Marketing Material

Step by Step Guide to Natural Language Processing: Extract ESG Sentiment from Company Reports

RAM Systematic Equity Team

06.10.2022

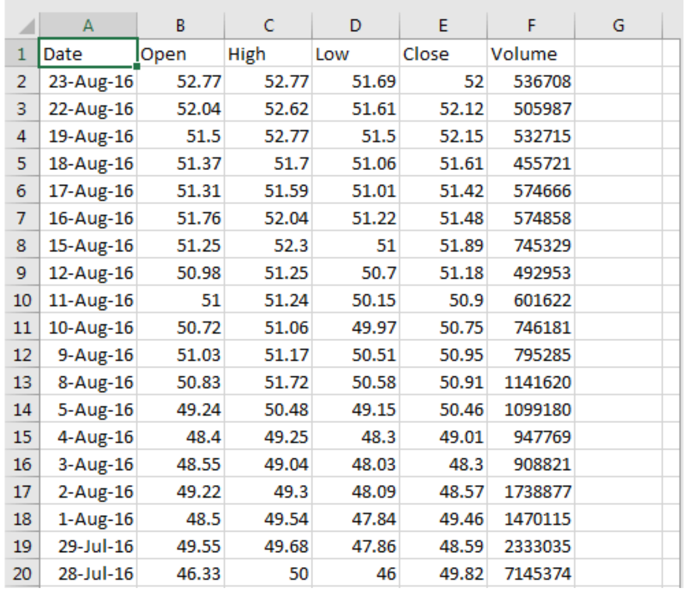
Table of Content

- Natural Language Processing
  - Rule based
  - Statistics based
  - Machine learning based
- Applications to Sustainable Finance
  - Sentiment extraction
  - Implementation demo

Demo on Google Colaboratory

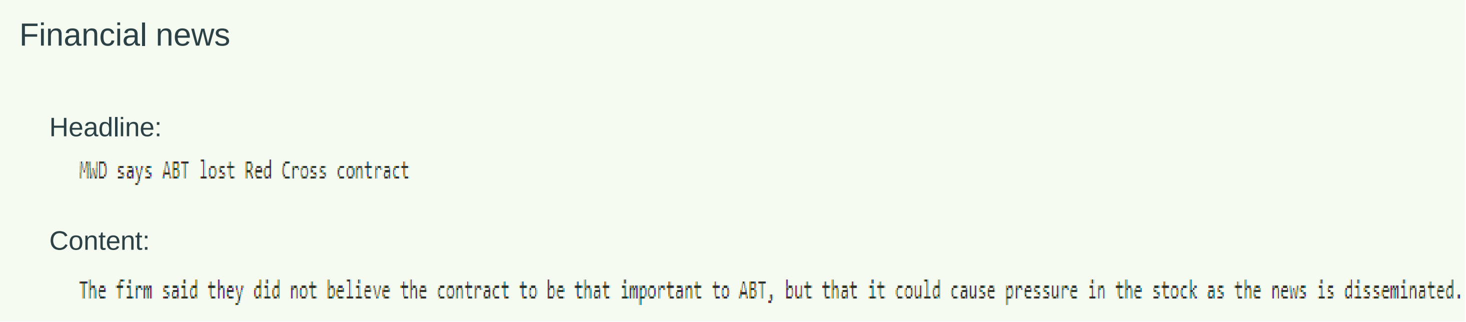
Structured and Unstructured Data

Structured Data



	A	B	C	D	E	F	G
1	Date	Open	High	Low	Close	Volume	
2	23-Aug-16	52.77	52.77	51.69	52	536708	
3	22-Aug-16	52.04	52.62	51.61	52.12	505987	
4	19-Aug-16	51.5	52.77	51.5	52.15	532715	
5	18-Aug-16	51.37	51.7	51.06	51.61	455721	
6	17-Aug-16	51.31	51.59	51.01	51.42	574666	
7	16-Aug-16	51.76	52.04	51.22	51.48	574858	
8	15-Aug-16	51.25	52.3	51	51.89	745329	
9	12-Aug-16	50.98	51.25	50.7	51.18	492953	
10	11-Aug-16	51	51.24	50.15	50.9	601622	
11	10-Aug-16	50.72	51.06	49.97	50.75	746181	
12	9-Aug-16	51.03	51.17	50.51	50.95	795285	
13	8-Aug-16	50.83	51.72	50.58	50.91	1141620	
14	5-Aug-16	49.24	50.48	49.15	50.46	1099180	
15	4-Aug-16	48.4	49.25	48.3	49.01	947769	
16	3-Aug-16	48.55	49.04	48.03	48.3	908821	
17	2-Aug-16	49.22	49.3	48.09	48.57	1738877	
18	1-Aug-16	48.5	49.54	47.84	49.46	1470115	
19	29-Jul-16	49.55	49.68	47.86	48.59	2333035	
20	28-Jul-16	46.33	50	46	49.82	7145374	

Unstructured Data



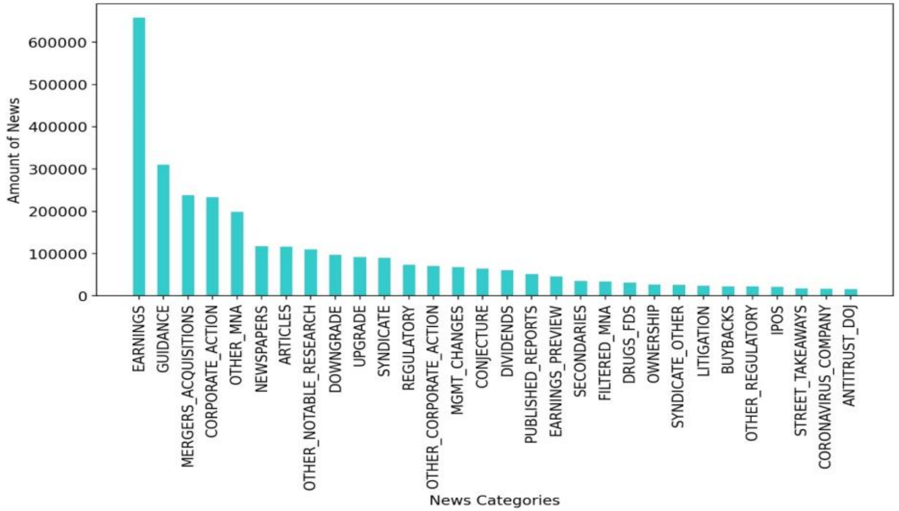
Financial news

Headline:

```
MWD says ABT lost Red Cross contract
```

Content:

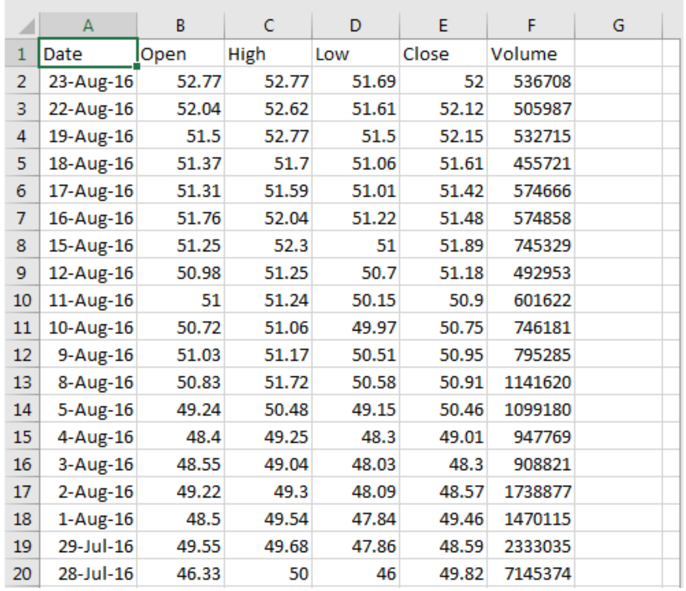
```
The firm said they did not believe the contract to be that important to ABT, but that it could cause pressure in the stock as the news is disseminated.
```



Source: RAM Active Investments, StreetAccount, Factset.

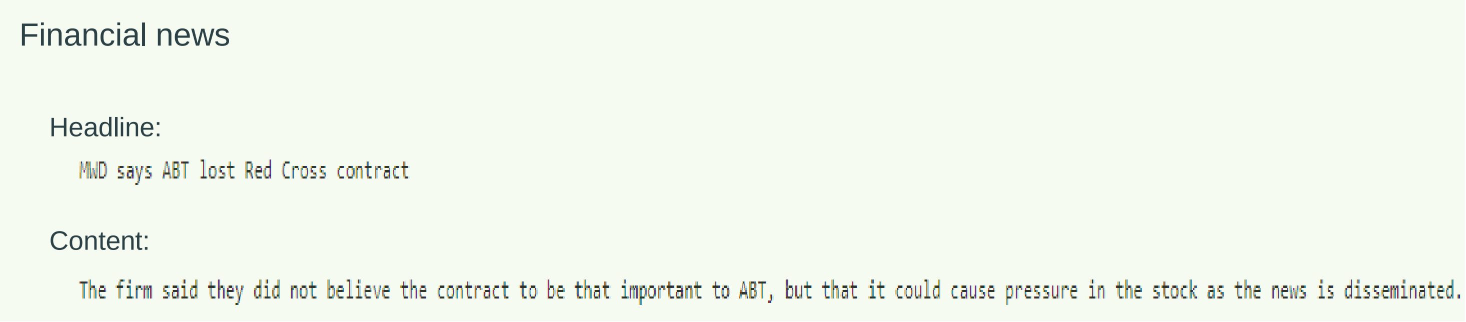
Structured and Unstructured Data

Structured Data



	A	B	C	D	E	F	G
1	Date	Open	High	Low	Close	Volume	
2	23-Aug-16	52.77	52.77	51.69	52	536708	
3	22-Aug-16	52.04	52.62	51.61	52.12	505987	
4	19-Aug-16	51.5	52.77	51.5	52.15	532715	
5	18-Aug-16	51.37	51.7	51.06	51.61	455721	
6	17-Aug-16	51.31	51.59	51.01	51.42	574666	
7	16-Aug-16	51.76	52.04	51.22	51.48	574858	
8	15-Aug-16	51.25	52.3	51	51.89	745329	
9	12-Aug-16	50.98	51.25	50.7	51.18	492953	
10	11-Aug-16	51	51.24	50.15	50.9	601622	
11	10-Aug-16	50.72	51.06	49.97	50.75	746181	
12	9-Aug-16	51.03	51.17	50.51	50.95	795285	
13	8-Aug-16	50.83	51.72	50.58	50.91	1141620	
14	5-Aug-16	49.24	50.48	49.15	50.46	1099180	
15	4-Aug-16	48.4	49.25	48.3	49.01	947769	
16	3-Aug-16	48.55	49.04	48.03	48.3	908821	
17	2-Aug-16	49.22	49.3	48.09	48.57	1738877	
18	1-Aug-16	48.5	49.54	47.84	49.46	1470115	
19	29-Jul-16	49.55	49.68	47.86	48.59	2333035	
20	28-Jul-16	46.33	50	46	49.82	7145374	

Unstructured Data



Financial news

Headline:

```
MWD says ABT lost Red Cross contract
```

Content:

```
The firm said they did not believe the contract to be that important to ABT, but that it could cause pressure in the stock as the news is disseminated.
```

Quantitative Analysis

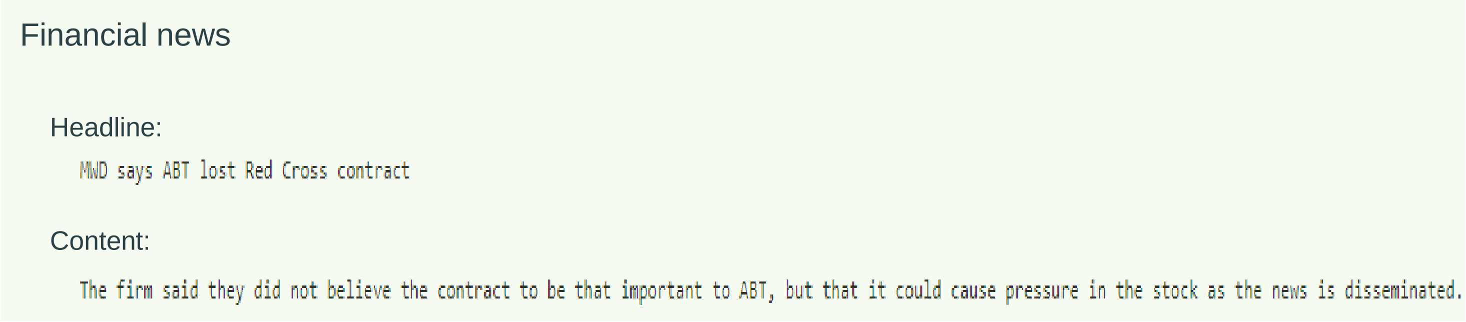
$$\mathbf{x} \in \mathbb{R}^d \quad y \in \mathbb{R} \quad \hat{y} = f(\mathbf{X})$$

$$\hat{y} = \sum_{i=1}^{d} w_i x_i + b$$

$$\hat{y} = \mathrm{w}^\top \sigma(\mathbf{W}\mathbf{x} + \mathbb{B}) + b$$

Structured and Unstructured Data

Unstructured Data



Financial news

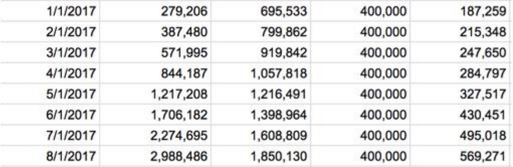
Headline:

```
MND says ABT lost Red Cross contract
```

Content:

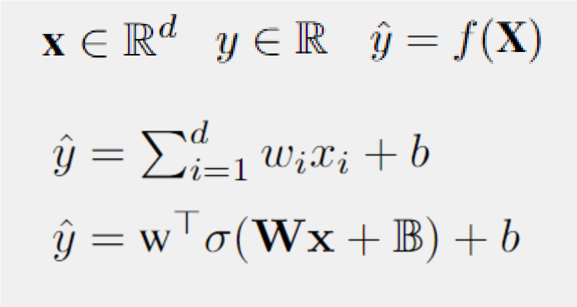
```
The firm said they did not believe the contract to be that important to ABT, but that it could cause pressure in the stock as the news is disseminated.
```

Transformation



1/1/2017	279,206	695,533	400,000	187,259
2/1/2017	387,480	799,862	400,000	215,348
3/1/2017	571,995	919,842	400,000	247,650
4/1/2017	844,187	1,057,818	400,000	284,797
5/1/2017	1,217,208	1,216,491	400,000	327,517
6/1/2017	1,706,182	1,398,964	400,000	430,451
7/1/2017	2,274,695	1,608,809	400,000	495,018
8/1/2017	2,988,486	1,850,130	400,000	569,271

Quantitative Analysis

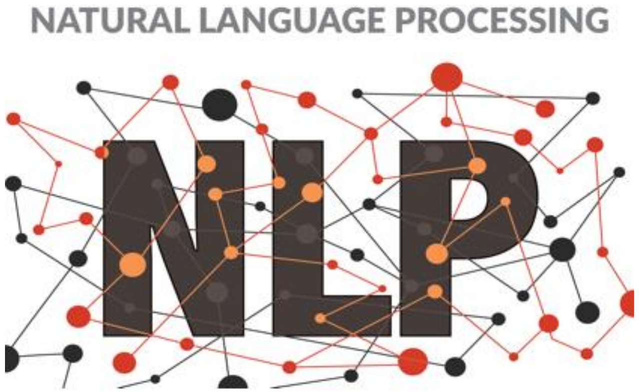


$$\mathbf{x} \in \mathbb{R}^d \quad y \in \mathbb{R} \quad \hat{y} = f(\mathbf{X})$$

$$\hat{y} = \sum_{i=1}^{d} w_i x_i + b$$

$$\hat{y} = \mathrm{w}^\top \sigma(\mathbf{W}\mathbf{x} + \mathbb{B}) + b$$

Natural Language Processing



Unstructured Data

Financial news

Headline:

`MND says ABT lost Red Cross contract`

Content:

`The firm said they did not believe the contract to be that important to ABT, but that it could cause pressure in the stock as the news is disseminated.`

Transformation

Quantitative Analysis

$$\mathbf{x} \in \mathbb{R}^d \quad y \in \mathbb{R} \quad \hat{y} = f(\mathbf{X})$$

$$\hat{y} = \sum_{i=1}^{d} w_i x_i + b$$

$$\hat{y} = \mathrm{w}^\top \sigma(\mathbf{W}\mathbf{x} + \mathbb{B}) + b$$

Source: Chris Kuo/Dr. Dataman, "Looking into Natural Language Processing
", https://medium.com/dataman-in-ai/natural-language-processing-nlp-for-electronic-health-record-ehr-part-i-4cb1d4c2f24b, 2018

Natural Language Processing

NLP development phases:

- Rule based
- Statistics based
- Machine learning based

“In the early 1900s, a Swiss linguistics professor named Ferdinand de Saussure almost deprived the world of the concept of “Language as a Science.”

