

# Deep Learning Extraction for Counterparty Risk Signals from a Corpus of Millions of Documents



---

**S&P Global**

# Deep Learning Extraction for Counterparty Risk Signals from a Corpus of Millions of Documents

## Abstract

China has been experiencing rapid growth over the last decade due to economically friendly reforms and a growing skilled and young population. With this increasing growth, China's interconnectedness with the global economy has increased significantly. In parallel to this economic evolution, technology has experienced rapid acceleration, which has enabled firms and governments to track and record vast amounts of data. The side effect of this unstructured big data growth is that datasets may be polluted, meaning information can be conflicting, missing, and/or unreliable. This creates a gap in the ability to provide transparency to the exposed firms importing from China: both timely early warning signals and wide coverage of small- and medium-sized enterprises (SMEs). We have been able to address this problem for our end-users by using deep learning\* to extract information value and opinion from a public corpus to create the needed transparency.

### Moody Hadi

Group Manager – Innovation & Product Research  
Risk Services  
S&P Global Market Intelligence

\*Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised, or unsupervised.

# Outline

- Overview of S&P Global
- Commercial Problem & Solution Overview
- Process Explained
  - Preprocessing
  - Training Set
  - Tokenization
  - Vectorization
  - Neural Networks
  - Benchmarking
  - Model Explanation
- Conclusion

# Overview of S&P Global

- Four major business lines:
  - S&P Global Ratings
  - S&P Global Indices
  - S&P Global Market Intelligence
  - S&P Global Platts
- Within S&P Global Market Intelligence
  - S&P Global Risk Services focuses on the varied form of risk management applications including:
    - Credit Risk
    - Market Risk
    - Operational Risk

# Commercial Problem & Solution Overview

# Current credit risk models primarily rely on financial statements

Financials are not always available or timely, so there is a need for alternative risk indicators

- Most firms, whether financial or non-financial corporates, have exposures to SMEs worldwide
- Current data sources include financial statements and trade payments
- These sources are not always available or timely
- The challenges are compounded in China, given language barriers and data availability
- There is a need for alternative timely risk signals

We are seeking useful, trusty, and timely first-party source information from companies with a broad reach

# A common problem identified by S&P Global Market Intelligence is that our clients need to review large bodies of unstructured data to find a domain-specific signal

- Problem:
  - Clients need to manage a significant amount of information to extract useful signals
  - The process today is manual, inefficient, and dependent on human expertise
  - While not specific to China, China presents its own idiosyncratic challenges
- Solution:
  - We started with Chinese mandatory company announcements:
    - In line with our strategic focus on China
    - Deals with language complexity
  - We created a robust technology solution to automate the processing, extraction, and opinion classification of documents
  - Our end-to-end workflow processes PDFs and extracts sentiment specifically around a company's financial health
  - The results are signals that clients can use to help sort through massive amounts of data, set their own notifications, and identify insights where they should focus their time
  - Domain specificity is handled by multiple modules in S&P Global Market Intelligence's offering within the same tech stack conditional on the domains being targeted
- The process is not intended to be a credit model:
  - First phase is to extract the opinion that a company discloses about the business within a financial health context
  - Second phase is to combine that opinion with credit scoring from their financial statements

# Systematically identifying real-time sentiment from announcements has many advantages

- Efficiency: A firm monitoring only 20 companies would need to sift through over 2,000 announcements and 33,000 pages\*
- Effectiveness: One individual reading 130 pages a day will see their performance reduce over time
- Timeliness: Identify signals more frequently in order to analyze information faster (as compared to financial statements that are only available once a year)

\*This data is in the public domain.



# CAESAR\* is an opinion extraction system focused on the Chinese Market.

CAESAR = Chinese Automated Extraction for Sentiment Analysis Reporting

- China is a unique market:
  - Documents are difficult to handle
  - Getting reliable information on the Chinese market is a unique and non-trivial problem
  - Difficult language for non-native speakers
- CAESAR handles the above by:
  - Optical Character Recognition (OCR) engine has image enhancements to handle noisy documents
  - Focused on a corpus of the most reliable documents in China, those directly from the companies' public announcements
  - Language difficulty is handled by a context-sensitive system:
    - OCR extraction on the Chinese corpus
    - Sentence and word detection on Chinese financial corpus, where words and sentence beginnings and endings are language dependent
    - The conversion of tokens via word2vec is Chinese financial health specific
- Reusable assets include:
  - A series of systems that cover the gamut of Natural Language Processing (NLP) from BagOfWords to Neural Networks
  - Context sensitive vectorization
  - A modular extraction system, where components can be replaced to handle other languages and contexts
  - Sentiment attribution to create efficiencies for the analyst, as well as transparency and self-consistency\*\* of the extraction
  - Simplifies localization to the end-users
  - Internal know-how and experience

# Detect and reflect: From document to sentiment

Our first phase is to transform unstructured text into sentiment signals relating to financial health of companies

Large body of PDFs



Structured data

level	page_num	block_num	par_num	line_num	word_num	left	top	width	height	conf	text
1	1	0	0	0	0	0	0	2480	3507	-1	
2	1	1	0	0	0	1656	185	445	49	-1	
3	1	1	1	0	0	1656	185	445	49	-1	
4	1	1	1	1	0	1656	185	445	49	-1	
5	1	1	1	1	1	1656	185	123	48	93	公称
5	1	1	1	1	2	1779	185	89	49	93	编号
5	1	1	1	1	3	0	0	2480	3507	71	:
5	1	1	1	1	4	1905	193	196	36	71	2018-001
2	1	2	0	0	0	377	307	1703	419	-1	
3	1	2	1	0	0	377	307	1703	419	-1	
4	1	2	1	1	0	377	307	1703	50	-1	
5	1	2	1	1	1	377	307	96	49	93	证
5	1	2	1	1	2	0	0	2480	3507	79	券
5	1	2	1	1	3	479	308	94	48	92	代码
5	1	2	1	1	4	582	311	9	21	88	:
5	1	2	1	1	5	627	315	145	36	91	839771
5	1	2	1	1	6	979	310	47	42	93	证
5	1	2	1	1	7	1032	307	44	49	92	券
5	1	2	1	1	8	1083	308	91	48	92	编
5	1	2	1	1	9	1185	331	9	21	91	称
5	1	2	1	1	10	0	0	2480	3507	34	:
5	1	2	1	1	11	1231	308	45	47	26	称
5	1	2	1	1	12	1281	308	46	47	72	编

Sentiment

相关提及“煤飞色舞”再现沪指走高 适当关注成长性好的创业板 对于目前行情,华龙证券投资顾问郑红伟认为,投资者无需过多关注指数表现,积极把握中小创业板中成长性良好的个股。相关提及 采购助理个人工作总结三篇 重点采购的大项目有:健桥证券北京营业部UPS、发电机设备;健桥证券丹阳营业部建设项目所需设备;华龙证券日常设备供应;湖南省农行小型机以及附属设备;银河证券西安集中交易所需设备;项目以及华龙证券设备供应等采购任务。相关提及 创业板先礼后兵 领跌变领涨 华龙证券李小刚:创业板指在前期高点附近有回落整理的要求,做好调仓换股的准备。亏损盈利 低位震荡沪指收跌 总体向上状态未遭破坏 华龙证券兰州合水路营业部 投资顾问狄红梅认为,目前市场总体向上的状态并未遭到破坏,周三创业板的上涨充分说明了市场资金风格的转化,投入相对超跌的中小品种,且创业板虽遇到半年线压力出现回落,但其下方均线已形成明显多头排列,其短期在量能不能立即放大的情况下不会出现大涨,但下跌空间也不大。

