Financial News Sentiment Classification & Analysis

Project Implementation

Team Members

- Zijie Ku (zijieku2) - Coordinator
    - Seeking Alpha scrapper
    - Google AutoML model training
    - Documentation
    - Presentation
- Yanbin Zhang (zhang50)
    - Forbes scrapper
    - Google AutoML model training, testing, validation
    - Documentation

Introduction

As described in the Course Proposal, our main goal is to create a sentiment analysis tool for financial news. In this course project, we utilized several available Financial APIs to achieve this. Even though we countered various technical difficulties, including but not limited to Google's service disruptions in certain regions throughout the development stages, service throttling from the news source, etc., we are able reach our goal to train models to classify and to analyze the sentiment of the financial news. With our tool, user can get a general sense of the sentiment of new article(s).

Implementation Phases

The summary of the whole implementation process can be broken down in the following phases:

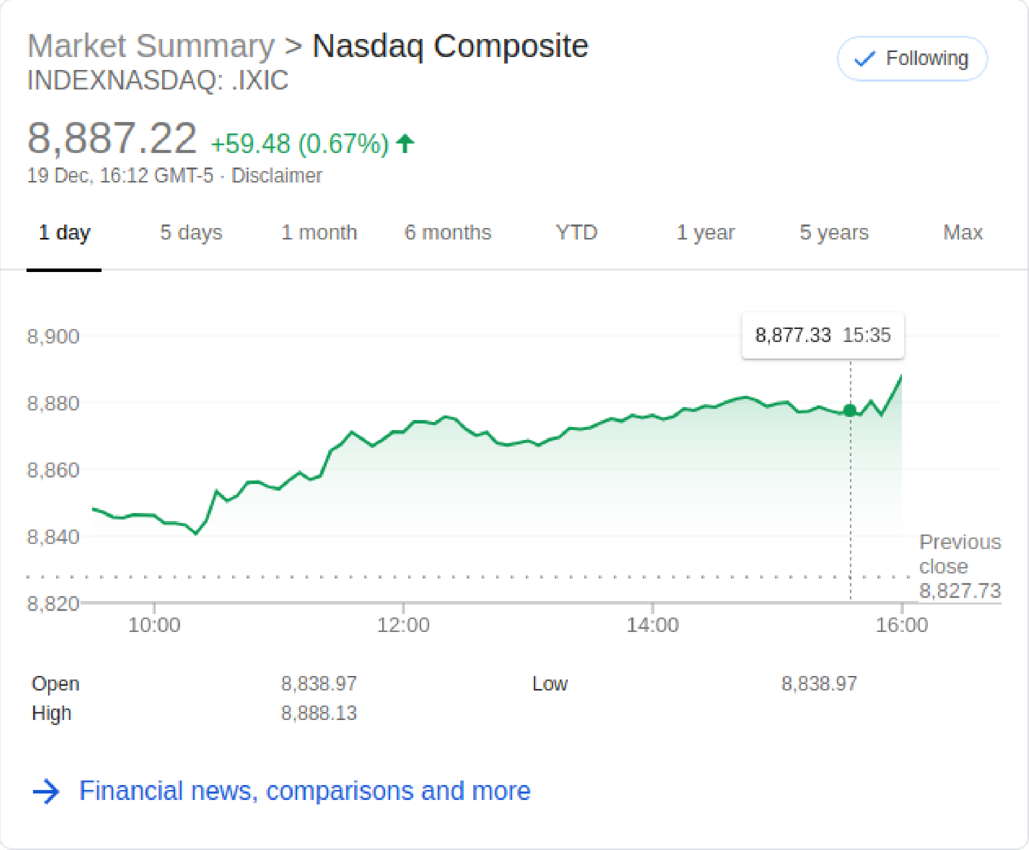
- Phase I - Fetch top mentioned tickers with Stock News API from the past 30 days.
- Phase II - Retrieve relevant historical news along with labeled sentiment data from the past 60 days.
- Phase III - Use machine learning techniques to train models with Google AutoML.
- Phase IV - Test and validation on the trained models.

Analysis Results

According to the text classification, three categories are classified, namely "positive", "neutral", and "negative". In our project, the sentiment analysis model has three categories, 0 - "positive", 1 - "neutral", and 2 - "negative" correspondingly.

While the Nasdaq composite has a positive gain at the moment of writing this documentation, but Uber's stock price is clearly affected by this 'negative' (2) news about its Ex-CEO, Travis.