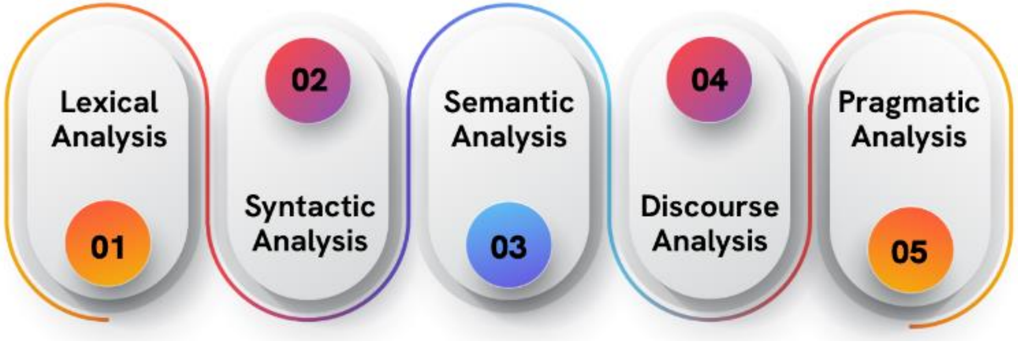
“NLP makes computers capable of ‘understanding’ the contents of documents”

Source: Keith D. Foote, “A Brief History of Natural Language Processing”, https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/, 2019.

Natural Language Processing

Rule-based

- Automatic parsing and information extraction
- Discretionary analysis



Source: OptiSol, “The 5 phases of natural language processing”, https://www.optisolbusiness.com/insight/the-5-phases-of-natural-language-processing, 2022

NLP development phases:

- Rule based
- Statistics based
- Machine learning based

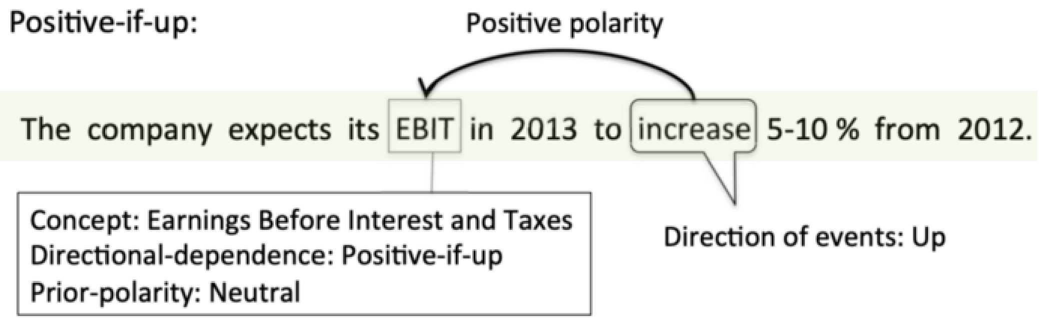
Natural Language Processing

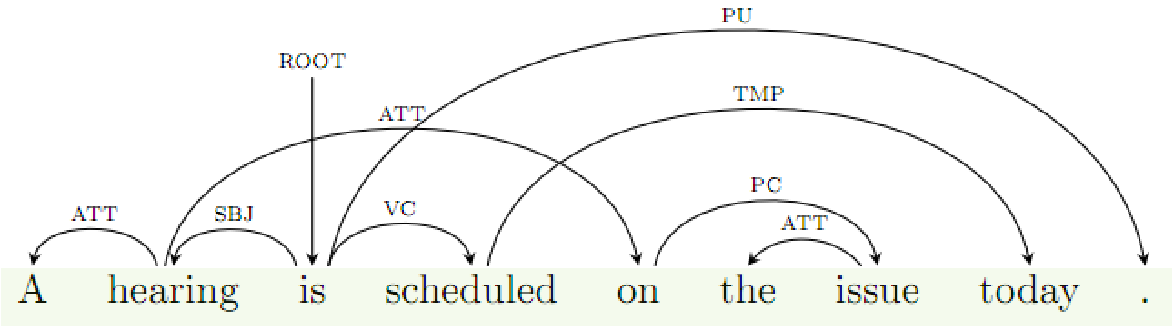
Rule-based

- Automatic parsing and information extraction
- Discretionary analysis

NLP development phases:

- Rule based
- Statistics based
- Machine learning based


Positive-if-up:
Positive polarity
The company expects its EBIT in 2013 to increase 5-10 % from 2012.
Concept: Earnings Before Interest and Taxes
Directional-dependence: Positive-if-up
Prior-polarity: Neutral
Direction of events: Up



Source: Neo Yi Peng," How NLP has evolved for Financial Sentiment Analysis", "https://towardsdatascience.com/how-nlp-has-evolved-for-financial-sentiment-analysis-fb2990d9b3ed, 2020.
Matt Payne, "7 NLP Techniques for Extracting Information from Unstructured Text using Algorithms", https://www.width.ai/post/extracting-information-from-unstructured-text-using-algorithms, 2021.

Natural Language Processing

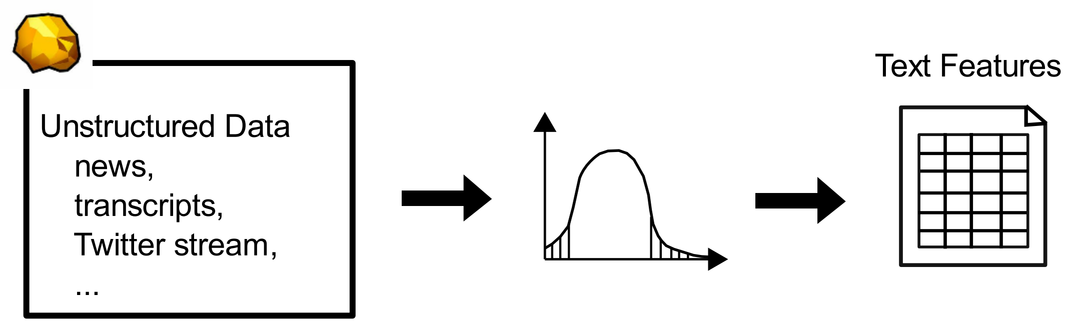
Statistics-based

- Language models
- Topic models
- Linguistic feature extraction

NLP development phases:

- Rule based
- Statistics based
- Machine learning based

Data



Natural Language Processing

Statistics-based

- Language models
- Topic models
- Linguistic feature extraction

NLP development phases:

- Rule based
- Statistics based
- Machine learning based

Language models

(probability) Generative models

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

$$P(\text{next word } = ? | \text{The company expects its})$$

e.g., EBIT, annual return, etc.

Applications: machine translation,
speech recognition,
spelling correction, etc.

Tower of Babel

Source: Wikipedia

Natural Language Processing

Statistics-based

- Language models
- Topic models
- Linguistic feature extraction

Topics models

Generative latent variable models

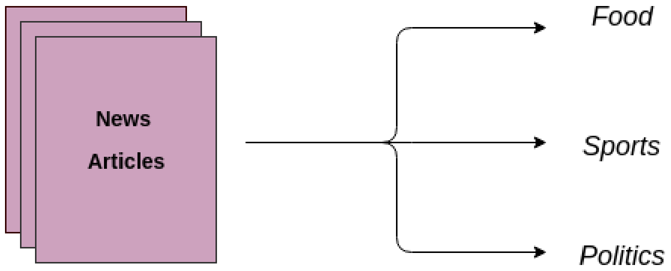
Latent Dirichlet allocation

$$P(\boldsymbol{Z}, \boldsymbol{W}; \alpha, \beta) = \int_{\theta} \int_{\varphi} P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta)\, d\boldsymbol{\varphi}\, d\boldsymbol{\theta}$$

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) P(W_{j,t} \mid \varphi_{Z_{j,t}}),$$

NLP development phases:

- Rule based
- Statistics based
- Machine learning based



Source: Khuyen Tran, ”pyLDAvis: Topic Modelling Exploration Tool That Every NLP Data Scientist Should Know”, https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know, 2022.

Natural Language Processing

Statistics-based

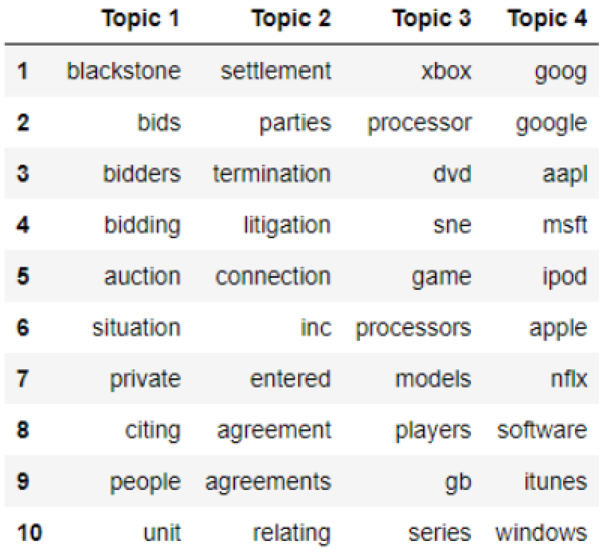
- Language models
- Topic models
- Linguistic feature extraction

NLP development phases:

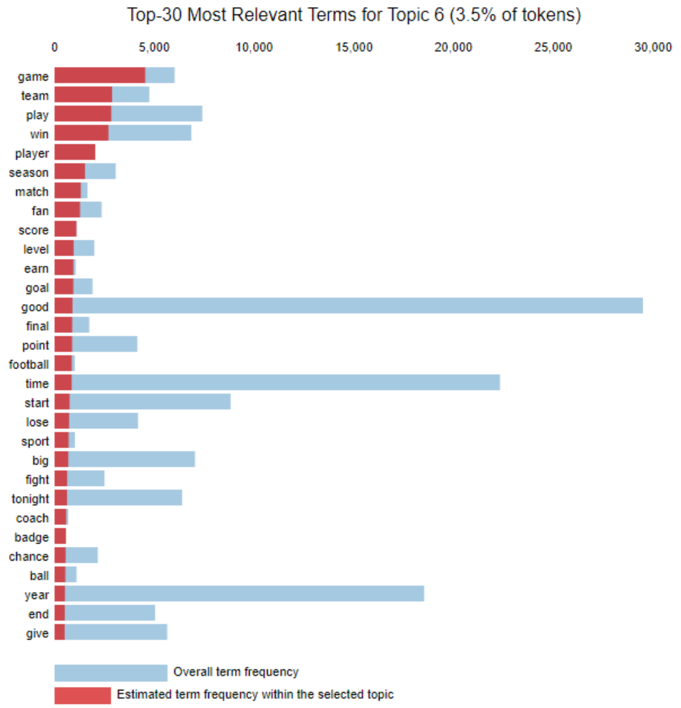
- Rule based
- Statistics based
- Machine learning based

Topics models

Generative latent variable models



	Topic 1	Topic 2	Topic 3	Topic 4
1	blackstone	settlement	xbox	goog
2	bids	parties	processor	google
3	bidders	termination	dvd	aapl
4	bidding	litigation	sne	msft
5	auction	connection	game	ipod
6	situation	inc	processors	apple
7	private	entered	models	nflx
8	citing	agreement	players	software
9	people	agreements	gb	itunes
10	unit	relating	series	windows



Source: Khuyen Tran, "pyLDAvis: Topic Modelling Exploration Tool That Every NLP Data Scientist Should Know", https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know, 2022.

Natural Language Processing

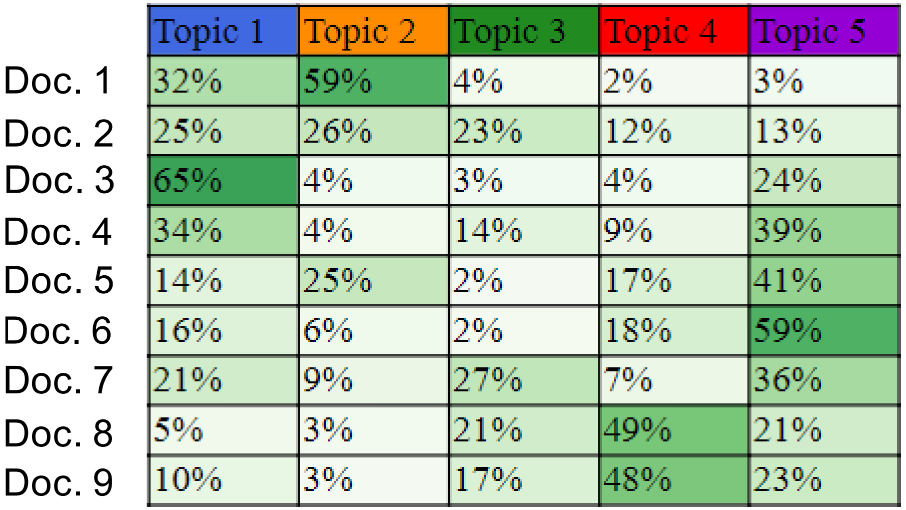
Statistics-based

- Language models
- Topic models
- Linguistic feature extraction

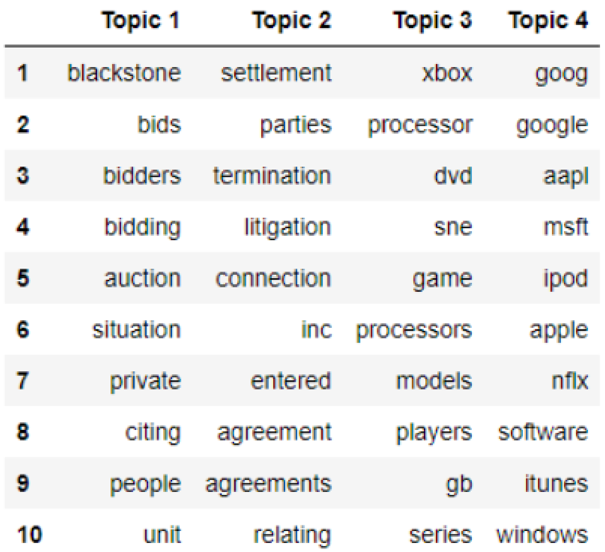
NLP development phases:

- Rule based
- Statistics based
- Machine learning based

Topics models: Generative latent variable models



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Doc. 1	32%	59%	4%	2%	3%
Doc. 2	25%	26%	23%	12%	13%
Doc. 3	65%	4%	3%	4%	24%
Doc. 4	34%	4%	14%	9%	39%
Doc. 5	14%	25%	2%	17%	41%
Doc. 6	16%	6%	2%	18%	59%
Doc. 7	21%	9%	27%	7%	36%
Doc. 8	5%	3%	21%	49%	21%
Doc. 9	10%	3%	17%	48%	23%



	Topic 1	Topic 2	Topic 3	Topic 4
1	blackstone	settlement	xbox	goog
2	bids	parties	processor	google
3	bidders	termination	dvd	aapl
4	bidding	litigation	sne	msft
5	auction	connection	game	ipod
6	situation	inc	processors	apple
7	private	entered	models	nflx
8	citing	agreement	players	software
9	people	agreements	gb	itunes
10	unit	relating	series	windows

Source: Boemer, Dominik. "Topic modeling of investment style news." (2020).

