chicken."

Figure 3: The Chinese language is a tonal language; different words can sound phonetically similar.[13]

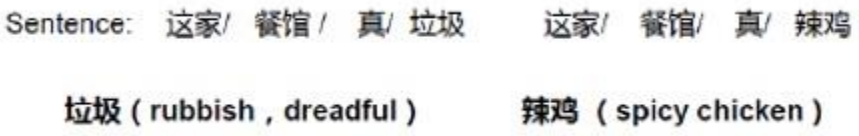


Figure 4 illustrates other examples of Chinese retail investors' cyber-slang. The two words shown on the top row under the headings each actually have three meanings. The first word in the Cyberspeak column, representing *river crab*, is phonetically similar to the Chinese word for *harmony*, which appears as the corresponding word in the Actual Meaning column. *Harmony* is a euphemism for government censorship, a

[12] Source [PCH, 2017]
[13] Source: PanAgora

term that first became popular during Chinese President Hu Jintao's administration, so the cyber-slang word for *river crab* actually refers to government censorship. The second word in the Cyberspeak column can be interpreted as a Chinese person's name, "Guo," but it is phonetically similar to the word representing *national team*, in the corresponding Actual Meaning column. The term *national team* denotes organizations that act on behalf of the Chinese government to ensure stability in the A-shares stock market, especially during times of volatility and stress.[14]

Figure 4: Examples of Chinese retail investors' cyber-slang[15]

Cyberspeak		Actual Meaning
河蟹 (river crab)	≈	和谐 (harmony)
郭嘉队 (a person)	≈	国家队 (national team)

NLP Methodology

Until four or five years ago, sentiment detection with natural language processing (NLP) was most commonly done using the Bag-of-Words (BoW) approach.[16] In this model, dictionary entries are categorized as either positive or negative words, and sentiment in a document is measured simply by counting those with positive sentiments and those with negative. The most famous of the sentiment dictionaries in English for financial documents is the *Loughran-McDonald Master Dictionary*.[17] Using the simple Bag-of-Words approach, and with the right dictionary, anyone can make a crude estimate of document sentiments.

However, establishing a sentiment dictionary for the Chinese language is difficult—especially for texts written on the internet discussion forums favored by retail Chinese investors, where the use of cyber-slang is prevalent—because:

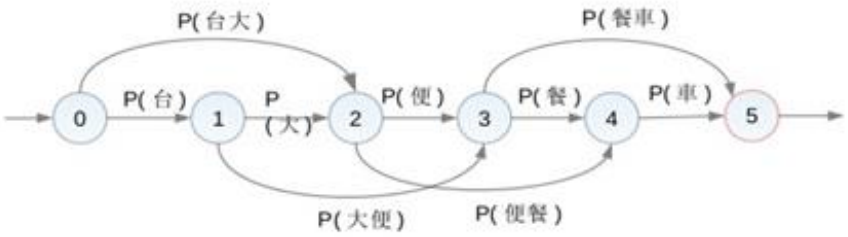
- The possible combination of words composed of single or multiple Chinese characters is very large.
- The presence of cyber-slang makes the possible combinations of sentiment words even larger.
- Cyber-slang, like all popular slangs, is rapidly evolving.

[14] See Bloomberg, "China's Plunge-Protection Team Is Poised to Save Stock Markets," https://www.bloomberg.com/news/articles/2018-09-18/china-s-plunge-protection-team-is-poised-to-save-stock-markets.
[15] Source: PanAgora
[16] See [LM, 2017] for more details.
[17] See [LM, 2010] for more details.

Because a dictionary-based approach to analyzing Chinese retail texts is labor intensive and highly arbitrary, we use another, more promising approach—word embedding. With this method, words (and sentences and documents) are embedded as vectors in a fixed, high-dimensional vector space through the use of neural networks. The corpus we use for embedding words in the neural network and to perform sentiment detection includes posts from Chinese internet equity discussion forums.

Because there are no spaces to separate Chinese words in sentences, we use a computational linguistic technique called *segmentation* to break down a given sentence into words.[18] The segmentation technique we used computes all the possible meaning permutations for a given sentence by creating a directed acyclic graph (DAG).[19] The technique then computes all possible paths in the DAG and, using dynamic programming[20] techniques, chooses the path with the greatest probability.

Figure 5: An example of a DAG through a single Chinese sentence[21]



After segmentation, we now have individual Chinese words to work with. The next step is to convert these individual words as high-dimensional vector space vectors via embedding. The reason vectors are preferred over words is that vectors are mathematical objects, and after converting words to vectors, we can use well-developed mathematical tools to analyze them.