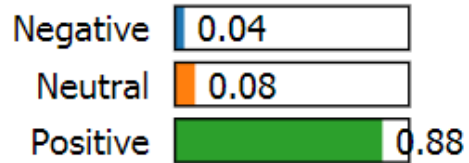


# We are identifying relevant pieces of text and its embedded opinion

## Context Sensitive Extraction, Attribution, and Actionable Notifications via Sentiment Analysis

True class is: Positive

Prediction probabilities



### Text with highlighted words

证券代码: ⓪ 证券简称: 力尊信通 公告编号  
主办券商: 中航证券 北京力尊信通科技股份有限公司关于获得高新技术企业证书的公告  
本公司及董事会全体成员保证公告内容的真实、准确和完整, 没有虚假记载、误导性陈述或者重大遗漏, 并对其内容的真实性、准确性和完整性承担个别及连带法律责任  
北京力尊信通科技股份有限公司(以下简称“公司”)于 ⓪年 ⓪月 ⓪日收到由北京市科学技术委员会、北京市财政局、北京市国家税务局、北京市地方税务局联合颁发的《高新技术企业证书》(证书编号: α⓪), 发证时间: ⓪年 ⓪月 ⓪日, 有效期三年  
根据相关规定, 公司自通过高新技术企业认定并向主管税务机关办理完减免税手续后三年内, 可享受国家关于高新技术企业的相关优惠政策, 即按 ⓪税率缴纳企业所得税

Positive Connotation – Green  
Negative Connotation – Blue

### Google Translate:

Stock code: 证券 Securities abbreviation: Lizun Xintong Bulletin number

Sponsored Broker: AVIC Securities Beijing Lizun Xintong Technology Co., Ltd. Announcement on Obtaining High-tech Enterprise Certificate

The company and all members of the board of directors guarantee the truthfulness, accuracy and completeness of the contents of the announcement, and there are no false records, misleading statements or major omissions, and bear individual and joint legal responsibility for the truthfulness, accuracy and completeness of the contents.

Beijing Lizun Xintong Technology Co., Ltd. (hereinafter referred to as “the company”) was jointly awarded by the Beijing Municipal Science and Technology Commission, the Beijing Municipal Finance Bureau, the Beijing Municipal State Taxation Bureau and the Beijing Municipal Local Taxation Bureau on the following day. “High-tech Enterprise Certificate” (Certificate No.: α⓪), issuance time: ⓪ Year ⓪ Month Day, valid for three years

According to the relevant regulations, the company can enjoy the relevant preferential policies of the state on high-tech enterprises within three years after passing the high-tech enterprise certification and handling the tax reduction and exemption procedures with the competent tax authorities, that is, paying the enterprise income tax at the tax rate.

**Note:** the translation is provided for context only. The process deals with the native language announcements.

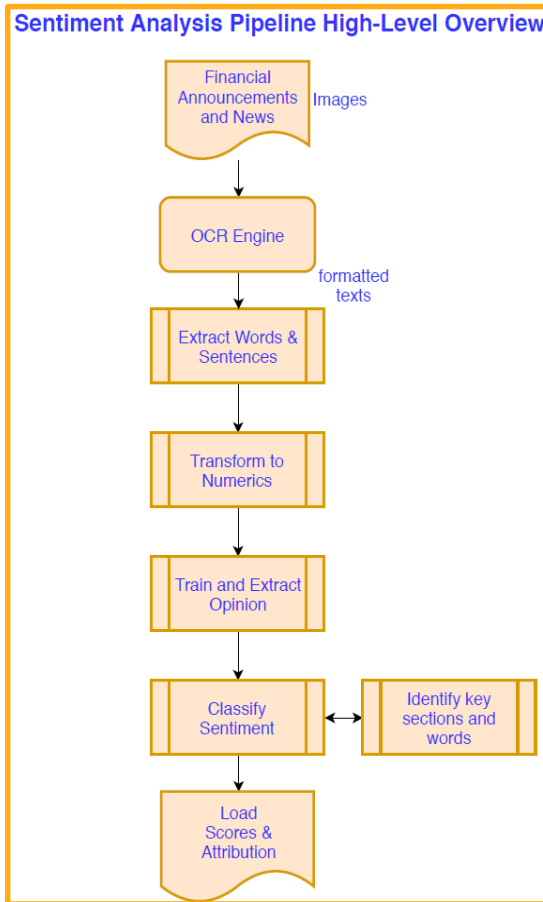
# Data set

We selected a broad set of company announcements

- There is a large body of public disclosures by Chinese companies
- These disclosures are mandatory and factual
- They cover:
  - Firmographics: e.g., board and management resolutions
  - Legal changes: e.g., legal opinions, litigations
  - Financial information: e.g., annual reports, shareholder meetings
- These factual and timely disclosures can be used to extract valuable signals about the financial health of a company

# Process

The end-to-end process accesses, parses, processes, and analyzes announcements to extract each article's embedded opinion on the financial health of a company – sentiment classification



- Automated collection of announcement documents
- Transform unstructured to structured text via OCR and document analysis
- Pre-process the structured body of text to define sentences, paragraphs, and sections
- Vectorization model
- Model training and validation
- Sentiment classification model
- Sentiment indicators, visualization

# Summary of Technical Considerations for NLP techniques

## Technical considerations

Approach	Advantage	Disadvantage
Bag of Words	Deterministic Tractable Transparent (Simple Attribution)	Dictionary Maintenance Hard to capture sequencing High Bias Context insensitive
Regularized Bag of Words	More targeted to extract veracity Deterministic Tractable Transparent (Simple Attribution) Can control context sensitivity	Less Transparent Requires Recalibration/Higher Maintenance Target Variables need to be defined
Neural Networks	Captures Sequencing Context sensitive Low Bias (highly adaptable) Reinforcement allows for automated re-training	Low Transparency (black box) Training is complex (computationally... eg : H/W, Algo)

# Summary of Business Considerations for NLP techniques

## Business considerations

Use Case	Advantage	Disadvantage
Public opinions on companies (e.g. forums, surveys, Twitter....)	<ul style="list-style-type: none"> <li>Public data (no privacy considerations)</li> <li>Easy to understand the output</li> <li>Attribution</li> <li>Complements existing scoring methods</li> </ul>	<ul style="list-style-type: none"> <li>Large body of prior work</li> <li>Relatively new to extract veracity</li> <li>Public opinion data is suspect</li> </ul>
Public disclosures directly from "informed audience" (e.g. company, analyst research ...)	<ul style="list-style-type: none"> <li>More reliable data - easier to extract information value</li> <li>Provides wider coverage for companies that do not disclose financials or generally have unreliable information</li> <li>Public data (no privacy considerations)</li> <li>Easy to understand the output</li> <li>Attribution</li> <li>Complements existing scoring methods</li> <li>Creates new methods for classification given a target</li> </ul>	<ul style="list-style-type: none"> <li>To extract veracity more complex approaches are required</li> <li>Target variable needs to be defined (e.g.. credit ranking ...)</li> <li>Non-deterministic approaches are complex computationally</li> </ul>

# Process Explained



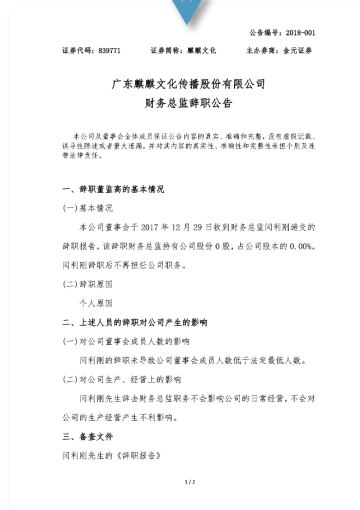
# Pre-processing: A PDF to enhanced images

Image enhancements include removing borders, background colors, and adjusting brightness and contrasts of the text.