Natural Language Processing

Statistics-based

- Language models
- Topic models
- Linguistic feature extraction

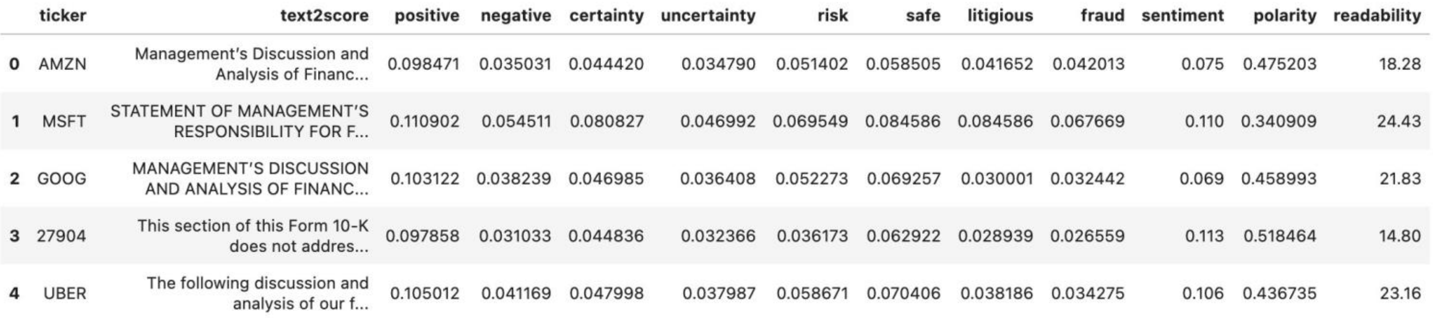
NLP development phases:

- Rule based
- Statistics based
- Machine learning based

Linguistic feature extraction

Domain-specific lexicons

e.g., positive, negative, litigious, polarity, risk, readability, fraud, safe, certainty, uncertainty, and sentiment.



	ticker	text2score	positive	negative	certainty	uncertainty	risk	safe	litigious	fraud	sentiment	polarity	readability
0	AMZN	Management's Discussion and Analysis of Financ...	0.098471	0.035031	0.044420	0.034790	0.051402	0.058505	0.041652	0.042013	0.075	0.475203	18.28
1	MSFT	STATEMENT OF MANAGEMENT'S RESPONSIBILITY FOR F...	0.110902	0.054511	0.080827	0.046992	0.069549	0.084586	0.084586	0.067669	0.110	0.340909	24.43
2	GOOG	MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANC...	0.103122	0.038239	0.046985	0.036408	0.052273	0.069257	0.030001	0.032442	0.069	0.458993	21.83
3	27904	This section of this Form 10-K does not addres...	0.097858	0.031033	0.044836	0.032366	0.036173	0.062922	0.028939	0.026559	0.113	0.518464	14.80
4	UBER	The following discussion and analysis of our f...	0.105012	0.041169	0.047998	0.037987	0.058671	0.070406	0.038186	0.034275	0.106	0.436735	23.16

Source: Sanjiv Das, Bodhisatta Saha, Daniel Zhu, and Derrick Zhang, ”Create a dashboard with SEC text for financial NLP in Amazon SageMaker JumpStart”, https://aws.amazon.com/blogs/machine-learning/create-a-dashboard-with-sec-text-for-financial-nlp-in-amazon-sagemaker-jumpstart/, 2021.

Natural Language Processing

Rule-based

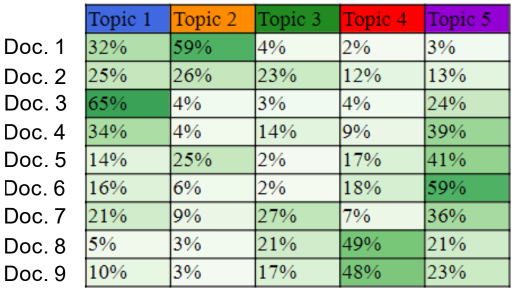
- Automatic parsing and information extraction
- Discretionary analysis

Statistics-based

- Language models
- Topic models
- Linguistic feature extraction

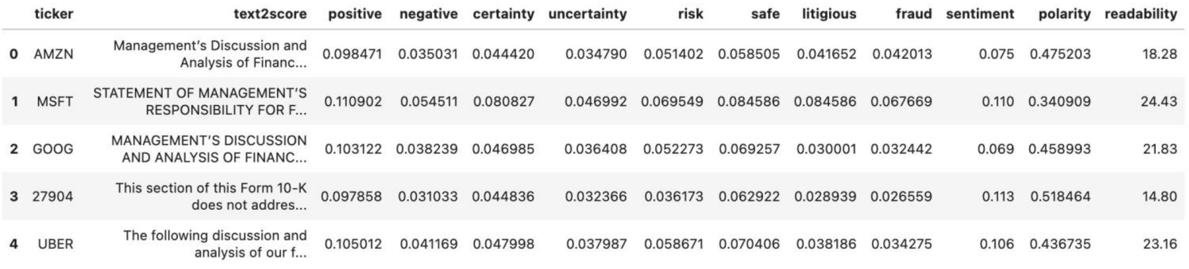
Drawbacks

- Lack of semantic distinguishability
- Lack of end-to-end development
- …



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Doc. 1	32%	59%	4%	2%	3%
Doc. 2	25%	26%	23%	12%	13%
Doc. 3	65%	4%	3%	4%	24%
Doc. 4	34%	4%	14%	9%	39%
Doc. 5	14%	25%	2%	17%	41%
Doc. 6	16%	6%	2%	18%	59%
Doc. 7	21%	9%	27%	7%	36%
Doc. 8	5%	3%	21%	49%	21%
Doc. 9	10%	3%	17%	48%	23%

Source: Boemer, Dominik. "Topic modeling of investment style news." (2020).



	ticker	text2score	positive	negative	certainty	uncertainty	risk	safe	litigious	fraud	sentiment	polarity	readability
0	AMZN	Management's Discussion and Analysis of Financ...	0.098471	0.035031	0.044420	0.034790	0.051402	0.058505	0.041652	0.042013	0.075	0.475203	18.28
1	MSFT	STATEMENT OF MANAGEMENT'S RESPONSIBILITY FOR F...	0.110902	0.054511	0.080827	0.046992	0.069549	0.084586	0.084586	0.067669	0.110	0.340909	24.43
2	GOOG	MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANC...	0.103122	0.038239	0.046985	0.036408	0.052273	0.069257	0.030001	0.032442	0.069	0.458993	21.83
3	27904	This section of this Form 10-K does not addres...	0.097858	0.031033	0.044836	0.032366	0.036173	0.062922	0.028939	0.026559	0.113	0.518464	14.80
4	UBER	The following discussion and analysis of our f...	0.105012	0.041169	0.047998	0.037987	0.058671	0.070406	0.038186	0.034275	0.106	0.436735	23.16

Source: Sanjiv Das, Bodhisatta Saha, Daniel Zhu, and Derrick Zhang, "Create a dashboard with SEC text for financial NLP in Amazon SageMaker JumpStart", https://aws.amazon.com/blogs/machine-learning/create-a-dashboard-with-sec-text-for-financial-nlp-in-amazon-sagemaker-jumpstart/, 2021.

"NOK hopes its N-Gage mobile phone will boost sales and attract younger generation…"

"HBC and BCS upgraded to outperform from peer perform at Bear Stearns… "

"BEAS upgraded to overweight from equal weight at ThinkEquity…"

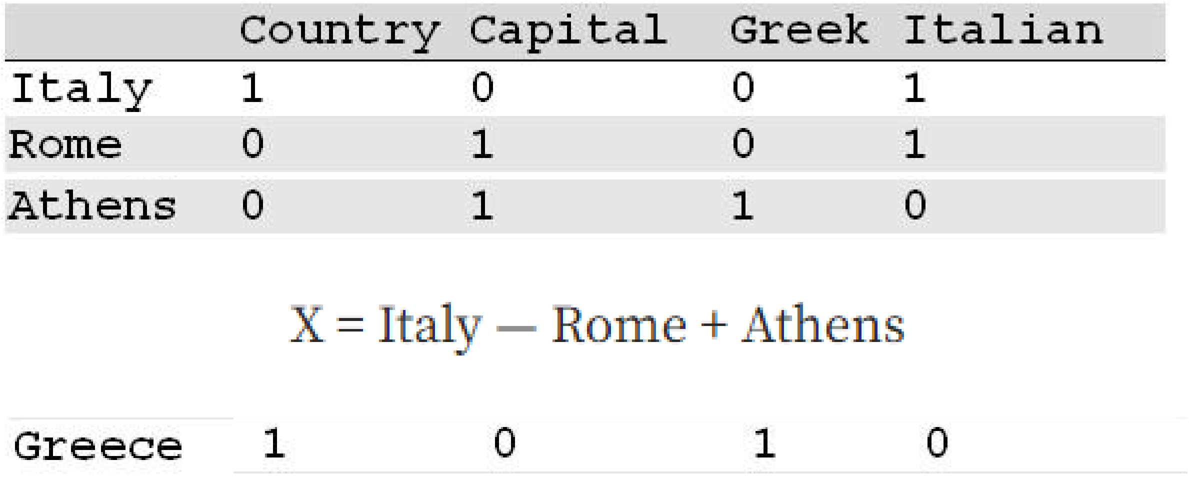
Source: StreetAccount

Natural Language Processing

Machine learning based

Text embedding: define meaning with coordinates/vectors

An introductory example:



	Country	Capital	Greek	Italian
Italy	1	0	0	1
Rome	0	1	0	1
Athens	0	1	1	0

$$X = Italy - Rome + Athens$$

Greece	1	0	1	0

NLP development phases:

- Rule based
- Statistics based
- Machine learning based

Source: Yuli Vasiliev, “Using Word Vectors to Mathematically Find Words with Similar Meanings”, https://towardsdatascience.com/using-word-vectors-to-mathematically-find-words-with-similar-meanings-fc9eadac3dbe, 2020.

Natural Language Processing

Machine learning based

Text embedding

- High-dimensional vectors in the semantic space
- hundreds of dimensions
- Downstream tasks

Data:

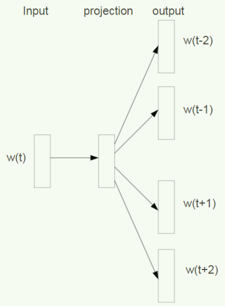
- Large generic corpus
- Domain-specific data

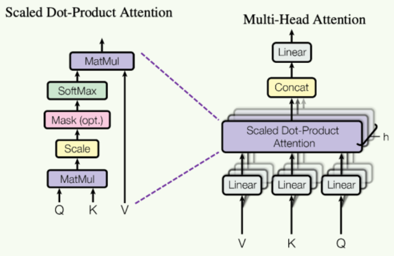
Wikipedia

Unstructured Data
news
transcripts
Twitter stream
...

Model architecture:

- Similarity
- Relevance
- …



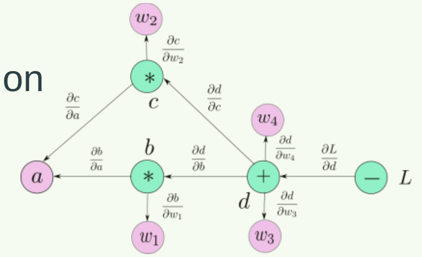


NLP development phases:

- Rule based
- Statistics based
- Machine learning based

Model training:

- Automatic differentiation
- Computing power



w_2
$\frac{\partial c}{\partial w_2}$
$*$
c
$\frac{\partial c}{\partial a}$
$\frac{\partial d}{\partial c}$
w_4
$\frac{\partial d}{\partial w_4}$
b
$\frac{\partial b}{\partial a}$
$\frac{\partial d}{\partial b}$
$\frac{\partial L}{\partial d}$
a
$*$
$+$
$-$
L
$\frac{\partial b}{\partial w_1}$
d
$\frac{\partial d}{\partial w_3}$
w_1
w_3

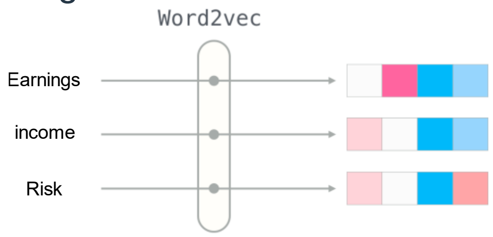
Source:
(1) Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013).
(2) Yasuto Tamura, ”Multi-head attention mechanism: “queries”, “keys”, and “values,” over and over again”, https://data-science-blog.com/blog/2021/04/07/multi-head-attention-mechanism/, 2021.
(3) Ayoosh Kathuria, ”PyTorch 101, Part 1: Understanding Graphs, Automatic Differentiation and Autograd”, https://blog.paperspace.com/pytorch-101-understanding-graphs-and-automatic-differentiation/, 2020.

Natural Language Processing

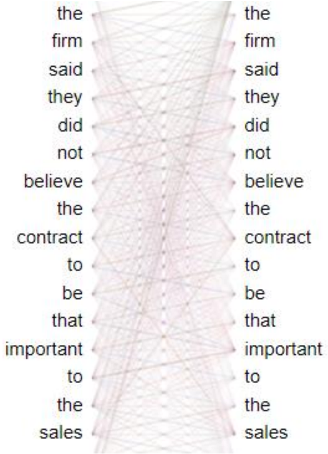
Machine learning based

Text embedding

- Word embedding



- Contextualized embedding
  - through large language models



Data:

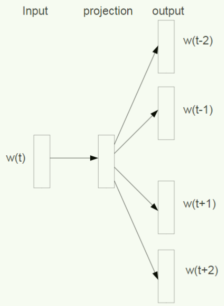
- Large generic corpus
- Domain-specific data

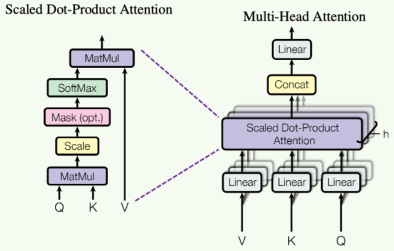
Wikipedia

Unstructured Data
news
transcripts
Twitter stream
...