Vector space embedding is preferred over other NLP techniques because:

- It takes word order into consideration.
- It understands semantics. That is, it recognizes the context in which a given word appears.
- It can identify the right context for words that are spelled the same but have multiple meanings. For example, recognizing other words that appear near the word *bank*, the technique we used can identify if the reference is to a river bank or a financial institution.

[18] There are many approaches to Chinese language segmentation. For recent examples, see [SSD, 2017] and [DLZ, 2018].
[19] Directed acyclic graph: https://en.wikipedia.org/wiki/Directed_acyclic_graph
[20] Dynamic programming: https://en.wikipedia.org/wiki/Dynamic_programming
[21] Source: https://www.slideshare.net/ckmarkohchang/chinese-words-segmentation-tutorial

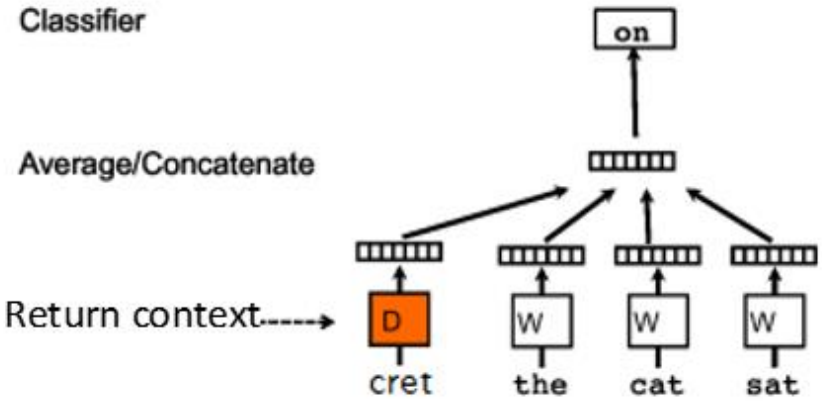
- The algorithm is unsupervised. That is, samples do not need to be labeled, classified, or categorized to train the neural network. This makes it more suitable for real-world applications.[22]

The standard vector space embedding approach does have a shortcoming, however, and that is its inability to distinguish words that are semantically similar yet have different meanings. As an example, in a financial context, *increase* and *decrease* are semantically similar words, but they have opposite meanings. In financial texts it is not uncommon to encounter the following two sentences:

- This stock has increased due to the company's operating results.
- This stock has decreased due to the company's operating results.

Because machine-learning algorithms can learn only from the data they are provided, the similarity of the above two sentences would lead the standard embedding algorithm to conclude that *increase* and *decrease* have similar meanings. To overcome this problem, we modified the standard approach to use a given stock's specific return as an additional context, so that words that are semantically similar yet different in meaning can be distinguished from each other. In high-dimensional vector space, what we did was to essentially increase the distance between vectors representing *increase* and *decrease*. Figure 6 below shows our neural network's topology.

Figure 6: PanAgora word-embedding neural network's topology [23]



Research Results

The following is a famous example from Google that illustrates a relationship in high-dimensional vector space:

King – Man + Woman ≈ Queen

[22] Unsupervised learning: https://en.wikipedia.org/wiki/Unsupervised_learning
[23] Source: PanAgora

Using our new algorithm, we found similar relationships for embedded Chinese word vectors. For example, Figure 7 shows that in the Chinese words vector space that we created, "floating red – rise + fall ≈ floating green." (Note, in the Chinese stock market, gains are colored red, while losses are colored green.)

Figure 7: Examples of Chinese word vector arithmetic[24]

Figure 8 shows a comparison between the embedded word *institution* and its most similar words, as measured by cosine similarity.[25] Note that our NLP algorithm learned that *institution* is similar to *large investors* and *hot money*. However, the algorithm also notes that it is similar to *chickens and dogs* and *chicken spirit*. That's because the Chinese word for *chickens and dogs* is phonetically similar to the word for *institution*, and the word for *chicken spirit* is phonetically similar to the word for *mutual funds*. Thus, our algorithm has learned Chinese cyber-slang.

Figure 8: We can learn Chinese cyber-slang using our word-embedding technique.[26]

Similarity Rank	机构 (Institution)	Cos Similarity
#1	大户 (large investors)	0.72
#2	游资 (hot money)	0.72
#3	鸡狗 (chickens & dogs)	0.66
#4	鸡精 (chicken spirit/essence)	0.62

Lastly, our modified network allows us to distinguish words that are semantically similar yet have different meanings. Figure 9 illustrates cosine similarity score for the words *rise* and *fall* as determined by standard word-embedding techniques on the left (without financial return context) and by our new technique on the right (with financial return context). Note that the new algorithm makes much finer distinctions between the two words (lower cosine similarity score) than does the standard vector-embedding technique.