Example of a good quality source PDF file (PDF that is constructed from a text document)



Image enhancement routine



Example of a medium quality source PDF file (scanned pages that have borders, shadows, and blurred text)

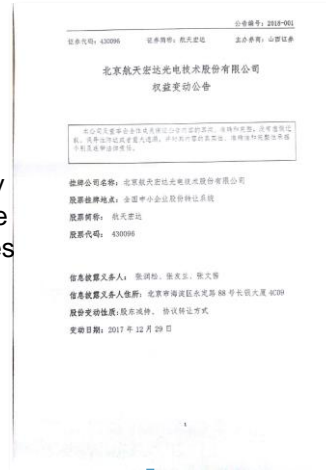
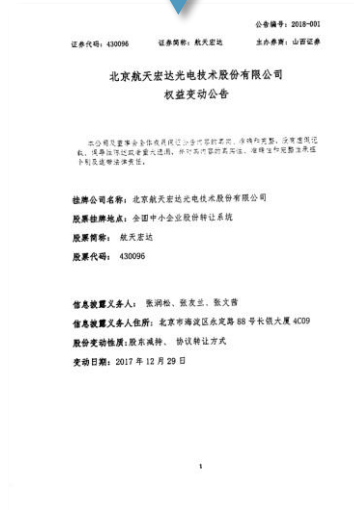


Image enhancement routine



Example of a poor quality source PDF file (low quality scanned pages, with noisy background and text)

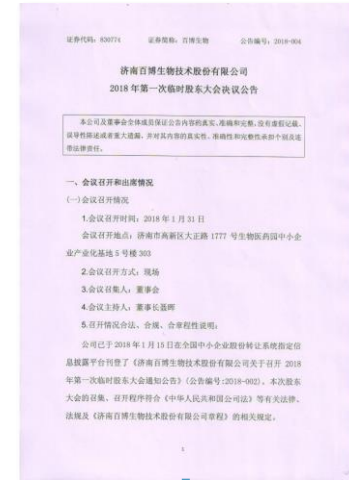
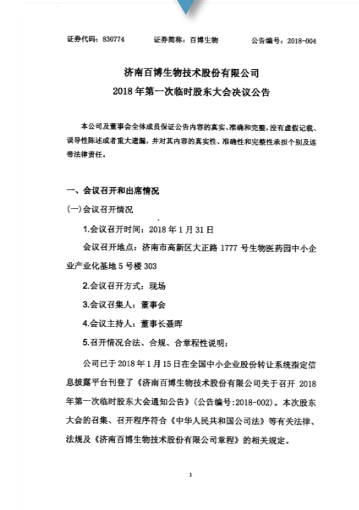
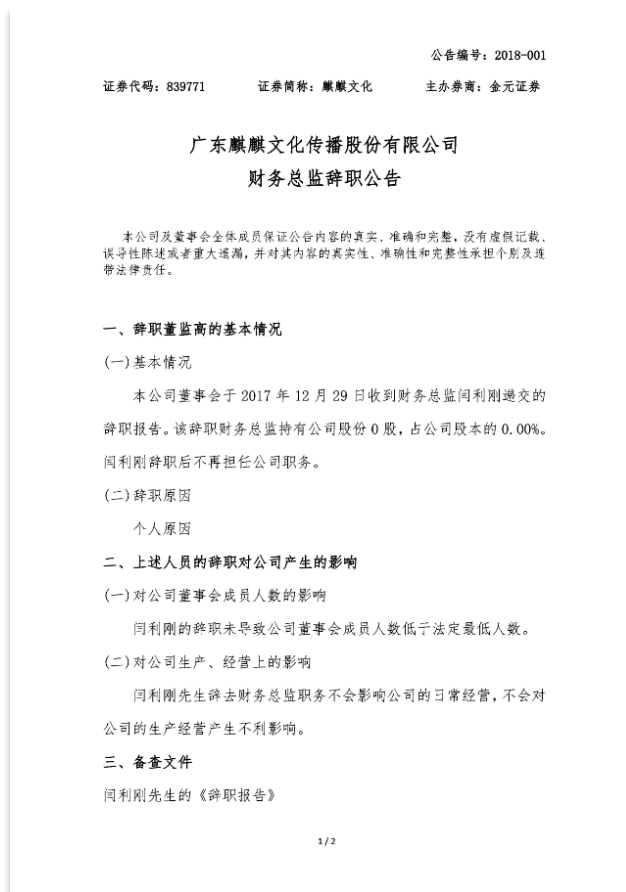


Image enhancement routine



# Pre-processing: Enhanced images to 'structured' text

Enhanced images are converted to a table with coordinates of each token (page, block, paragraph, and line number, in addition to the position and width and height of each token)



level	page_num	block_num	par_num	line_num	word_num	left	top	width	height	conf	text
1	1	0	0	0	0	0	0	2480	3507	-1	
2	1	1	0	0	0	1656	185	445	49	-1	
3	1	1	1	0	0	1656	185	445	49	-1	
4	1	1	1	1	0	1656	185	445	49	-1	
5	1	1	1	1	1	1656	185	123	48	93	公告
5	1	1	1	1	2	1779	185	89	49	93	编号
5	1	1	1	1	3	0	0	2480	3507	71	:
5	1	1	1	1	4	1905	193	196	36	71	2018-001
2	1	2	0	0	0	377	307	1703	419	-1	
3	1	2	1	0	0	377	307	1703	419	-1	
4	1	2	1	1	0	377	307	1703	50	-1	
5	1	2	1	1	1	377	307	96	49	93	证
5	1	2	1	1	2	0	0	2480	3507	79	券
5	1	2	1	1	3	479	308	94	48	92	代码
5	1	2	1	1	4	582	331	9	21	88	:
5	1	2	1	1	5	627	315	145	36	91	839771
5	1	2	1	1	6	979	310	47	42	93	证
5	1	2	1	1	7	1032	307	44	49	92	券
5	1	2	1	1	8	1083	308	91	48	82	简
5	1	2	1	1	9	1185	331	9	21	91	称
5	1	2	1	1	10	0	0	2480	3507	34	:
5	1	2	1	1	11	1231	308	45	47	26	麒
5	1	2	1	1	12	1281	308	46	47	72	腊



# Pre-processing: Structured text to text without tables and headers/footers

Tables should be removed from the documents. Using the constructed structured text, we are able to extract all tables from documents with circa 100% accuracy.

Additionally, all headers and footers (like company logo, title of document, legal disclaimer, etc.) that are being repeated across all pages and have no information value have been extracted from the documents.

公告编号: 2018-001

宏达工作,任公司董事、董事会秘书,任期三年。

二、信息披露义务人一致行动关系说明

张润松为公司控股股东,张润松和张友兰为夫妻关系,张文蕾和张润松为父女关系,张文蕾和张友兰为母女关系。

三、信息披露义务人持股情况

截至2017年12月29日(本次转让前),信息披露义务人持股情况如下:

信息披露义务人	股份名称	股份种类	股份数量(股)	股份比例(%)	所持股份性质及性质变动情况	限售股变动达到规定比例日期	权益变动方式
张润松	航天宏达	人民币普通股	3,104,000	47.75%	流通股7,328,000股,可转让股份776,000股	2017年12月29日	协议转让
张友兰	航天宏达	人民币普通股	956,812	14.72%	流通股717,609股,可转让股份239,203股	-	-
张文蕾	航天宏达	人民币普通股	132,000	2.03%	流通股99,000股,可转让股份33,000股	-	-
合计			4,192,812	64.50%			

6

公告编号: 2018-001

宏达工作,任公司董事、董事会秘书,任期三年。

二、信息披露义务人一致行动关系说明

张润松为公司控股股东,张润松和张友兰为夫妻关系,张文蕾和张润松为父女关系,张文蕾和张友兰为母女关系。

三、信息披露义务人持股情况

截至2017年12月29日(本次转让前),信息披露义务人持股情况如下:

信息披露义务人	股份名称	股份种类	股份数量(股)	股份比例(%)	所持股份性质及性质变动情况	限售股变动达到规定比例日期	权益变动方式
张润松	航天宏达	人民币普通股	3,104,000	47.75%	流通股7,328,000股,可转让股份776,000股	2017年12月29日	协议转让
张友兰	航天宏达	人民币普通股	956,812	14.72%	流通股717,609股,可转让股份239,203股	-	-
张文蕾	航天宏达	人民币普通股	132,000	2.03%	流通股99,000股,可转让股份33,000股	-	-
合计			4,192,812	64.50%			

6

width	height	conf	text
0	2480	3509	-1
07	431	46	-1
07	431	46	-1
07	431	46	-1
07	139	46	95 公告
10	58	43	93 编号
0	2480	3509	89 ;
15	188	37	89 2018-001
01	1676	17	-1
01	1676	17	-1
4	1	2	1 1 0 398 301 1676 17 -1
5	1	2	1 1 1 398 301 1676 17 95
2	1	3	0 398 431 1193 6 -1
3	1	3	1 0 398 431 1193 6 -1
4	1	3	1 1 0 398 431 1193 6 -1
5	1	3	1 1 1 398 431 1193 6 95
2	1	4	0 2072 320 3 130 -1
3	1	4	1 0 2072 320 3 130 -1

# Pre-processing: Cleaned documents to 'sectioned' documents

## Section 1:

doc_id	page_num	doc_size	doc_type	doc_status	doc_date	doc_title	doc_content	doc_keywords	doc_topics	doc_sentiment	doc_sentiment_score	doc_sentiment_label
1.1	1	0	0	0	1	0	240	357	-1			
2.1	1	0	0	0	2018-01-01	0	2018-01-01	45	-1			
3.1	1	1	0	0	2018-01-01	0	2018-01-01	45	-1			
4.1	1	1	1	1	1	1	104	185	45	-1		
5.1	1	1	1	1	1	1	104	185	45	-1		

## Section 2:

doc_id	page_num	doc_size	doc_type	doc_status	doc_date	doc_title	doc_content	doc_keywords	doc_topics	doc_sentiment	doc_sentiment_score	doc_sentiment_label
1.1	1	1	1	1	1	1	104	185	45	-1		
2.1	1	1	1	1	1	1	104	185	45	-1		
3.1	1	1	1	1	1	1	104	185	45	-1		
4.1	1	1	1	1	1	1	104	185	45	-1		
5.1	1	1	1	1	1	1	104	185	45	-1		

## Section n:

doc_id	page_num	doc_size	doc_type	doc_status	doc_date	doc_title	doc_content	doc_keywords	doc_topics	doc_sentiment	doc_sentiment_score	doc_sentiment_label
1.1	1	0	0	0	1	0	240	357	-1			
2.1	1	0	0	0	2018-01-01	0	2018-01-01	45	-1			
3.1	1	1	0	0	2018-01-01	0	2018-01-01	45	-1			
4.1	1	1	1	1	1	1	104	185	45	-1		
5.1	1	1	1	1	1	1	104	185	45	-1		

Document Type	Value	Why?
Long documents (>=100 pages)	Dividing the document into chapters allows for identification of meaningful sections.	<ul style="list-style-type: none"> <li>Provides transparency and self-consistency to the end-user on the reasoning of why a document has been assigned a certain classification.</li> <li>Allows the end-user to preform actions with the notifications. They can "skim-read" large documents, especially when they have a vast body to review.</li> </ul>
Short documents	Identifying the key words and phrases that had the highest influence on the assigned classification allows for easy to understand reasoning of the classification.	<ul style="list-style-type: none"> <li>Provides transparency and self-consistency to the end user on the reasoning of why a document has been assigned a certain classification.</li> <li>Allows for easy interpretation of the classification, similar to the deterministic approach (Bag of Words).</li> </ul>

公告编号: 2018-002  
证券代码: 839771 证券简称: 顺帆文化 主办券商: 金元证券

**广东顺帆文化传播股份有限公司**  
财务总监辞职公告

本公司及董事会全体成员保证公告内容的真实、准确和完整,没有虚假记载、误导性陈述或者重大遗漏,并对其内容的真实性、准确性和完整性承担个别及连带责任。

一、辞职董事的基本情况

(一)基本情况

本公司董事会于2017年12月29日收到财务总监李益国先生递交的辞职报告。该辞职财务总监李益国先生持有公司股份0股,占公司股本的0.00%,同利刚辞职后不再担任公司职务。

(二)辞职原因

个人原因

二、上述人员的辞职对公司产生的影响

(一)对公司董事会成员人数的影响

同利刚的辞职未导致公司董事会成员人数低于法定最低人数。

(二)对公司生产、经营上的影响

同利刚先生辞去财务总监职务不会影响公司的日常经营,不会对公司的生产经营产生不利影响。

三、备查文件

同利刚先生的《辞职报告》

doc_id	page_num	doc_size	doc_type	doc_status	doc_date	doc_title	doc_content	doc_keywords	doc_topics	doc_sentiment	doc_sentiment_score	doc_sentiment_label
1.1	1	0	0	0	1	0	240	357	-1			
2.1	1	0	0	0	2018-01-01	0	2018-01-01	45	-1			
3.1	1	1	0	0	2018-01-01	0	2018-01-01	45	-1			
4.1	1	1	1	1	1	1	104	185	45	-1		
5.1	1	1	1	1	1	1	104	185	45	-1		

➔

业产业化基地5号楼303

2.会议召开方式: 现场

3.会议召集人: 董事会

4.会议主持人: 董事长李基理

5.召开情况合法、合规、合章程性说明:

公司已于2018年1月15日在全国中小企业股份转让系统指定信息披露平台刊登了《济南百博生物技术股份有限公司关于召开2018年第一次临时股东大会通知公告》(公告编号:2018-002)。本次股东大会的召集、召开程序符合《中华人民共和国公司法》等有关法律、法规及《济南百博生物技术股份有限公司章程》的相关规定。