Model architecture:

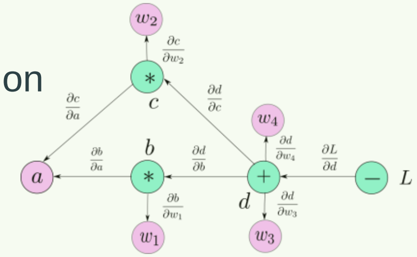
- Similarity
- Relevance
- …





Model training:

- Automatic differentiation
- Computing power


w_2
$\frac{\partial c}{\partial w_2}$
$*$
c
$\frac{\partial c}{\partial a}$
$\frac{\partial d}{\partial c}$
w_4
a
$\frac{\partial b}{\partial a}$
b
$*$
$\frac{\partial d}{\partial b}$
$+$
$\frac{\partial d}{\partial w_4}$
$\frac{\partial L}{\partial d}$
$-$
L
$\frac{\partial b}{\partial w_1}$
d
$\frac{\partial d}{\partial w_3}$
w_1
w_3

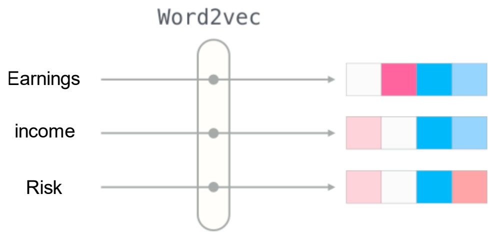
Source:
(1) Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013).
(2) Yasuto Tamura,"Multi-head attention mechanism: “queries”, “keys”, and “values,” over and over again”, https://data-science-blog.com/blog/2021/04/07/multi-head-attention-mechanism/, 2021.
(3) Ayoosh Kathuria, ”PyTorch 101, Part 1: Understanding Graphs, Automatic Differentiation and Autograd”, https://blog.paperspace.com/pytorch-101-understanding-graphs-and-automatic-differentiation/, 2020.
(4) Jay Alammar, “The Illustrated Word2vec”, https://jalammar.github.io/illustrated-word2vec/, 2019.

Natural Language Processing

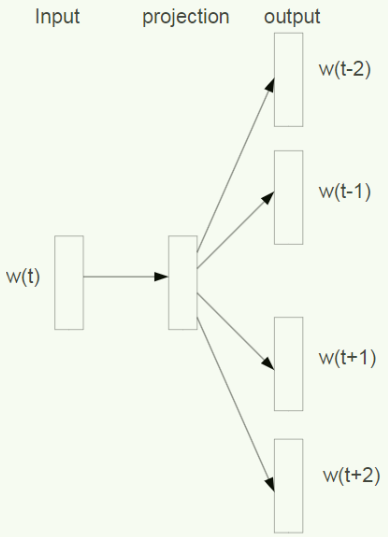
Machine learning based

Text embedding

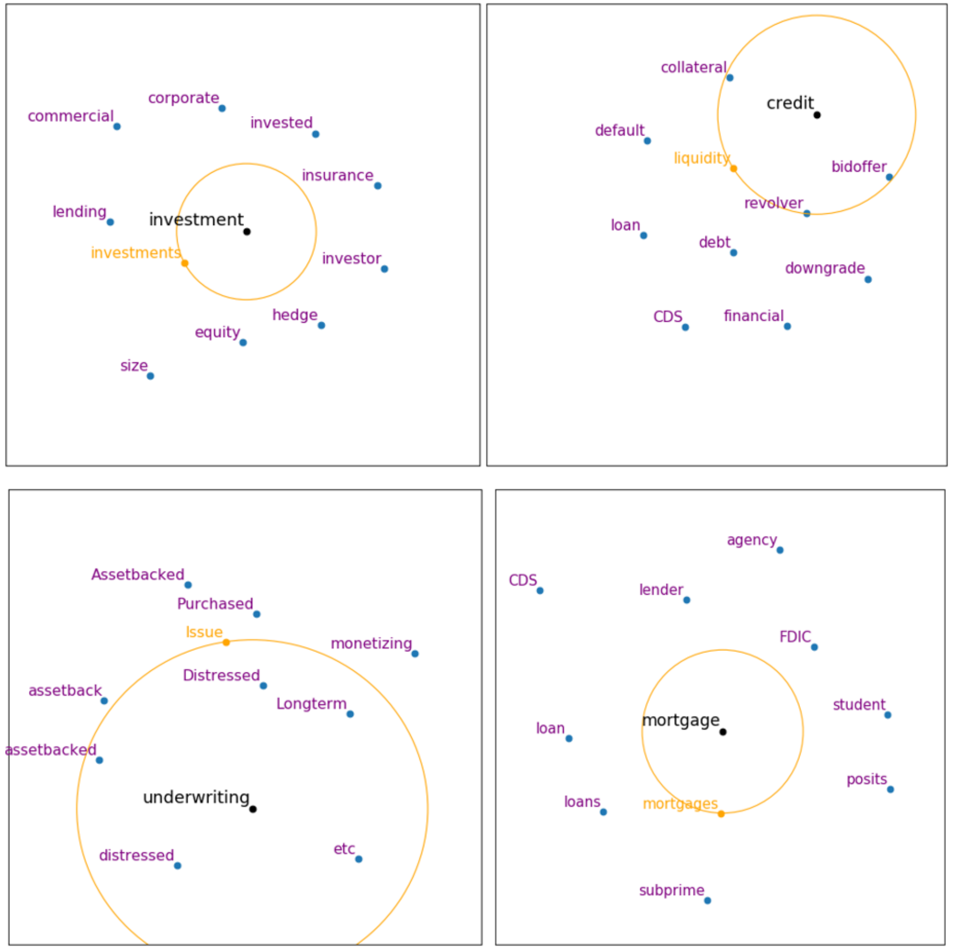
- Word embedding



“The firm said they did not believe the contract to be that important to the sales."



Minimize $\log \sigma({v'_{w_O}}^{\top} v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^{\top} v_{w_I}) \right]$



Source:
(1) Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013).
(2) Manu Siddhartha, “BankFin Embeddings : Customized word embeddings Pre-Trained on Financial Text corpus for Financial NLP tasks”, https://github.com/sid321axn/bank_fin_embedding, 2020.

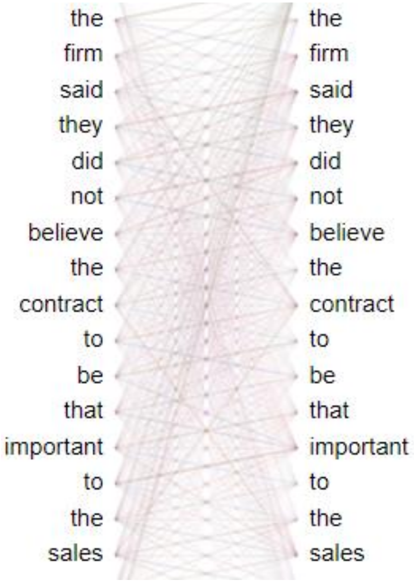
Natural Language Processing

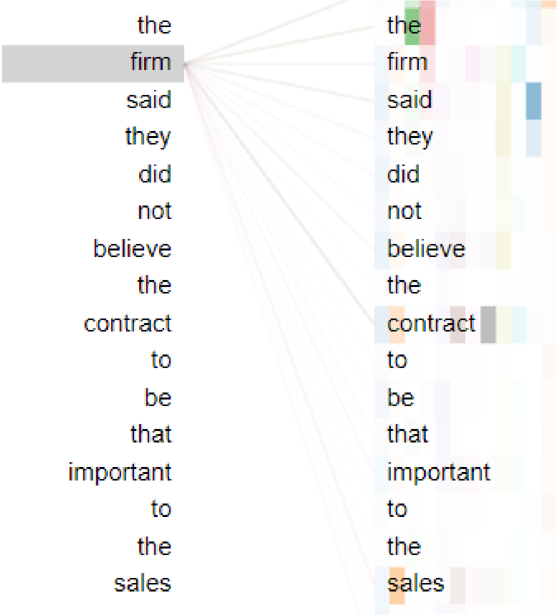
Machine learning based

Text embedding

- Contextualized embedding

“The firm said they did not believe the contract to be that important to the sales."





Source:
(1) Jesse Vig, ”Visualize Attention in NLP Models”, https://github.com/jessevig/bertviz, 2022.
(2) Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

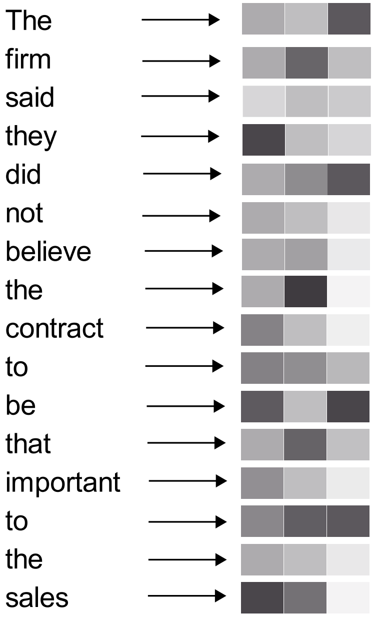
Natural Language Processing

Machine learning based

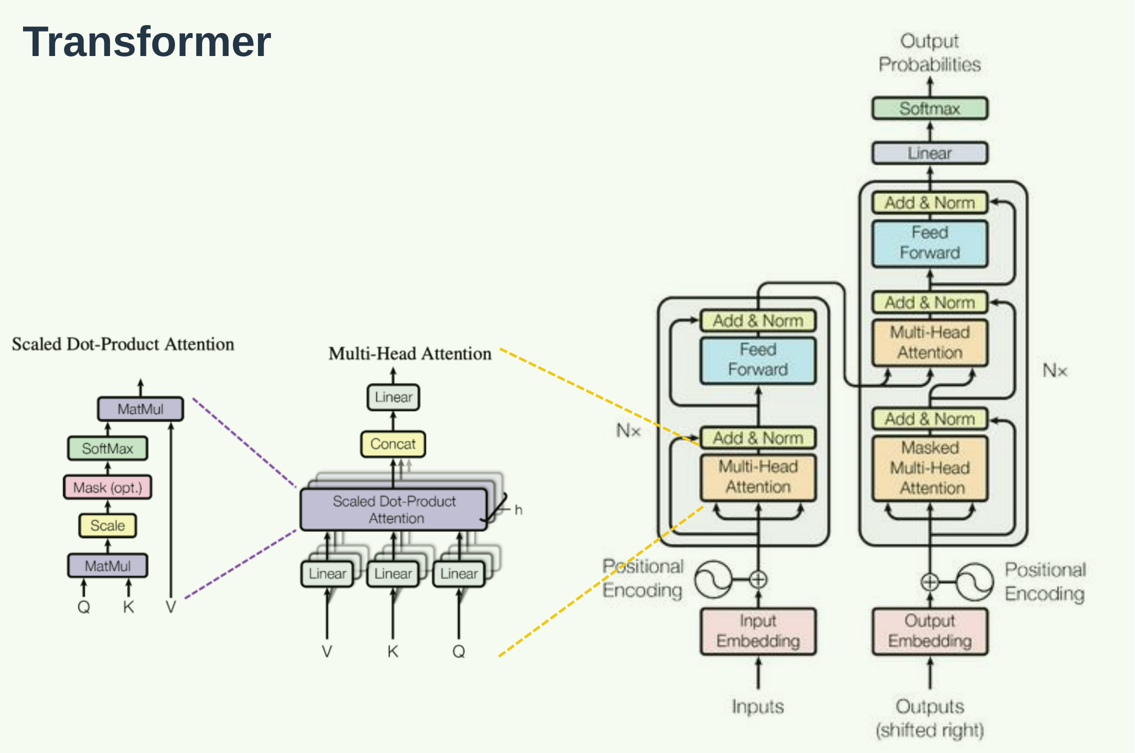
Text embedding

- Contextualized embedding

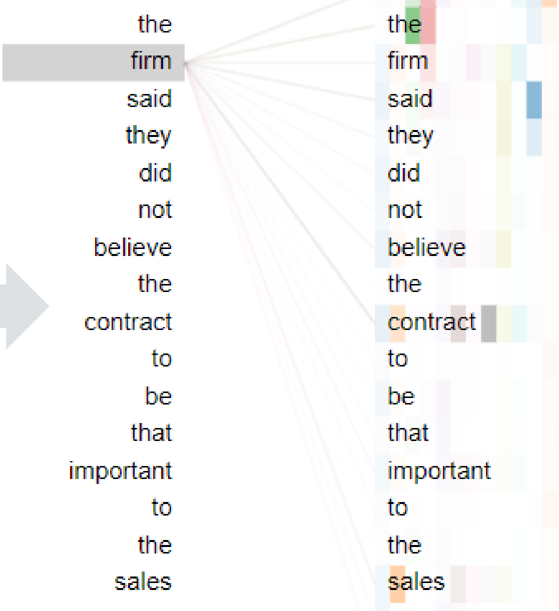
"The firm said they did not believe the contract to be that important to the sales."