[24] Source: PanAgora
[25] Cosine similarity: https://en.wikipedia.org/wiki/Cosine_similarity
[26] Source: PanAgora

Figure 9: Results from standard vector-embedding technique versus our modified technique[27]

Word Pair	Similarity
(涨, 跌)	0.733

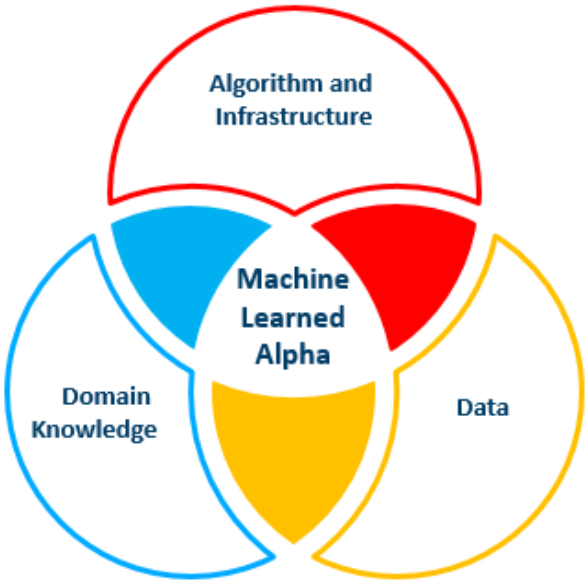
Word Pair	Similarity
(涨, 跌)	0.417

Summary

The Chinese stock market is large, liquid, and increasingly important on the global stage. It is also relatively inefficient and dominated by retail investors who exhibit behaviors and investment decisions that are hard to understand. This paper examines how to use NLP techniques to understand Chinese retail investor sentiments. Adapting the latest NLP technique and combining our informed insights into the financial domain with unique data, we can create algorithms to understand Chinese retail investor sentiments.

The research method discussed here is just one technique that PanAgora uses to understand the market. As shown in Figure 10, PanAgora believes that machine-learning expertise, domain knowledge, and data all must be combined in novel ways to extract alpha from markets around the world. In the coming months and years, machine-learning techniques will be one of many tools in our expanding toolbox that will help deliver long-term outstanding performance for our clients.

Figure 10: PanAgora's philosophy on machine learning[28]



[27] Source: PanAgora
[28] Source: PanAgora

References

[SL, 2018] Shrivastava R. and J. Lee, "Factor Investing in the China A-shares Market: Revelations from a Contextual Alpha Model." PanAgora white paper (2018).

[PCH, 2017] Peng H., E. Cambria, and A. Hussain, "A Review of Sentiment Analysis Research in Chinese Language." *Cognitive Computing* (2017): 9:423-435. http://sentic.net/chinese-sentiment-analysis-review.pdf

[LM, 2017] Loughran T., and B. McDonald, "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research*, Vol. 54, No. 4 (2016).

[LM, 2010] Loughran T., and B. McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *Journal of Finance* (2010).

[SSD, 2017] Sun Z., G. Shen, and Z. Deng, "A Gap-Based Framework for Chinese Word Segmentation via Very Deep Convolutional Networks." (2017).

[DLZ, 2018] Duan S., J. Li, and H. Zhao, "Fast Neural Chinese Word Segmentation for Long Sentences." (2018).

Legal Disclosures

The opinions expressed in this article represent the current, good faith views of the author(s) at the time of publication, are provided for limited purposes, are not definitive investment advice, and should not be relied on as such. The information presented in this article has been developed internally and/or obtained from sources believed to be reliable; however, PanAgora does not guarantee the accuracy, adequacy, or completeness of such information. Predictions, opinions, and other information contained in this article are subject to change continually and without notice of any kind and may no longer be true after the date indicated. Past performance is not a guarantee of future results. As with any investment there is a potential for profit as well as the possibility of loss. This material is directed exclusively at investment professionals.

Any investments to which this material relates are available only to or will be engaged in only with investment professionals.

PanAgora is exempt from the requirement to hold an Australian financial services license under the Corporations Act 2001 in respect of the financial services. PanAgora is regulated by the SEC under U.S. laws, which differ from Australian laws.