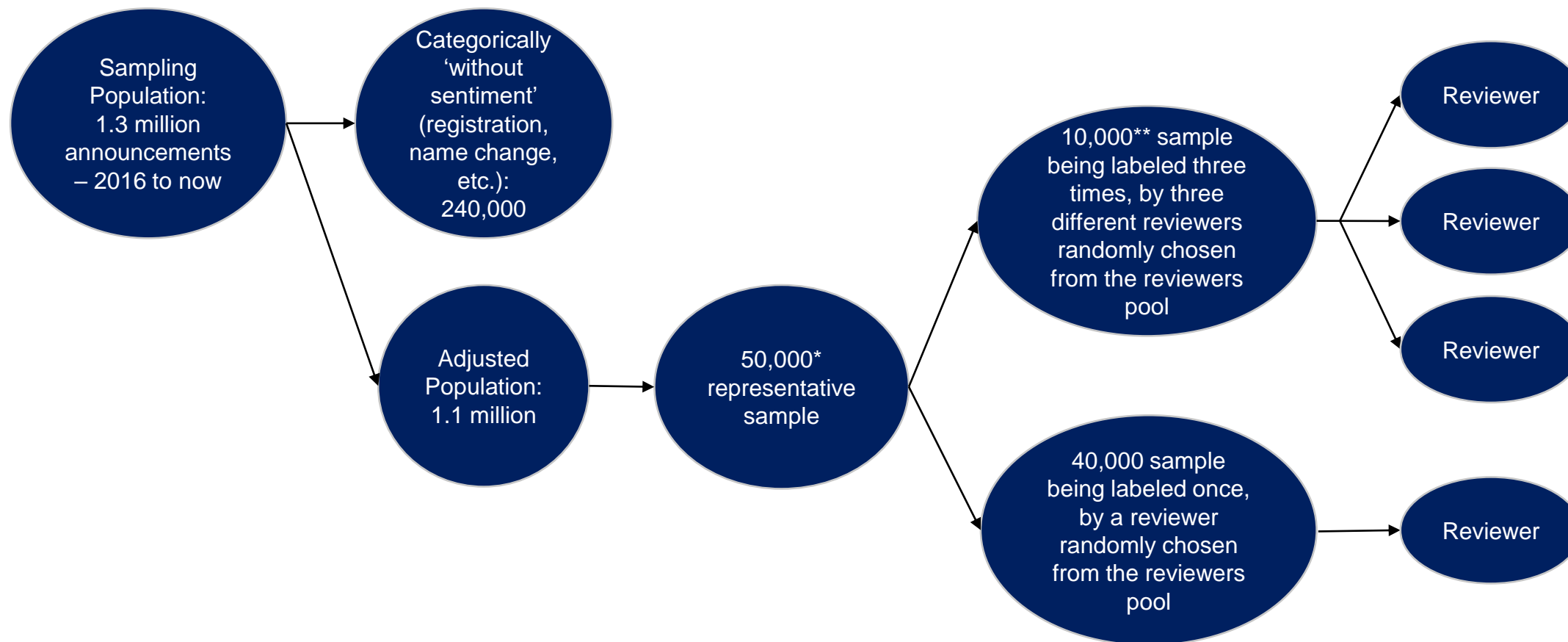
姓名	职务	人民币普通股	持股比例	流通股股份	可转让股份
张友忠	财务总监	956,812	14.72%	717,609股,可转让股份239,203股	-
张文普	财务总监	132,000	2.03%	132,000股,可转让股份33,000股	-
合计		4,192,812	64.50%		

S&P Global

Permission to reprint or distribute any content from this presentation requires the prior written approval of S&P Global Market Intelligence. Not for distribution to the public. Copyright © 2018 by S&P Global Market Intelligence. All rights reserved.

18

# Training set (end state)



\* First training set is 5,500 articles: we are adding 50,000 to the sample (in progress)

\*\* 1,000 of the original 5,500 articles were also shared with a second group

Labels	
- Positive	- Without Sentiment
- Negative	- Illegible
- Neutral	

# Building the end state training set

The following steps have been taken to build the training set:

1. Interview and select qualified native Chinese financial professionals, with experience in risk management and assessment of Chinese companies.
2. Identify categories of announcements that are 'without sentiment' by nature (e.g. legal registration, name change, etc.) based on reviewers feedback.
3. Sample 50,000 announcements from the population, after removing items identified in Step 2. Samples are randomly drawn from year 2016 to present, while maintaining the overall distribution of announcements across categories.
4. From the 50,000 sampled announcements, randomly choose 10,000 of them as a 'hold-out'\* sample. Assign each hold-out sample to three random reviewers, in order to control quality the human reviews.
5. Send sectioned documents to reviewers and receive back labels for each document under the following categories: positive, negative, neutral, without sentiment, and illegible.

\*The hold-out sample is a sample of data not used in fitting a model; used to assess performance.

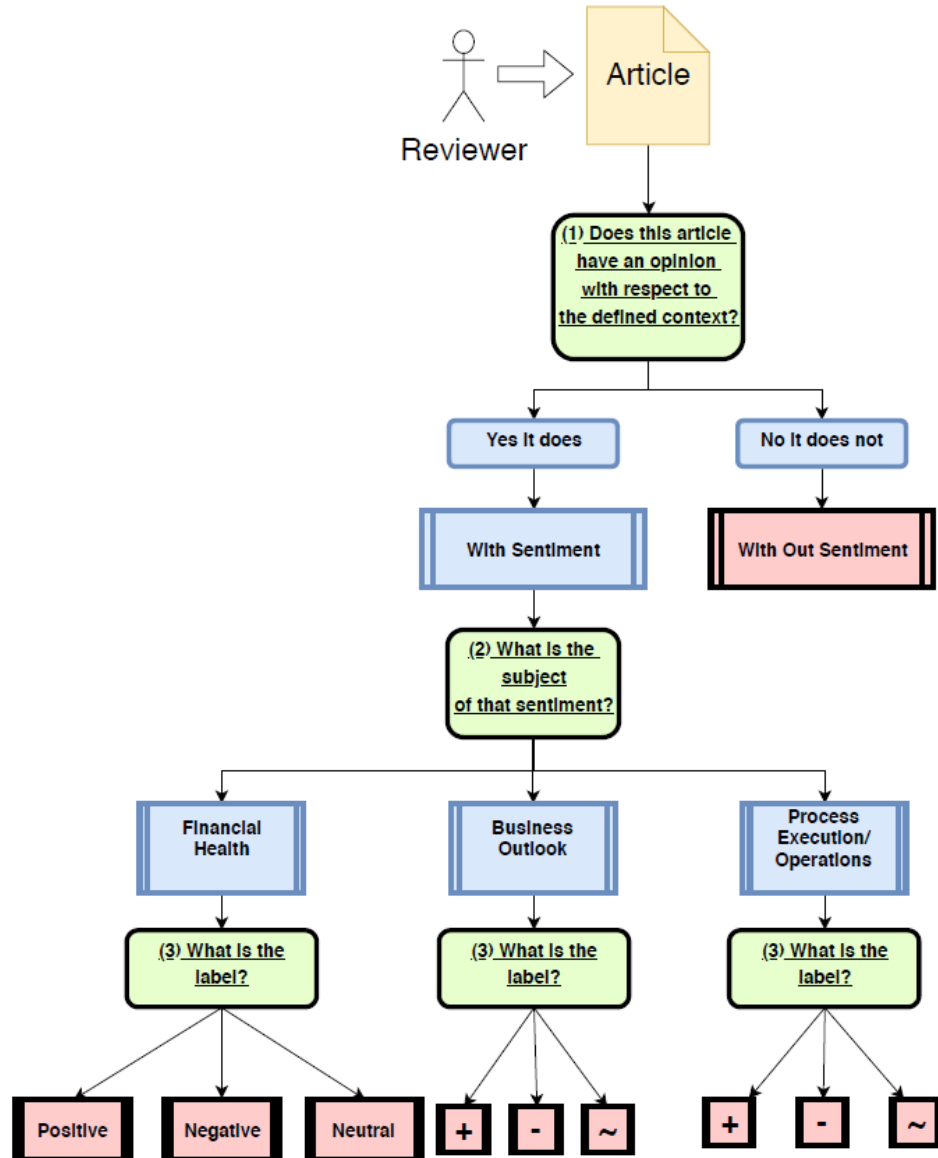
# Definitions for training set design and labeling

Objective:

This product extracts and then labels the company's explicit message in its own disclosures on activities that relate to the following context (a group of subjects):

- *Financial Health:*
  - This is the company's opinion on the subject of financial risk and health. In its opinion (if it exists), is there financial risk on liquidity i.e. the company's ability to pay future obligations?
  - This covers connotation on its capital structure, revenues, assets, liabilities.
  - Meritable observations on its financials, for example has there been fraud...?
- *Business Outlook:*
  - This is the company's opinion (if it exists) on the subject of its business model.
  - Business model activities are strategic and tactical activities, such as new products, partnerships/alliances.
- *Process Execution:*
  - This is the company's opinion (if it exists) on the subject of its business process execution.
  - These are operational activities that have no impact on the strategic or tactical activities.