The
firm
said
they
did
not
believe
the
contract
to
be
that
important
to
the
sales



Transformer



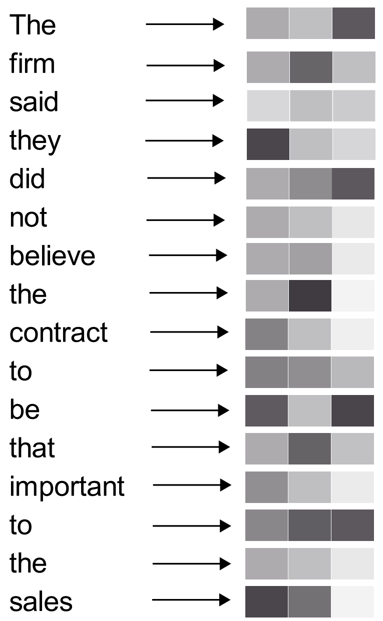
Source:
(1) Jesse Vig, "Visualize Attention in NLP Models", https://github.com/jessevig/bertviz, 2022.
(2) Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
(3) Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

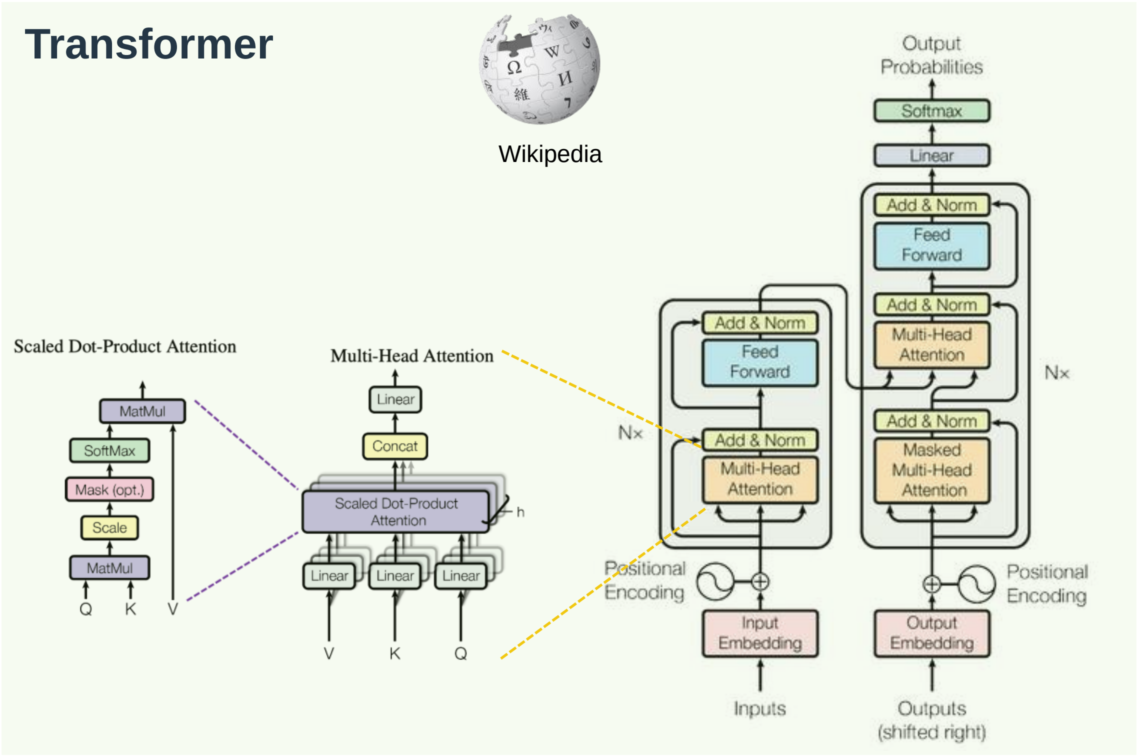
Natural Language Processing

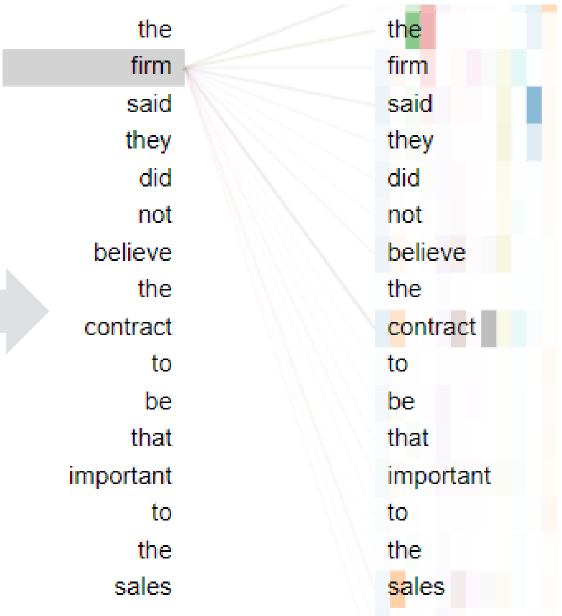
Machine learning based

Large language models

- Training tasks: (randomly) masked words, next sentence prediction, etc.







Source:
(1) Jesse Vig, "Visualize Attention in NLP Models", https://github.com/jessevig/bertviz, 2022.
(2) Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
(3) Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

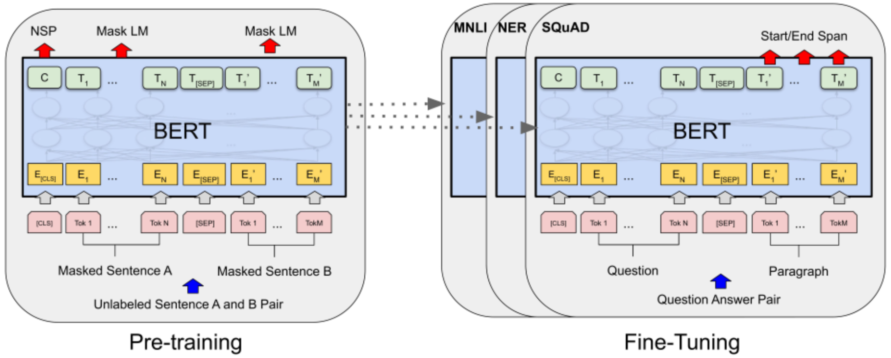
Natural Language Processing

Machine learning based

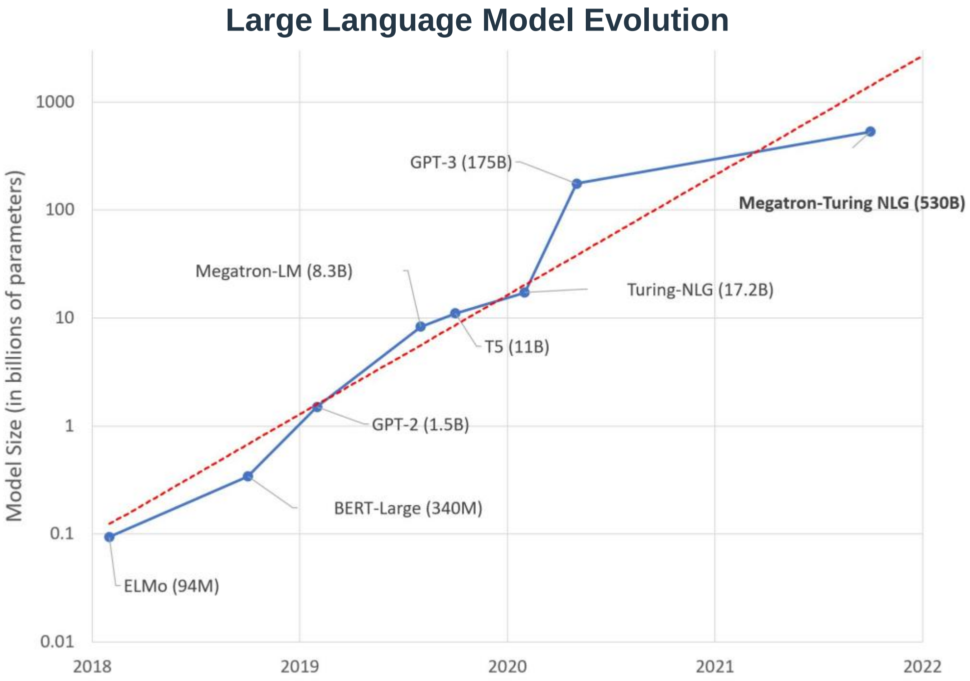
Large language models

- Pre-trained on large and general language corpus
- Fine-tuned on application specific data

Wikipedia



Source: Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).


Large Language Model Evolution

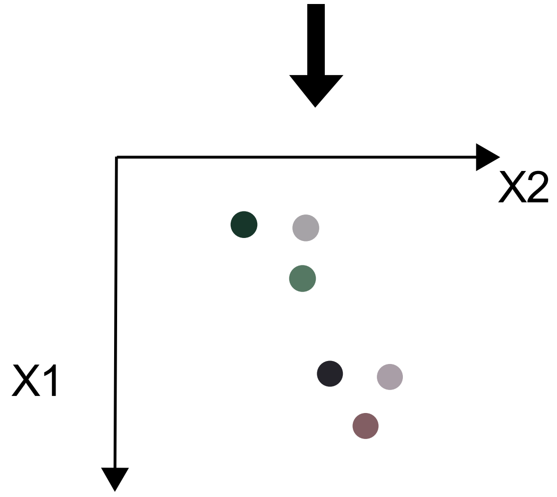
Source: Julien Simon, https://huggingface.co/blog/large-language-models, 2021.

Natural Language Processing

Applications of text embedding

- Text semantic similarity
- Text clustering
- …

- "...warms up to idea of potential acquisitions in new markets..."
- "...reports preliminary Q1 adjusted EBITDA €535M; raises FY outlook..."
- "Fundamentals in Residential Systems segment continue to be strong, driven by the new housebuild sector..."
- "A third of dealer network is closed and a third operating with limited capacity..."
- "...have cooled down merger negotiation talks due to the uncertainties surrounding coronavirus..."
- "...comments on a significant decrease of its share price over the last few days..."

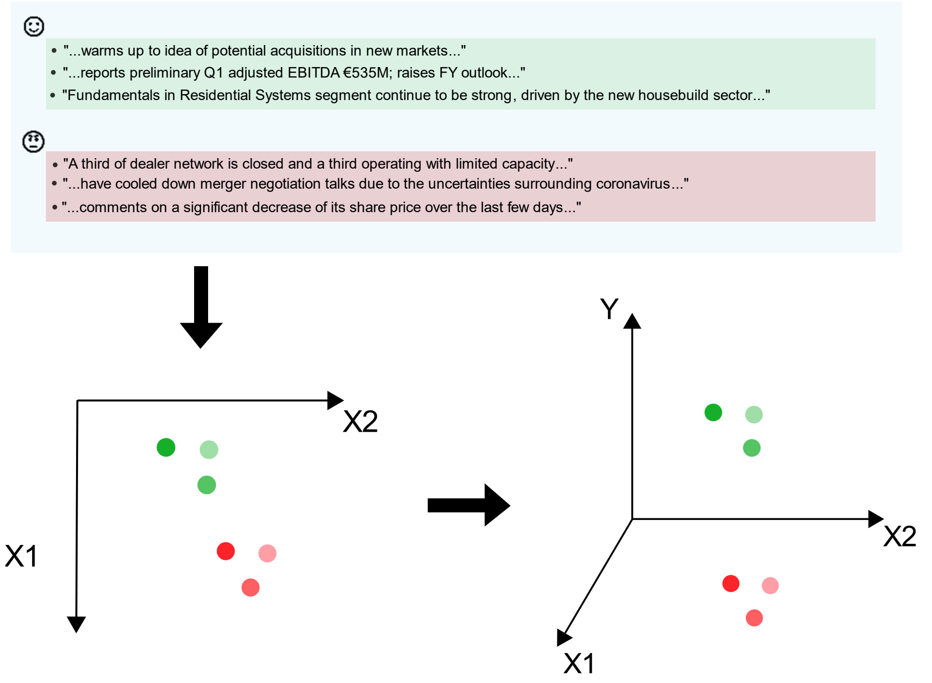


Natural Language Processing

Applications of text embedding

- Text semantic similarity
- Text clustering
- …

- Supervised learning and fine-tuning
  - ○ Sentiments
  - ○ ESG controversies
  - ○ Stock Movements
  - ○ …

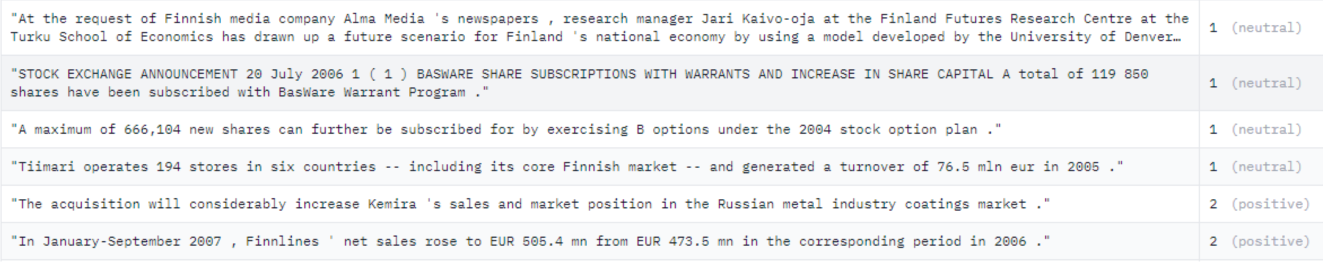


☺

- "...warms up to idea of potential acquisitions in new markets..."
- "...reports preliminary Q1 adjusted EBITDA €535M; raises FY outlook..."
- "Fundamentals in Residential Systems segment continue to be strong, driven by the new housebuild sector..."

☹

- "A third of dealer network is closed and a third operating with limited capacity..."
- "...have cooled down merger negotiation talks due to the uncertainties surrounding coronavirus..."
- "...comments on a significant decrease of its share price over the last few days..."



"At the request of Finnish media company Alma Media 's newspapers , research manager Jari Kaivo-oja at the Finland Futures Research Centre at the Turku School of Economics has drawn up a future scenario for Finland 's national economy by using a model developed by the University of Denver…	1 (neutral)
"STOCK EXCHANGE ANNOUNCEMENT 20 July 2006 1 (1) BASWARE SHARE SUBSCRIPTIONS WITH WARRANTS AND INCREASE IN SHARE CAPITAL A total of 119 850 shares have been subscribed with BasWare Warrant Program ."	1 (neutral)
"A maximum of 666,104 new shares can further be subscribed for by exercising B options under the 2004 stock option plan ."	1 (neutral)
"Tiimari operates 194 stores in six countries -- including its core Finnish market -- and generated a turnover of 76.5 mln eur in 2005 ."	1 (neutral)
"The acquisition will considerably increase Kemira 's sales and market position in the Russian metal industry coatings market ."	2 (positive)
"In January-September 2007 , Finnlines ' net sales rose to EUR 505.4 mn from EUR 473.5 mn in the corresponding period in 2006 ."	2 (positive)

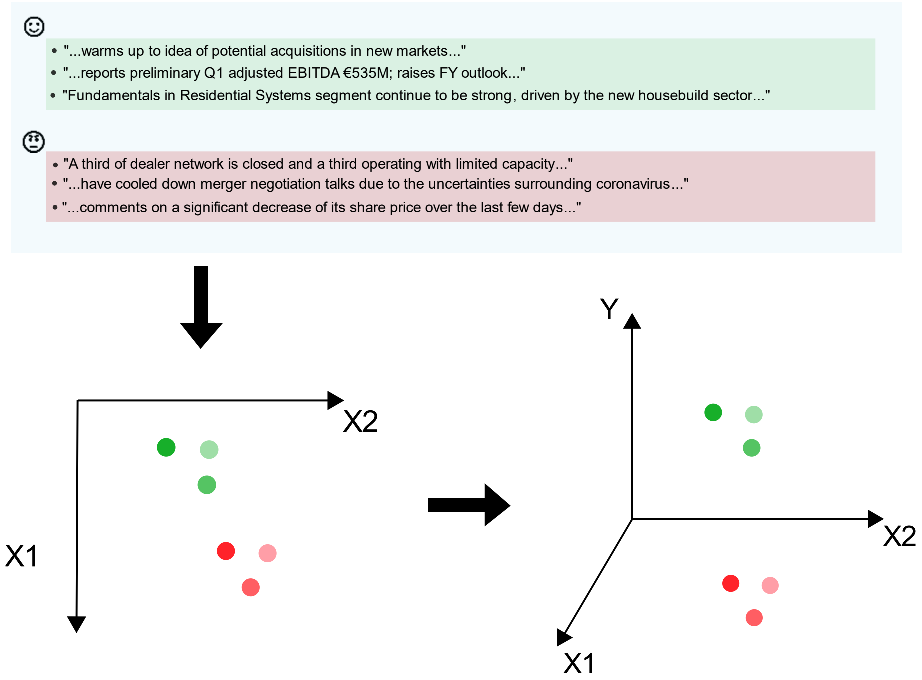
Source: P. Malo and A. Sinha and P. Korhonen and J. Wallenius and P. Takala, https://huggingface.co/datasets/financial_phrasebank/, 2020.