# Reviewer labeling process flow



- Simple examples of *extracting* a label:
  - “Our operations are strong and we see good growth forward ...” →Positive
  - “Our operations are proceeding as expected ...” →Neutral
  - “Our operations had weakness and the market is softening ...” →Negative
  - “Our operations are provided in this statement and we certify ...” →Without Sentiment
- For short articles, does the overall connotation in the article, while following the labeling process flow, have a label?
- For long articles, we section them and ask the reviewers to follow the above label process flow to define a label to the section.



# Tokenization: Sentence and word tokenization

Input to the models are sequences of sentences and words:

- Sentence Tokenization: From the text, identify the beginning and end of sentences.
- Word Tokenization: Break each sentence into sequences of words.

深圳市嘉兰图设计股份有限公司关于子公司破产申请已被法院受理的公告。本公司及董事会全体成员保证公告内容的真实、准确和完整，没有虚假记载、误导性陈述或者重大遗漏，并对其内容的真实性、准确性和完整性承担个别及连带法律责任。



```
[  
["深圳市", "嘉兰图", "设计", "股份", "有限公司", "关于", "子公司", "破  
产", "申请", "已", "被", "法院", "受理", "的", "公告"],  
  
["本", "公司", "及", "董事会", "全体成员" "保证", "公告", "内容", "的", "  
真实", "准确", "和", "完整", "没有", "虚假", "记载", "误导性", "陈述", "  
或者", "重大", "遗漏", "并", "对", "其", "内容", "的", "真实性", "准确性  
", "和", "完整性", "承担", "个别", "及", "连带", "法律责任"]  
]
```

# Vector representation of words

Once we have converted our text into sequences of sentences and words, we need to turn these sequences into numerical vectors. There are two main methods for this:

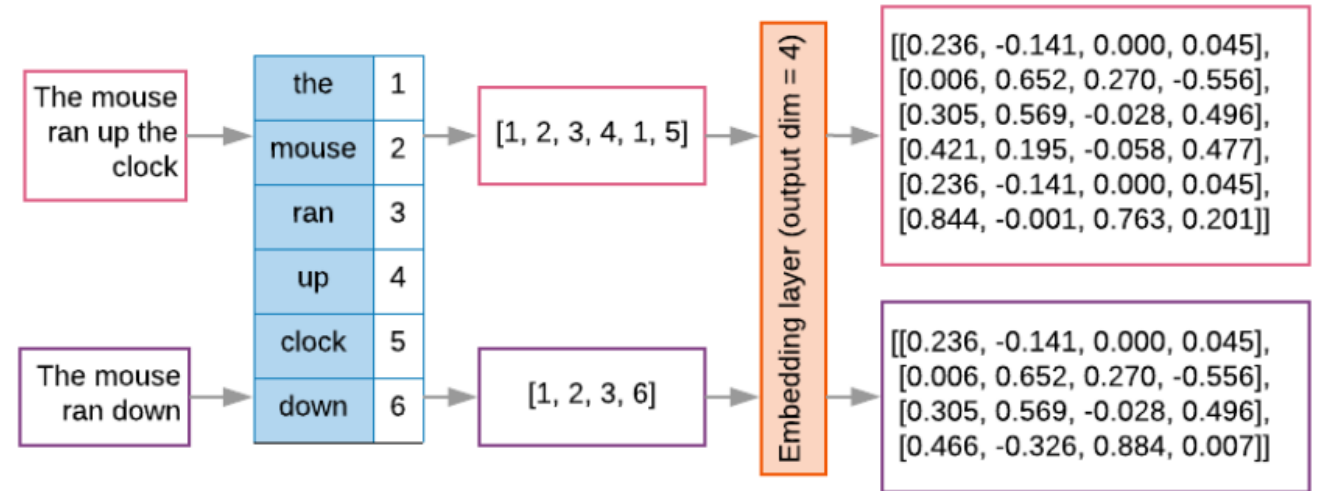
## 1. One-hot encoding

This representation works when vocabulary is small and so not useful, especially for the Chinese language.

```
'The mouse ran up the clock' = [  
  [0, 1, 0, 0, 0, 0, 0],  
  [0, 0, 1, 0, 0, 0, 0],  
  [0, 0, 0, 1, 0, 0, 0],  
  [0, 0, 0, 0, 1, 0, 0],  
  [0, 1, 0, 0, 0, 0, 0],  
  [0, 0, 0, 0, 0, 0, 1],  
]
```

## 2. Word embeddings

Model learns to turn word index sequences into vectors during the training process, such that each word index gets mapped to a dense vector of real values representing that word's location in semantic space.



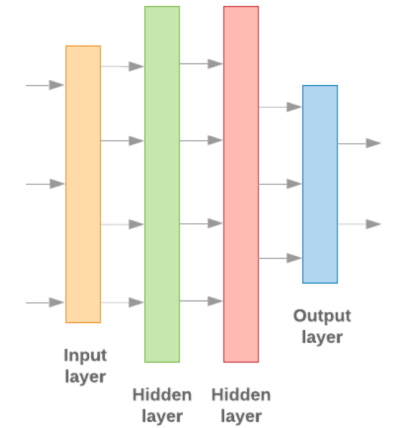
# Neural Networks

Neural Networks are nothing but a series of linear transformations from inputs to outputs.

For this system, we have assessed alternative methodologies and chose a framework that has two main components:

- Convolutional Neural Networks (CNN): They learn local patterns of the input sequence. However, they have no memory. Each input shown to them is processed independently, with no state kept in between inputs.
- Long Short-Term Memory (LSTM) networks: They iterate over the input sequence, and ‘remember’ what they have seen so far.

Our analysis shows that combining CNN and LSTM networks gives the system the ability to learn complexities of the Chinese language.



# Neural Networks - limitations

Neural Networks are revolutionizing our day-to-day life. However, they have two main drawbacks:

## 1. Lack of transparency:

- Utilized method for overcoming it:
  - Assess the effect of changing the model's input on its output. Effectively, utilize sensitivity analysis to understand how the model works and its main drivers

## 2. Overfitting the training set:

- Utilized methods for overcoming it:
  - Maintaining a hold-out validation set
  - Drop outs: randomly *dropping out* (setting to zero) a number of output features of the layer during training
  - K-fold cross validation (split data into K partitions of equal size. For each partition  $i$ , train a model on the remaining  $K - 1$  partitions, and evaluate it on partition  $i$ )

# Benchmarking the approach

Benchmarking a complex workflow and deep learning is challenging – it is context and language sensitive. We are focusing on the process and training set.

Approaches we leveraged

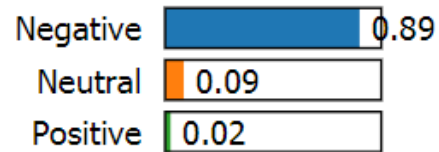
Approach	Result
Bag of Words on English translation	~40% accuracy
FaceBook FastText vectorization (native Chinese – non financial context)	~ 30% accuracy
Strong performing Deep Learning model	60%+ is considered a strong model

- Chosen architecture, compared to alternative architectures, shows robust out-of-sample performance. Additionally, results of conducted sensitivity analysis shows model is self consistent via Local Interpretable Model-agnostic Explanations (LIME).
- Performance will be monitored through capturing users' feedback within the User Interface, and quarterly performance monitoring against a challenger architecture (Bag of Words approach).
- Re-training is deterministic in case of significant divergence in model performance against the challenger systems and based on clients' feedback, or an introduction of a new announcement category by data provider not seen during the training.