Natural Language Processing

Applications of text embedding

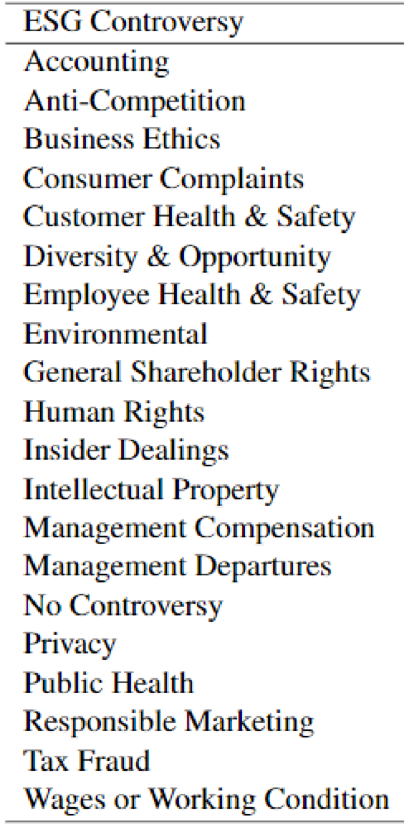
- Text semantic similarity
- Text clustering
- …

- Supervised learning and fine-tuning
  - Sentiments
  - ESG controversies
  - Stock Movements
  - …



☺
- "...warms up to idea of potential acquisitions in new markets..."
- "...reports preliminary Q1 adjusted EBITDA €535M; raises FY outlook..."
- "Fundamentals in Residential Systems segment continue to be strong, driven by the new housebuild sector..."

☹
- "A third of dealer network is closed and a third operating with limited capacity..."
- "...have cooled down merger negotiation talks due to the uncertainties surrounding coronavirus..."
- "...comments on a significant decrease of its share price over the last few days..."



ESG Controversy
Accounting
Anti-Competition
Business Ethics
Consumer Complaints
Customer Health & Safety
Diversity & Opportunity
Employee Health & Safety
Environmental
General Shareholder Rights
Human Rights
Insider Dealings
Intellectual Property
Management Compensation
Management Departures
No Controversy
Privacy
Public Health
Responsible Marketing
Tax Fraud
Wages or Working Condition

Source: Nugent, Tim, Nicole Stelea, and Jochen L. Leidner. "Detecting ESG topics using domain-specific language models and data augmentation approaches." *arXiv preprint arXiv:2010.08319* (2020).

Natural Language Processing

Applications of text embedding

- Text semantic similarity
- Text clustering
- …

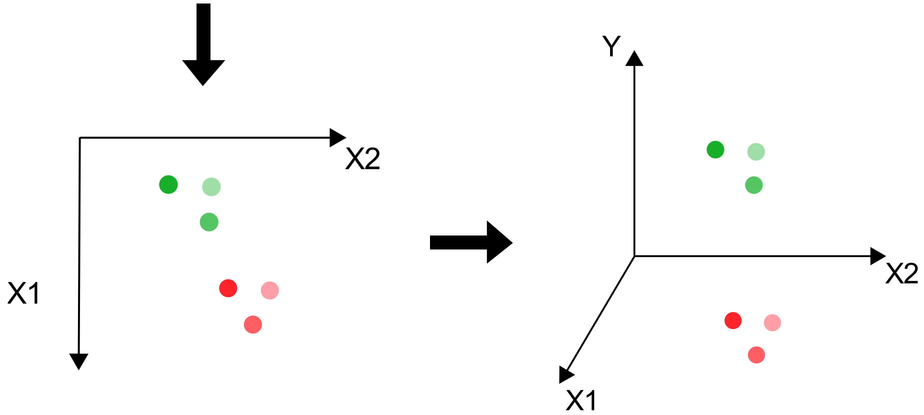
- Supervised learning fine-tuning
  - ○ Sentiments
  - ○ ESG controversies
  - ○ Stock Movements
  - ○ …

☺

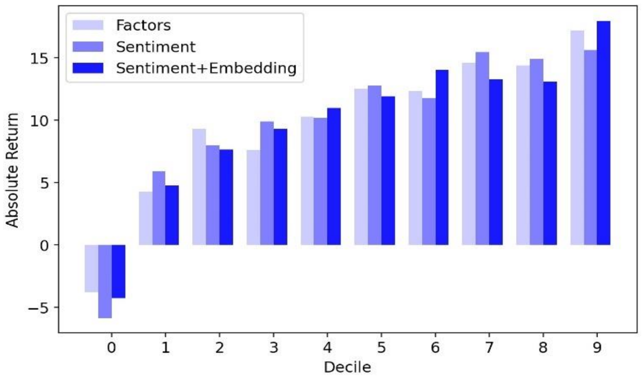
- "...warms up to idea of potential acquisitions in new markets..."
- "...reports preliminary Q1 adjusted EBITDA €535M; raises FY outlook..."
- "Fundamentals in Residential Systems segment continue to be strong, driven by the new housebuild sector..."

☹

- "A third of dealer network is closed and a third operating with limited capacity..."
- "...have cooled down merger negotiation talks due to the uncertainties surrounding coronavirus..."
- "...comments on a significant decrease of its share price over the last few days..."



Source: RAM Active Investments

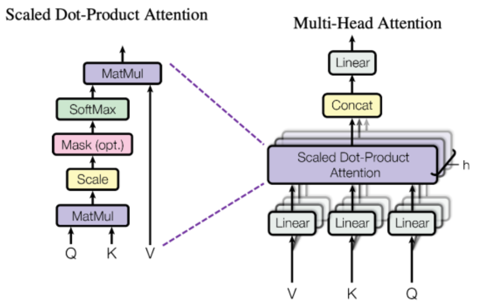


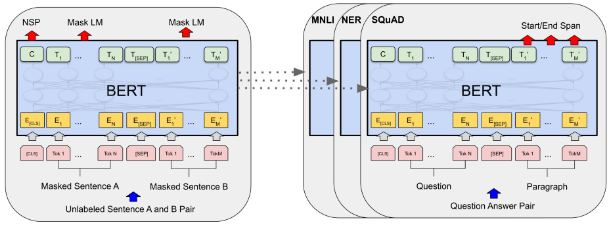
* Past performance is not a reliable indicator of future results.

Summary

$$P(\boldsymbol{Z},\boldsymbol{W};\alpha,\beta)=\int_{\boldsymbol{\theta}}\int_{\boldsymbol{\varphi}}P(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\varphi};\alpha,\beta)\,d\boldsymbol{\varphi}\,d\boldsymbol{\theta}$$

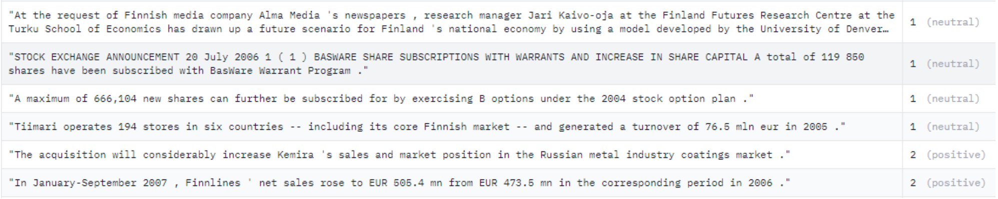
$$P(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\varphi};\alpha,\beta)=\prod_{i=1}^{K}P(\varphi_i;\beta)\prod_{j=1}^{M}P(\theta_j;\alpha)\prod_{t=1}^{N}P(Z_{j,t}\mid\theta_j)P(W_{j,t}\mid\varphi_{Z_{j,t}}),$$





Pre-training

Fine-Tuning



"At the request of Finnish media company Alma Media 's newspapers , research manager Jari Kaivo-oja at the Finland Futures Research Centre at the Turku School of Economics has drawn up a future scenario for Finland 's national economy by using a model developed by the University of Denver…	1 (neutral)
"STOCK EXCHANGE ANNOUNCEMENT 20 July 2006 1 (1) BASWARE SHARE SUBSCRIPTIONS WITH WARRANTS AND INCREASE IN SHARE CAPITAL A total of 119 850 shares have been subscribed with BasWare Warrant Program ."	1 (neutral)
"A maximum of 666,104 new shares can further be subscribed for by exercising B options under the 2004 stock option plan ."	1 (neutral)
"Tiimari operates 194 stores in six countries -- including its core Finnish market -- and generated a turnover of 76.5 mln eur in 2005 ."	1 (neutral)
"The acquisition will considerably increase Kemira 's sales and market position in the Russian metal industry coatings market ."	2 (positive)
"In January-September 2007 , Finnlines ' net sales rose to EUR 505.4 mn from EUR 473.5 mn in the corresponding period in 2006 ."	2 (positive)

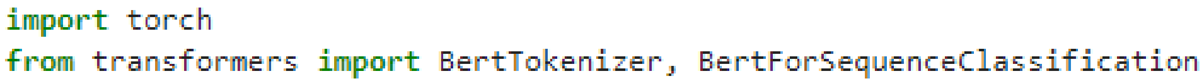
```
import torch
from transformers import BertTokenizer, BertForSequenceClassification
```

Source:https://commons.wikimedia.org/wiki/File:Python_logo_and_wordmark.svg

Demo

“If I had five minutes to chop down a tree, I'd spend the first three sharpening my axe.”

- Abraham Lincoln



```
import torch
from transformers import BertTokenizer, BertForSequenceClassification
```

Demo

Data resources and Development tools

Google Dataset Search

https://datasetsearch.research.google.com/

Ceres SEC Sustainability Disclosure Search Tool

https://tools.ceres.org/resources/tools/sec-sustainability-disclosure/

PyTorch

https://pytorch.org/

Hugging Face Transformers

https://github.com/huggingface/transformers

pandas

https://pandas.pydata.org/

NLTK

https://www.nltk.org/#

Demo

Data resources and Development tools

Ceres SEC Sustainability Disclosure Search Tool

https://tools.ceres.org/resources/tools/sec-sustainability-disclosure/

Hugging Face Transformers

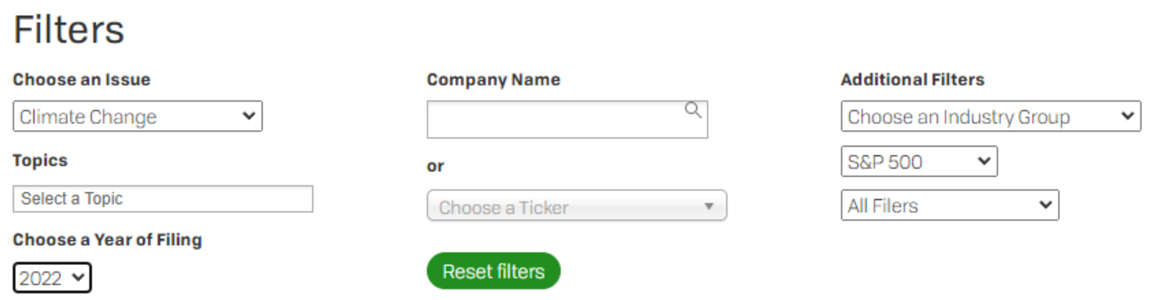
https://github.com/huggingface/transformers

Demo

Data resources and Development tools

Ceres SEC Sustainability Disclosure Search Tool

https://tools.ceres.org/resources/tools/sec-sustainability-disclosure/



Filters

Choose an Issue

Climate Change

Topics

Select a Topic

Choose a Year of Filing

2022

Company Name

or

Choose a Ticker

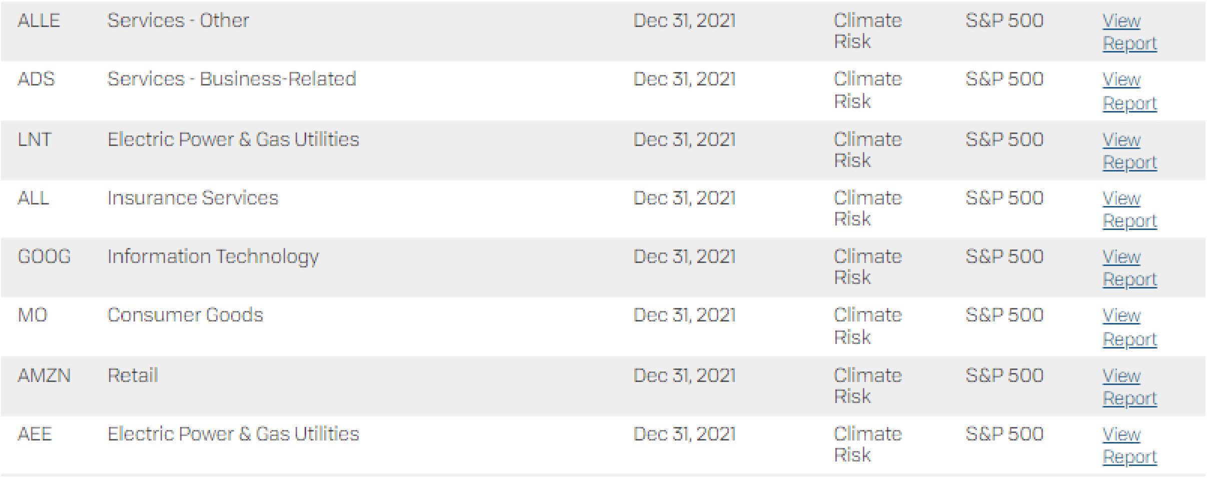
Reset filters

Additional Filters

Choose an Industry Group

S&P 500

All Filers



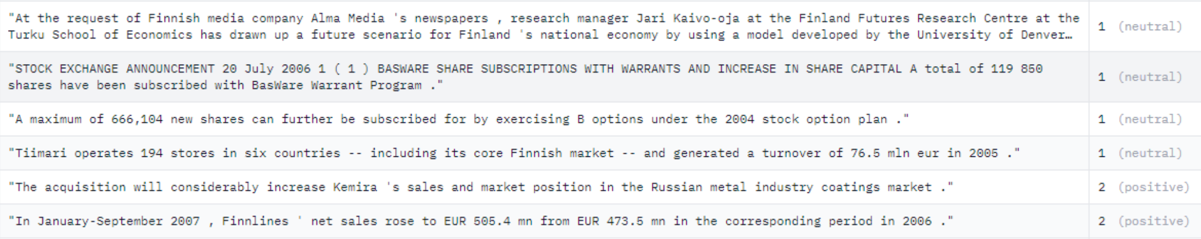
ALLE	Services - Other	Dec 31, 2021	Climate Risk	S&P 500	View Report
ADS	Services - Business-Related	Dec 31, 2021	Climate Risk	S&P 500	View Report
LNT	Electric Power & Gas Utilities	Dec 31, 2021	Climate Risk	S&P 500	View Report
ALL	Insurance Services	Dec 31, 2021	Climate Risk	S&P 500	View Report
GOOG	Information Technology	Dec 31, 2021	Climate Risk	S&P 500	View Report
MO	Consumer Goods	Dec 31, 2021	Climate Risk	S&P 500	View Report
AMZN	Retail	Dec 31, 2021	Climate Risk	S&P 500	View Report
AEE	Electric Power & Gas Utilities	Dec 31, 2021	Climate Risk	S&P 500	View Report