# UI: Transparency via LIME to show the Extraction Algorithm is Self Consistent

True class is: Negative

Prediction probabilities



## Text with highlighted words

证券代码: ⊙ 证券简称: 中艺股份 公告编号  
 南京中艺建筑设计院股份有限公司关于收到扬州市城乡建设局行政处罚决定书的公告  
 本公司及董事会全体成员保证公告内容的真实、准确和完整,没有虚假记载、误导性陈述或者重大遗漏,并对其内容的真实性、准确性和完整性承担个别及连带法律责任  
 南京中艺建筑设计院股份有限公司(以下简称“公司”)近日收到扬州市城乡建设局开具的《行政处罚决定书》(扬建罚  
 扬州市城乡建设局城市建设监察支队调查发现公司未 按工程建设强制性标准(条文)对承揽的扬子江世贸商业综合体工程进行设计,该行为违反了《建设工程质量管理条例》第十九条第一款  
 扬州市城乡建设局依据《建设工程质量管理条例》第六十三条第一款第(四)项以及第七十三条对当事人进行处罚  
 鉴于该工程尚未开工,公司积极配合行政机关查处违法行为并整改完成,主动消除违法行为危害后果,根据《中华人民共和国  
 公告编号: ⊙ 根据《中华人民共和国行政处罚法》第二十七条第一款的规定,给予公司减轻行政处罚  
 综上,扬州市城乡建设局作出对中艺公司处⊙元罚款、对直接负责的主管人员王刚处⊙元罚款、对其他直接责任人员祁岭处⊙元罚款的行政处罚  
 公司将根据《行政处罚决定书》的规定,在缴款期限内 足额缴纳人民币伍万元罚款,该笔罚款占营业收入比重较小,情节较轻  
 此外,该工程尚未开工,我公司积极配合行政机关查处违法行为并整改完成,主动消除了危害后果,本次事件不会对公司经营产生重大影响  
 公司管理层高度重视本次事件:由总经理召开紧急会议,进行全公司通告,对相关责任人进行严肃地批评和处罚,督促全公司认真吸取本次行政处罚的教训,积极整改;责令公司员工对公司承接的所有项目进行全面排查,避免此类问题再次发生;总结本次事项的教训,制定预防措施,特别是进一步加强对此类外地工程设计项目成果质量的严格把控,定期对公司员工进行工程建设强制性标准及相关法律法规的培训  
 扬州市城乡建设局行政处罚决定书》(扬建罚

Positive Connotation – Green  
 Negative Connotation – Blue

## Google Translate:

Stock code: 证券 Securities abbreviation: Zhongyi Shares Announcement No.

Announcement of Nanjing Zhongyi Architectural Design Institute Co., Ltd. on Receiving the Administrative Penalty Decision of Yangzhou Urban and Rural Construction Bureau

[...]

Nanjing Zhongyi Architectural Design Institute Co., Ltd. (hereinafter referred to as “the company”) recently received the “Administrative Punishment Decision” issued by Yangzhou Urban and Rural Construction Bureau.

[...]

Yangzhou Urban and Rural Construction Bureau will impose penalties on the parties in accordance with Article 63, paragraph 1 (4), and Article 73 of the Regulations on Quality Management of Construction Projects.

[...]

Announcement No.: 规定 The provisions of the first paragraph of Article 27 of the Administrative Punishment Law of the Republic give the company less administrative penalties

In summary, the Yangzhou Urban and Rural Construction Bureau made a fine for the Chinese company, a fine for the directly responsible person, Wang Gang, and a fine for other directly responsible personnel.

According to the “Administrative Punishment Decision”, the company will pay a fine of RMB 10,000 in full within the payment period. The fine accounts for a small proportion of the operating income, and the plot is lighter.

In addition, the project has not yet started, and our company actively cooperates with the administrative organs to investigate and deal with illegal acts and complete the rectification, and actively eliminates the harmful consequences.

This incident will not have a major impact on the company's operations.

The management of the company attaches great importance to this incident: the general meeting convened an emergency meeting, issued a company-wide notice, seriously criticized and punished the relevant responsible persons, urged the whole company to seriously absorb the lessons of this administrative punishment, and actively rectified;

ordered the employees of the company Conduct a comprehensive investigation of all the projects undertaken by the company to avoid such problems;; and develop preventive measures, especially to further strengthen summarize the lessons learned from this matter the strict control of the quality of such field engineering design projects, and regularly to the employees of the company. Training in mandatory construction standards and related laws and regulations

Yangzhou Urban and Rural Construction Bureau Administrative Punishment Decision

Yangzhou Urban and Rural Construction Bureau Administrative Punishment Decision

Note: the translation is provided for context only. The process deals with the native language announcements.

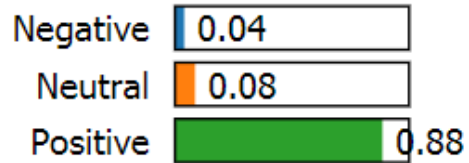


# We are identifying relevant pieces of text and its embedded opinion

## Context Sensitive Extraction, Attribution, and Actionable Notifications via Sentiment Analysis

True class is: Positive

Prediction probabilities



### Text with highlighted words

证券代码: Θ 证券简称: 力尊信通 公告编号  
主办券商: 中航证券 北京力尊信通科技股份有限公司关于获得高新技术企业证书的公告  
本公司及董事会全体成员保证公告内容的真实、准确和完整, 没有虚假记载、误导性陈述或者重大遗漏, 并对其内容的真实性、准确性和完整性承担个别及连带法律责任  
北京力尊信通科技股份有限公司(以下简称“公司”)于 Θ 年 Θ 月 Θ 日收到由北京市科学技术委员会、北京市财政局、北京市国家税务局、北京市地方税务局联合颁发的《高新技术企业证书》(证书编号: αΘ), 发证时间: Θ 年 Θ 月 Θ 日, 有效期三年  
根据相关规定, 公司自通过高新技术企业认定并向主管税务机关办理完减免税手续后三年内, 可享受国家关于高新技术企业的相关优惠政策, 即按 Θ 税率缴纳企业所得税

Positive Connotation – Green  
Negative Connotation – Blue

### Google Translate:

Stock code: 证券 Securities abbreviation: Lizun Xintong Bulletin number

Sponsored Broker: AVIC Securities Beijing Lizun Xintong Technology Co., Ltd. Announcement on Obtaining High-tech Enterprise Certificate

The company and all members of the board of directors guarantee the truthfulness, accuracy and completeness of the contents of the announcement, and there are no false records, misleading statements or major omissions, and bear individual and joint legal responsibility for the truthfulness, accuracy and completeness of the contents.

Beijing Lizun Xintong Technology Co., Ltd. (hereinafter referred to as “the company”) was jointly awarded by the Beijing Municipal Science and Technology Commission, the Beijing Municipal Finance Bureau, the Beijing Municipal State Taxation Bureau and the Beijing Municipal Local Taxation Bureau on the following day. “High-tech Enterprise Certificate” (Certificate No.: αΘ), issuance time: Θ Year Θ Month Day, valid for three years

According to the relevant regulations, the company can enjoy the relevant preferential policies of the state on high-tech enterprises within three years after passing the high-tech enterprise certification and handling the tax reduction and exemption procedures with the competent tax authorities, that is, paying the enterprise income tax at the tax rate.

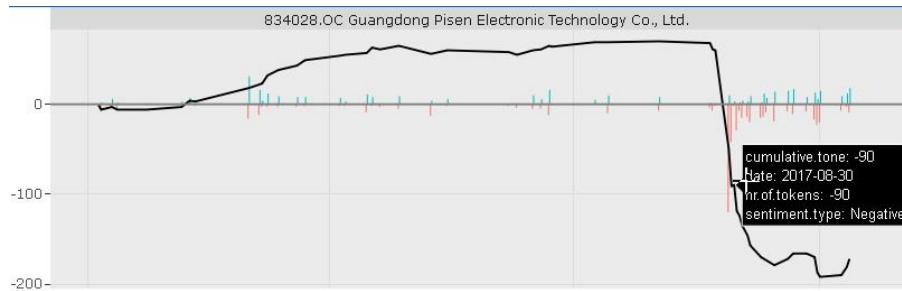
**Note:** the translation is provided for context only. The process deals with the native language announcements.