Demo

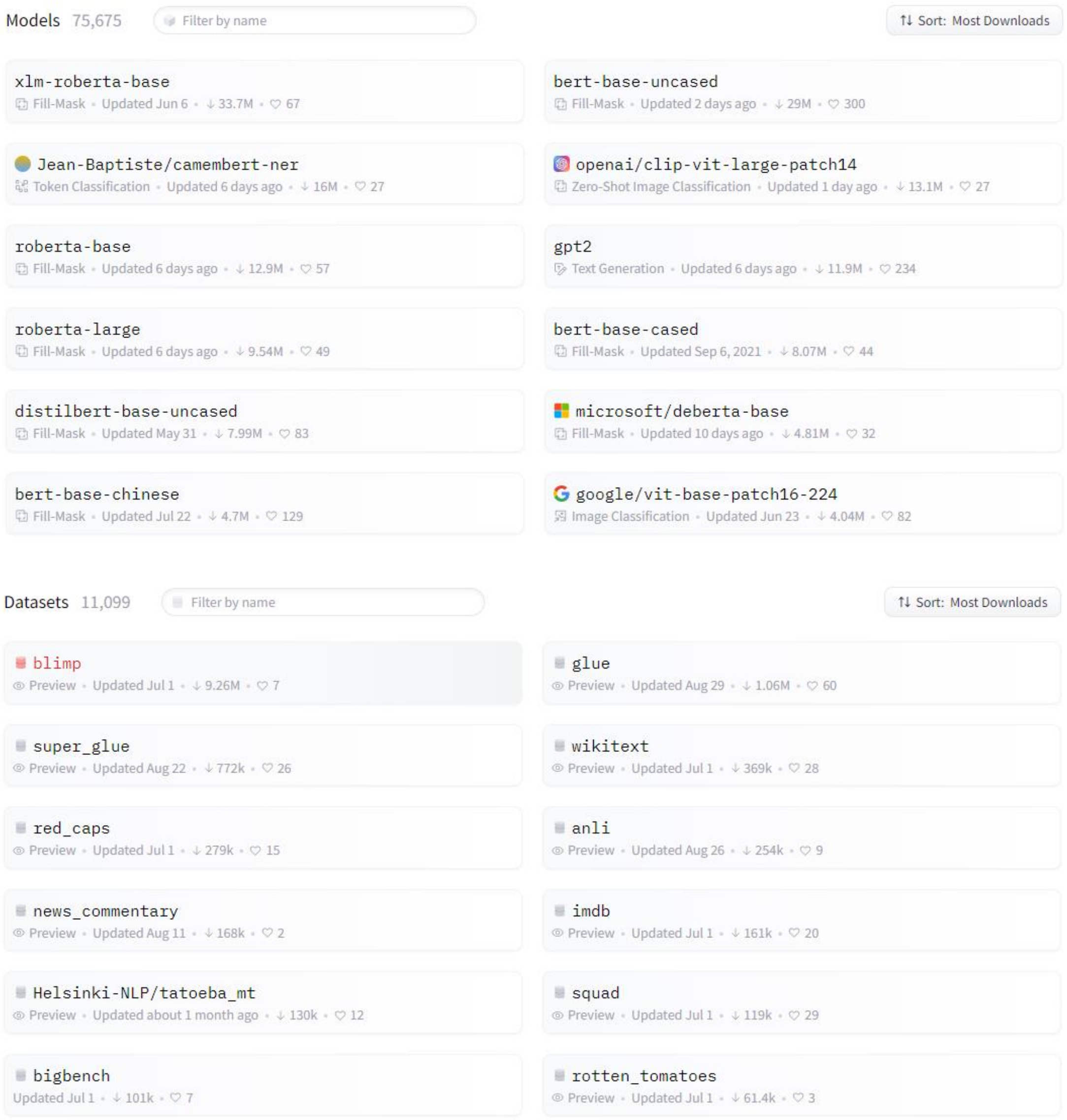
Data resources and Development tools

Hugging Face Transformers

https://github.com/huggingface/transformers



"At the request of Finnish media company Alma Media 's newspapers , research manager Jari Kaivo-oja at the Finland Futures Research Centre at the Turku School of Economics has drawn up a future scenario for Finland 's national economy by using a model developed by the University of Denver…	1 (neutral)
"STOCK EXCHANGE ANNOUNCEMENT 20 July 2006 1 (1) BASWARE SHARE SUBSCRIPTIONS WITH WARRANTS AND INCREASE IN SHARE CAPITAL A total of 119 850 shares have been subscribed with BasWare Warrant Program ."	1 (neutral)
"A maximum of 666,104 new shares can further be subscribed for by exercising B options under the 2004 stock option plan ."	1 (neutral)
"Tiimari operates 194 stores in six countries -- including its core Finnish market -- and generated a turnover of 76.5 mln eur in 2005 ."	1 (neutral)
"The acquisition will considerably increase Kemira 's sales and market position in the Russian metal industry coatings market ."	2 (positive)
"In January-September 2007 , Finnlines ' net sales rose to EUR 505.4 mn from EUR 473.5 mn in the corresponding period in 2006 ."	2 (positive)



Models 75,675

xlm-roberta-base
Fill-Mask • Updated Jun 6 • ↓ 33.7M • ♡ 67

bert-base-uncased
Fill-Mask • Updated 2 days ago • ↓ 29M • ♡ 300

Jean-Baptiste/camembert-ner
Token Classification • Updated 6 days ago • ↓ 16M • ♡ 27

openai/clip-vit-large-patch14
Zero-Shot Image Classification • Updated 1 day ago • ↓ 13.1M • ♡ 27

roberta-base
Fill-Mask • Updated 6 days ago • ↓ 12.9M • ♡ 57

gpt2
Text Generation • Updated 6 days ago • ↓ 11.9M • ♡ 234

roberta-large
Fill-Mask • Updated 6 days ago • ↓ 9.54M • ♡ 49

bert-base-cased
Fill-Mask • Updated Sep 6, 2021 • ↓ 8.07M • ♡ 44

distilbert-base-uncased
Fill-Mask • Updated May 31 • ↓ 7.99M • ♡ 83

microsoft/deberta-base
Fill-Mask • Updated 10 days ago • ↓ 4.81M • ♡ 32

bert-base-chinese
Fill-Mask • Updated Jul 22 • ↓ 4.7M • ♡ 129

google/vit-base-patch16-224
Image Classification • Updated Jun 23 • ↓ 4.04M • ♡ 82

Datasets 11,099

blimp
Preview • Updated Jul 1 • ↓ 9.26M • ♡ 7

glue
Preview • Updated Aug 29 • ↓ 1.06M • ♡ 60

super_glue
Preview • Updated Aug 22 • ↓ 772k • ♡ 26

wikitext
Preview • Updated Jul 1 • ↓ 369k • ♡ 28

red_caps
Preview • Updated Jul 1 • ↓ 279k • ♡ 15

anli
Preview • Updated Aug 26 • ↓ 254k • ♡ 9

news_commentary
Preview • Updated Aug 11 • ↓ 168k • ♡ 2

imdb
Preview • Updated Jul 1 • ↓ 161k • ♡ 20

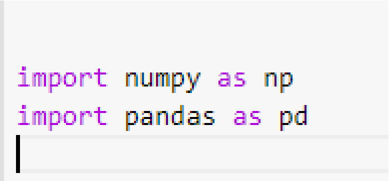
Helsinki-NLP/tatoeba_mt
Preview • Updated about 1 month ago • ↓ 130k • ♡ 12

squad
Preview • Updated Jul 1 • ↓ 119k • ♡ 29

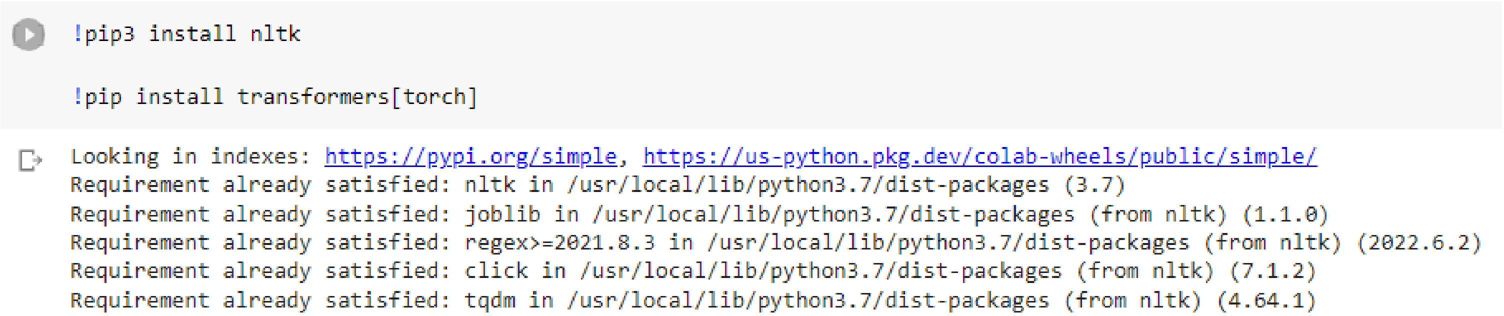
bigbench
Updated Jul 1 • ↓ 101k • ♡ 7

rotten_tomatoes
Preview • Updated Jul 1 • ↓ 61.4k • ♡ 3

Demo



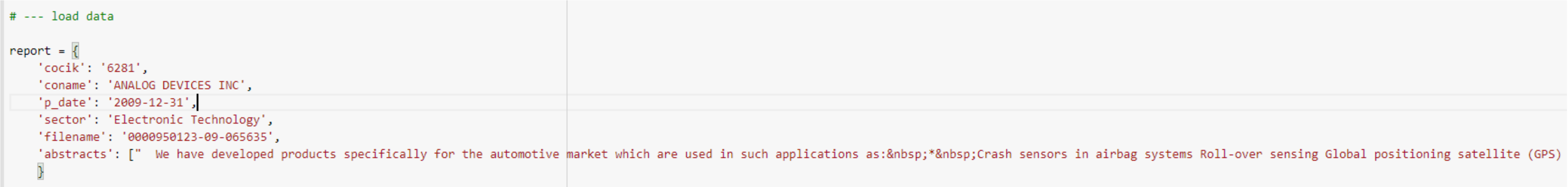
```
import numpy as np
import pandas as pd
```



```
!pip3 install nltk

!pip install transformers[torch]
```

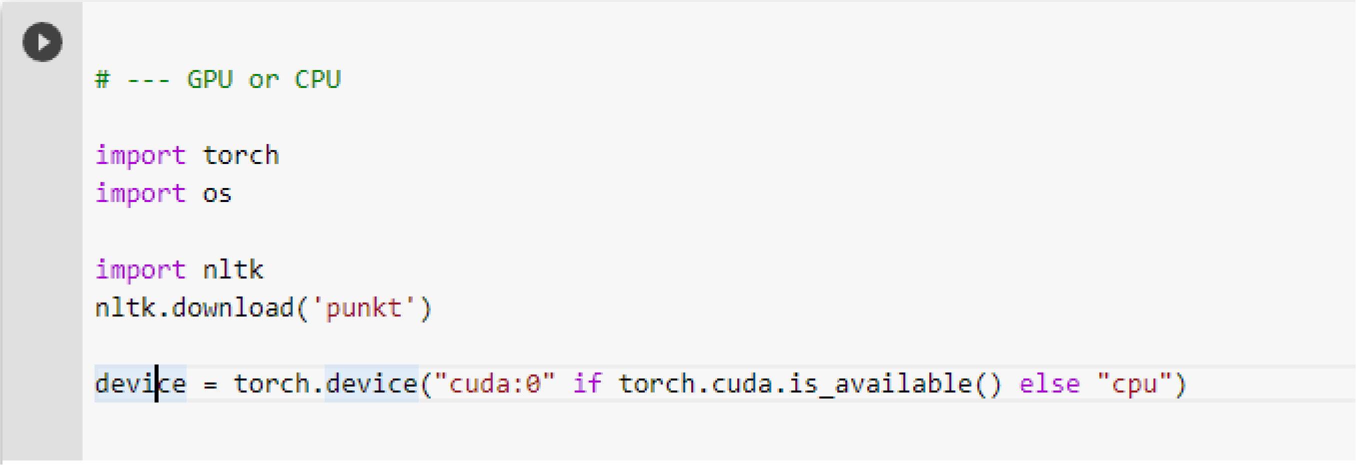
```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: nltk in /usr/local/lib/python3.7/dist-packages (3.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from nltk) (1.1.0)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.7/dist-packages (from nltk) (2022.6.2)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (from nltk) (7.1.2)
Requirement already satisfied: tqdm in /usr/local/lib/python3.7/dist-packages (from nltk) (4.64.1)
```

Demo on Google Colaboratory

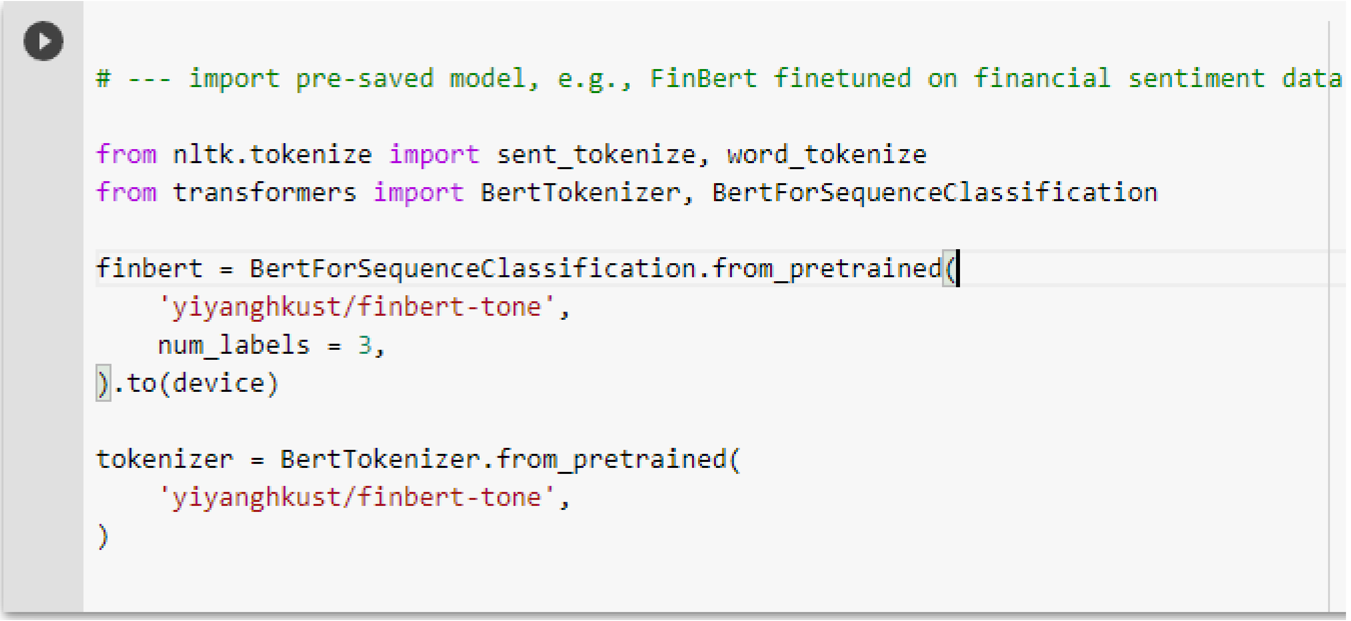


```
# --- load data

report = {
    'cocik': '6281',
    'coname': 'ANALOG DEVICES INC',
    'p_date': '2009-12-31',
    'sector': 'Electronic Technology',
    'filename': '0000950123-09-065635',
    'abstracts': ["  We have developed products specifically for the automotive market which are used in such applications as:&nbsp;*&nbsp;Crash sensors in airbag systems Roll-over sensing Global positioning satellite (GPS)
    }
```