Permission to reprint or distribute any content from this presentation requires the prior written approval of S&P Global Market Intelligence. Not for distribution to the public. Copyright © 2018 by S&P Global Market Intelligence. All rights reserved.

# UI: We have implemented a simple demo app that aggregates company-level sentiment and attributes the score to key words

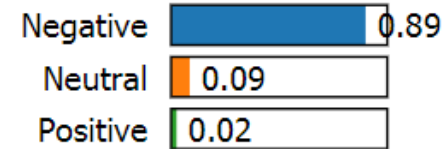
## Screenshots of actual indicative output

Time series of company-level aggregated sentiment



Highlighting of key sentiment driver words in original text

Prediction probabilities



Positive Connotation – Green  
Neutral Connotation – Orange  
Negative Connotation – Blue

### Text with highlighted words

证券代码: 0 证券简称: 中艺股份 公告编号  
南京中艺建筑设计院股份有限公司关于收到扬州市城乡建设局行政处罚决定书的公告  
本公司及董事会全体成员保证公告内容的真实、准确和完整, 没有虚假记载、误导性陈述或者重大遗漏, 并对其内容的真实性、准确性和完整性承担个别及连带法律责任  
南京中艺建筑设计院股份有限公司(以下简称“公司”)近日收到扬州市城乡建设局开具的《行政处罚决定书》(扬建罚  
扬州市城乡建设局城市建设监察支队调查发现公司未 按工程建设强制性标准(条文)对承揽的扬子江世贸商业综合体工程进行设计, 该行为违反了《建设工程质量管理条例》第十九条第一款  
扬州市城乡建设局依据《建设工程质量管理条例》第六十三条第一款第(四)项以及第七十三条对当事人进行处罚  
鉴于该工程尚未开工, 公司积极配合行政机关查处违法行为并整改完成, 主动消除违法行为危害后果, 根据《中华人民共和国行政处罚法》第二十七条第一款的规定, 给予公司减轻行政处罚  
公告编号: 0 共和国行政处罚法》第二十七条第一款的规定, 给予公司减轻行政处罚  
扬州市城乡建设局作出对中艺公司处0元罚款、对直接负责的主管人员王刚处0元罚款、对其他直接责任人员祁岭处0元罚款的行政处罚  
公司将根据《行政处罚决定书》的规定, 在缴款期限内 足额缴纳人民币伍万元罚款, 该笔罚款占营业收入比重较小, 情节较轻  
此外, 该工程尚未开工, 我公司积极配合行政机关查处违法行为并整改完成, 主动消除了危害后果, 本次事件不会对公司经营产生重大影响  
公司管理层高度重视本次事件: 由总经理召开紧急会议, 进行全公司通告, 对相关责任人进行严肃地批评和处罚, 督促全公司认真吸取本次行政处罚的教训, 积极整改; 责令公司员工对公司承接的所有项目进行全面排查, 避免此类问题再次发生; 总结本次事项的教训, 制定预防措施, 特别是进一步加强对此类外地工程设计项目成果质量的严格把控, 定期对公司员工进行工程建设强制性标准及相关法律法规的培训  
扬州市城乡建设局行政处罚决定书》(扬建罚

Word cloud: key drivers of positive/negative sentiment





# Further details on overcoming model transparency

- LIME: An algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model.
- Allows us to analyze the effect of removing a token within a corpus to see how it will impact the overall sentiment classification.
- We can rank these perturbations by article and across the entire corpus via Gini Importance to identify the most significant tokens (words) that have an impact on a certain polarity (positive, negative)
- The paper below explains the approach in detail:
- “Why Should I Trust You?” Explaining the Predictions of Any Classifier
  - <https://arxiv.org/abs/1602.04938>
  - <https://arxiv.org/pdf/1602.04938.pdf>
  - University of Washington:
    - Marco Tulio Ribeiro
    - Sameer Singh
    - Carlos Guestrin

# Conclusion

- **Hope we have demonstrated how we have converted unstructured data to reliable information using a Deep Learning Pipeline.**
- **We are actively using cutting-edge technologies and techniques to improve our clients' day-to-day workflow by building on top of S&P Global Market Intelligence's models and data to solve complex business problems for our clients.**
- **Thank You!**

**Moody Hadi**

Group Manager – Innovation & Product Research

Risk Services

S&P Global Market Intelligence

Copyright © 2019 by S&P Global Market Intelligence, a division of S&P Global Inc. All rights reserved.

These materials have been prepared solely for information purposes based upon information generally available to the public and from sources believed to be reliable. No content (including index data, ratings, credit-related analyses and data, research, model, software or other application or output therefrom) or any part thereof (Content) may be modified, reverse engineered, reproduced or distributed in any form by any means, or stored in a database or retrieval system, without the prior written permission of S&P Global Market Intelligence or its affiliates (collectively, S&P Global). The Content shall not be used for any unlawful or unauthorized purposes. S&P Global and any third-party providers, (collectively S&P Global Parties) do not guarantee the accuracy, completeness, timeliness or availability of the Content. S&P Global Parties are not responsible for any errors or omissions, regardless of the cause, for the results obtained from the use of the Content. THE CONTENT IS PROVIDED ON “AS IS” BASIS. S&P GLOBAL PARTIES DISCLAIM ANY AND ALL EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, ANY WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE OR USE, FREEDOM FROM BUGS, SOFTWARE ERRORS OR DEFECTS, THAT THE CONTENT’S FUNCTIONING WILL BE UNINTERRUPTED OR THAT THE CONTENT WILL OPERATE WITH ANY SOFTWARE OR HARDWARE CONFIGURATION. In no event shall S&P Global Parties be liable to any party for any direct, indirect, incidental, exemplary, compensatory, punitive, special or consequential damages, costs, expenses, legal fees, or losses (including, without limitation, lost income or lost profits and opportunity costs or losses caused by negligence) in connection with any use of the Content even if advised of the possibility of such damages.

S&P Global Market Intelligence's opinions, quotes and credit-related and other analyses are statements of opinion as of the date they are expressed and not statements of fact or recommendations to purchase, hold, or sell any securities or to make any investment decisions, and do not address the suitability of any security. S&P Global Market Intelligence assumes no obligation to update the Content following publication in any form or format. The Content should not be relied on and is not a substitute for the skill, judgment and experience of the user, its management, employees, advisors and/or clients when making investment and other business decisions. S&P Global Market Intelligence does not act as a fiduciary or an investment advisor except where registered as such. S&P Global keeps certain activities of its divisions separate from each other in order to preserve the independence and objectivity of their respective activities. As a result, certain divisions of S&P Global may have information that is not available to other S&P Global divisions. S&P Global has established policies and procedures to maintain the confidentiality of certain non-public information received in connection with each analytical process.

S&P Global Ratings does not contribute to or participate in the creation of credit scores generated by S&P Global Market Intelligence. Lowercase nomenclature is used to differentiate S&P Global Market Intelligence PD credit model scores from the credit ratings issued by S&P Global Ratings.

S&P Global may receive compensation for its ratings and certain analyses, normally from issuers or underwriters of securities or from obligors. S&P Global reserves the right to disseminate its opinions and analyses. S&P Global's public ratings and analyses are made available on its Web sites, [www.standardandpoors.com](http://www.standardandpoors.com) (free of charge) and [www.ratingsdirect.com](http://www.ratingsdirect.com) (subscription), and may be distributed through other means, including via S&P Global publications and third-party redistributors. Additional information about our ratings fees is available at [www.standardandpoors.com/usratingsfees](http://www.standardandpoors.com/usratingsfees).