Demo



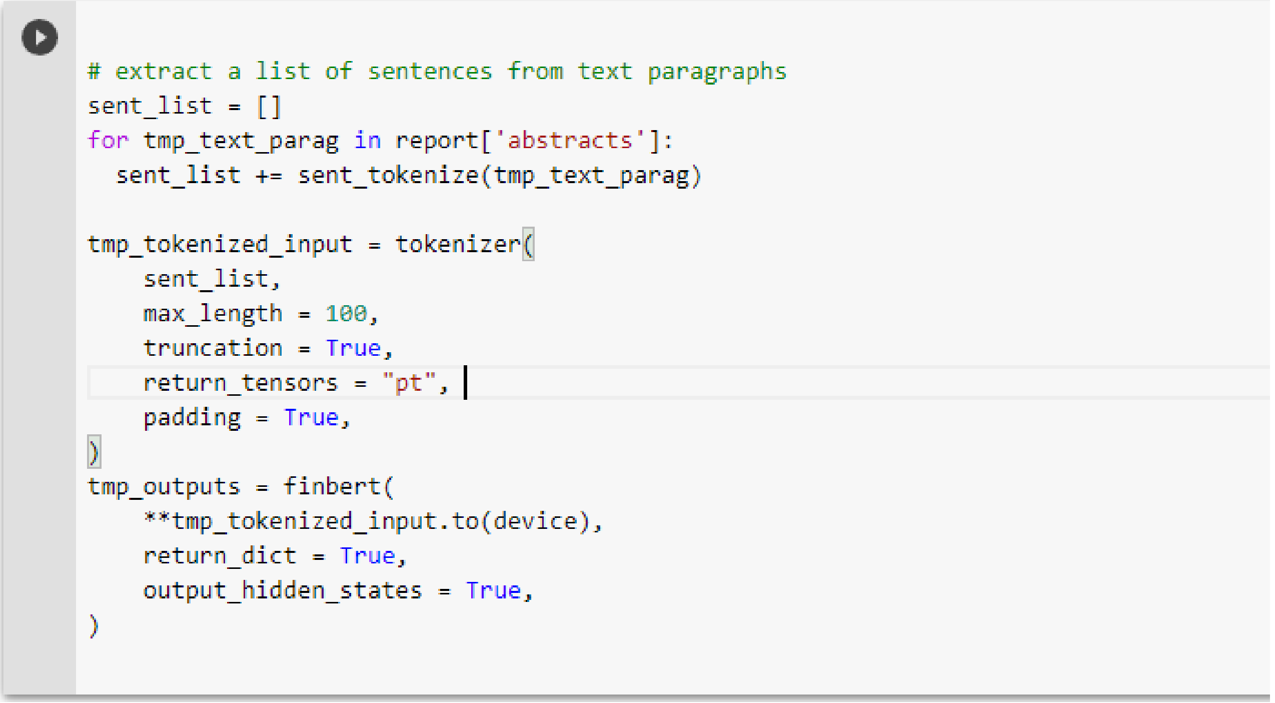
```
# --- GPU or CPU

import torch
import os

import nltk
nltk.download('punkt')

device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")
```



```
# --- import pre-saved model, e.g., FinBert finetuned on financial sentiment data

from nltk.tokenize import sent_tokenize, word_tokenize
from transformers import BertTokenizer, BertForSequenceClassification

finbert = BertForSequenceClassification.from_pretrained(
    'yiyanghkust/finbert-tone',
    num_labels = 3,
).to(device)

tokenizer = BertTokenizer.from_pretrained(
    'yiyanghkust/finbert-tone',
)
```

Demo on Google Colaboratory

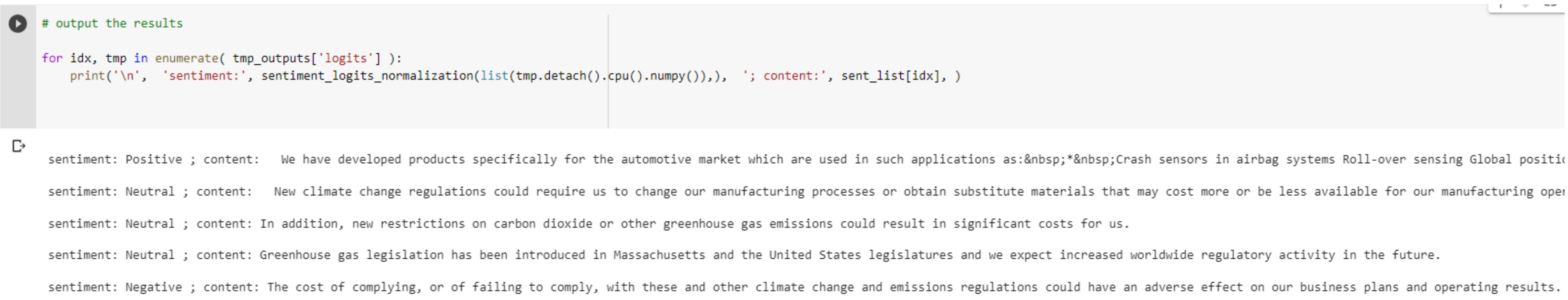
Demo



```
# extract a list of sentences from text paragraphs
sent_list = []
for tmp_text_parag in report['abstracts']:
  sent_list += sent_tokenize(tmp_text_parag)

tmp_tokenized_input = tokenizer(
    sent_list,
    max_length = 100,
    truncation = True,
    return_tensors = "pt",
    padding = True,
)
tmp_outputs = finbert(
    **tmp_tokenized_input.to(device),
    return_dict = True,
    output_hidden_states = True,
)
```

Demo on Google Colaboratory

Demo



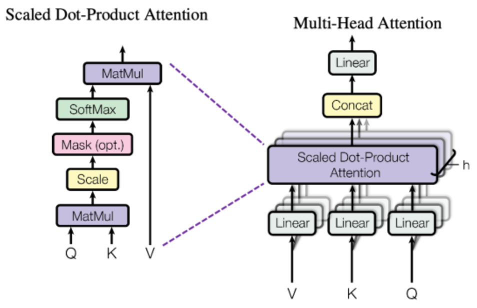
```
# output the results

for idx, tmp in enumerate( tmp_outputs['logits'] ):
    print('\n',  'sentiment:', sentiment_logits_normalization(list(tmp.detach().cpu().numpy()),),  '; content:', sent_list[idx], )
```

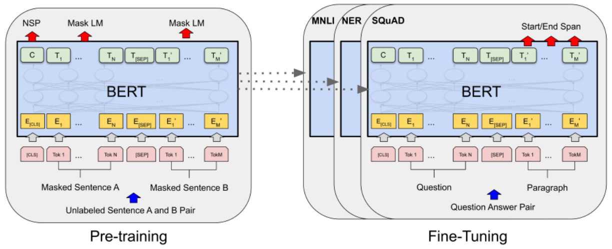
```
 sentiment: Positive ; content:   We have developed products specifically for the automotive market which are used in such applications as:&nbsp;*&nbsp;Crash sensors in airbag systems Roll-over sensing Global positi

 sentiment: Neutral ; content:   New climate change regulations could require us to change our manufacturing processes or obtain substitute materials that may cost more or be less available for our manufacturing ope

 sentiment: Neutral ; content: In addition, new restrictions on carbon dioxide or other greenhouse gas emissions could result in significant costs for us.

 sentiment: Neutral ; content: Greenhouse gas legislation has been introduced in Massachusetts and the United States legislatures and we expect increased worldwide regulatory activity in the future.

 sentiment: Negative ; content: The cost of complying, or of failing to comply, with these and other climate change and emissions regulations could have an adverse effect on our business plans and operating results.
```

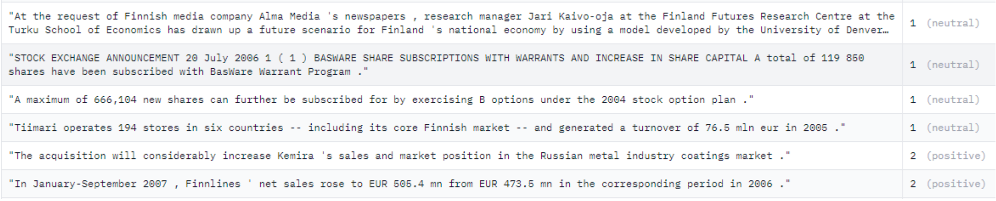
Conclusion

$$P(\boldsymbol{Z},\boldsymbol{W};\alpha,\beta)=\int_{\theta}\int_{\varphi}P(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\varphi};\alpha,\beta)\,d\boldsymbol{\varphi}\,d\boldsymbol{\theta}$$

$$P(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\varphi};\alpha,\beta)=\prod_{i=1}^{K}P(\varphi_i;\beta)\prod_{j=1}^{M}P(\theta_j;\alpha)\prod_{t=1}^{N}P(Z_{j,t}\mid\theta_j)P(W_{j,t}\mid\varphi_{Z_{j,t}}),$$

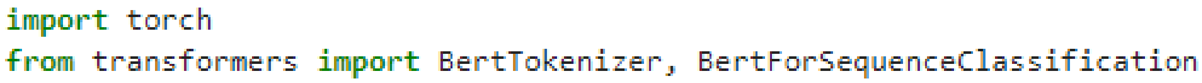






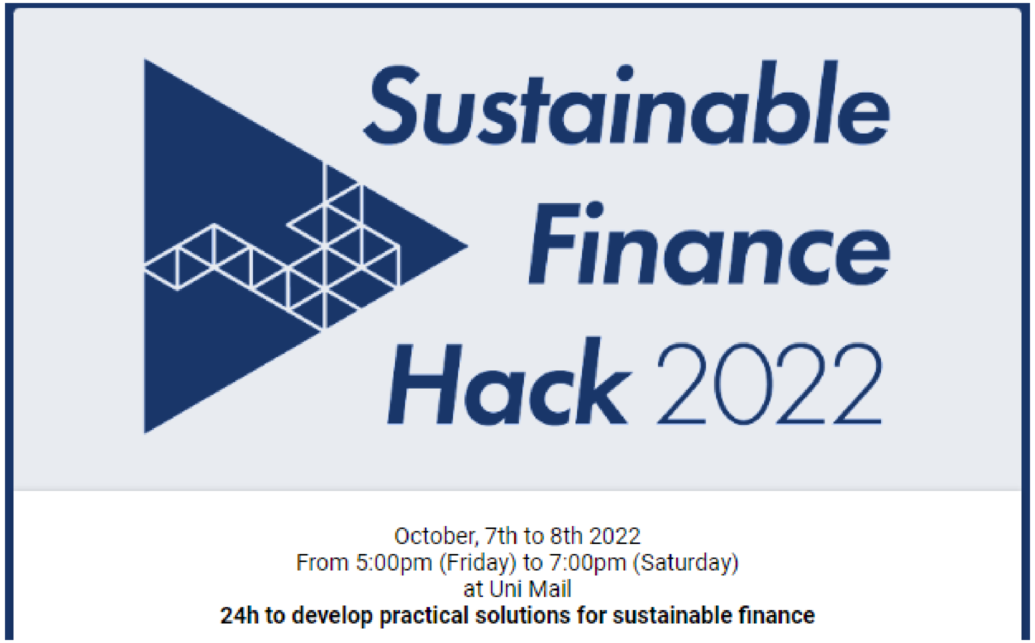
"At the request of Finnish media company Alma Media 's newspapers , research manager Jari Kaivo-oja at the Finland Futures Research Centre at the Turku School of Economics has drawn up a future scenario for Finland 's national economy by using a model developed by the University of Denver…	1 (neutral)
"STOCK EXCHANGE ANNOUNCEMENT 20 July 2006 1 (1) BASWARE SHARE SUBSCRIPTIONS WITH WARRANTS AND INCREASE IN SHARE CAPITAL A total of 119 850 shares have been subscribed with BasWare Warrant Program ."	1 (neutral)
"A maximum of 666,104 new shares can further be subscribed for by exercising B options under the 2004 stock option plan ."	1 (neutral)
"Tiimari operates 194 stores in six countries -- including its core Finnish market -- and generated a turnover of 76.5 mln eur in 2005 ."	1 (neutral)
"The acquisition will considerably increase Kemira 's sales and market position in the Russian metal industry coatings market ."	2 (positive)
"In January-September 2007 , Finnlines ' net sales rose to EUR 505.4 mn from EUR 473.5 mn in the corresponding period in 2006 ."	2 (positive)

Source:https://commons.wikimedia.org/wiki/File:Python_logo_and_wordmark.svg



```
import torch
from transformers import BertTokenizer, BertForSequenceClassification
```

Source: generated by https://beta.dreamstudio.ai/dream via the Stable Diffusion technique.



Sustainable Finance Hack 2022

October, 7th to 8th 2022
From 5:00pm (Friday) to 7:00pm (Saturday)
at Uni Mail
24h to develop practical solutions for sustainable finance

https://sfh22.sparkboard.com/projects

Build a Bad Buzz Factory for ESG Controversy Detection!

What is a Bad Buzz factory?

It is a machine learning infrastructure that generates artificial news articles about companies involved in ESG controversies.

What are ESG controversies?

They are negative events related to environmental, social and governance (ESG) topics.

Why do we need a Bad Buzz factory?

We aim to detect ESG controversies as soon as they occur to improve the sustainability of our equity portfolios. Training a machine learning model is an efficient method to detect such controversies within the news flow in an automatic and data-driven way. However, a large dataset is needed for the model training, and real-life news articles pose a real challenge in terms of data availability, labelling, copyrights, etc.

Could the solution be a Bad Buzz Factory that synthetically generates all the articles needed to train our model?

Disclaimer

This marketing document is only provided for information purposes to professional clients, and it does not constitute an offer, investment advice or a solicitation to subscribe shares in any jurisdiction where such an offer or solicitation would not be authorised or it would be unlawful. Past performance is not a guide to current or future results. There is no guarantee to get back the full amount invested. Particular attention is paid to the contents of this document but no guarantee, warranty or representation, express or implied, is given to the accuracy, correctness or completeness thereof. Issued in Switzerland by RAM Active Investments S.A. which is authorised and regulated in Switzerland by the Swiss Financial Market Supervisory Authority (FINMA).