- **Nasdaq composite**: Bullish market

- 

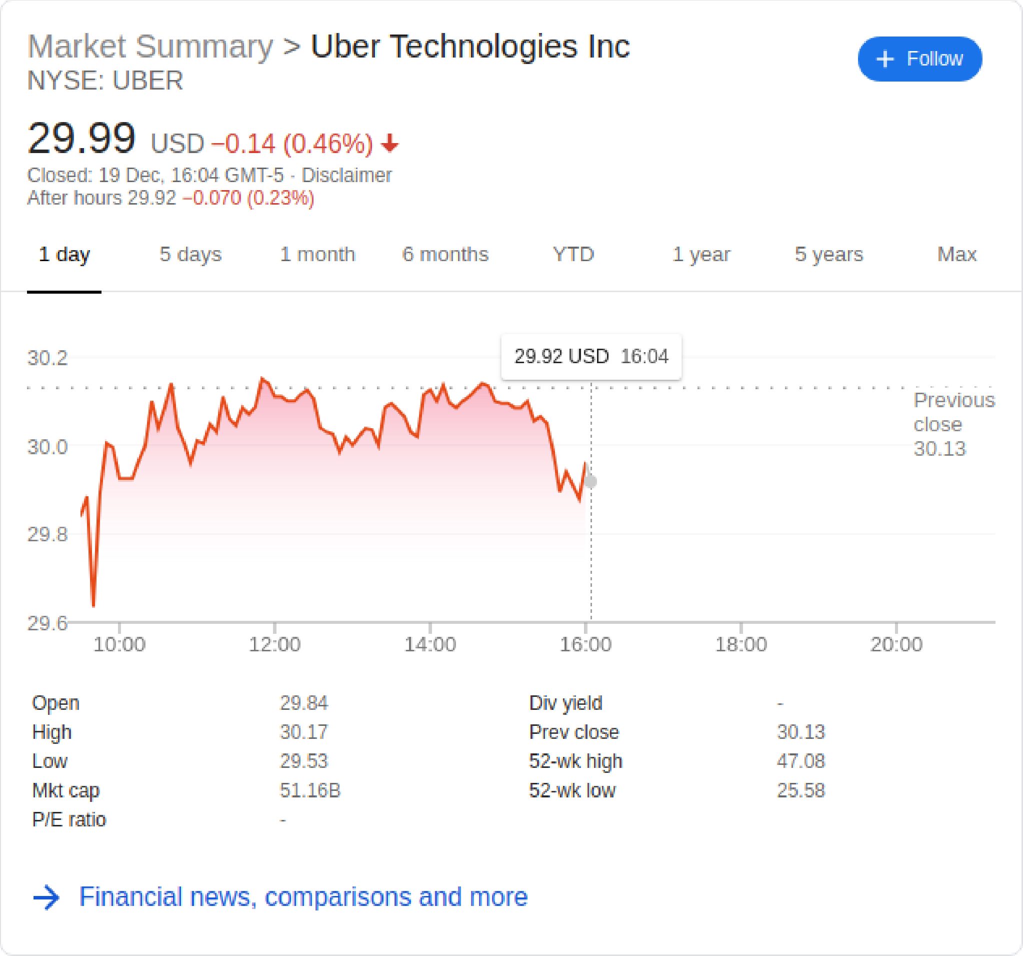
Market Summary > Nasdaq Composite
INDEXNASDAQ: .IXIC

8,887.22 +59.48 (0.67%) ↑
19 Dec, 16:12 GMT-5 · Disclaimer

Open	8,838.97	Low	8,838.97
High	8,888.13		

→ Financial news, comparisons and more

- **Uber**: Bearish market

- 

Market Summary > Uber Technologies Inc
NYSE: UBER

29.99 USD −0.14 (0.46%) ↓
Closed: 19 Dec, 16:04 GMT-5 · Disclaimer
After hours 29.92 −0.070 (0.23%)

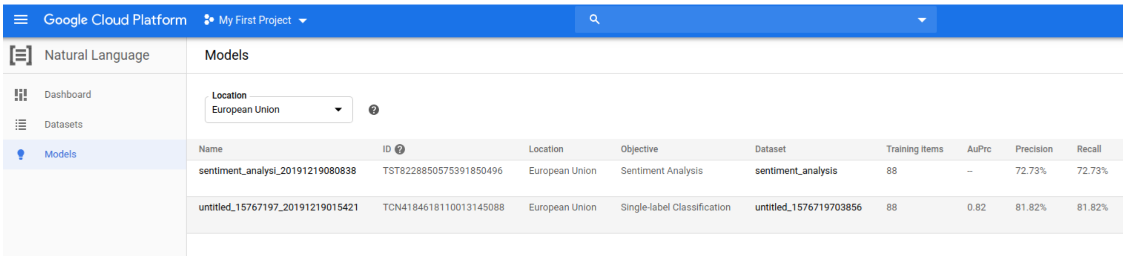
Open	29.84	Div yield	-
High	30.17	Prev close	30.13
Low	29.53	52-wk high	47.08
Mkt cap	51.16B	52-wk low	25.58
P/E ratio	-		

→ Financial news, comparisons and more

Testing Results

Models

All models will be available under the `Models` section as following.

- 

Models

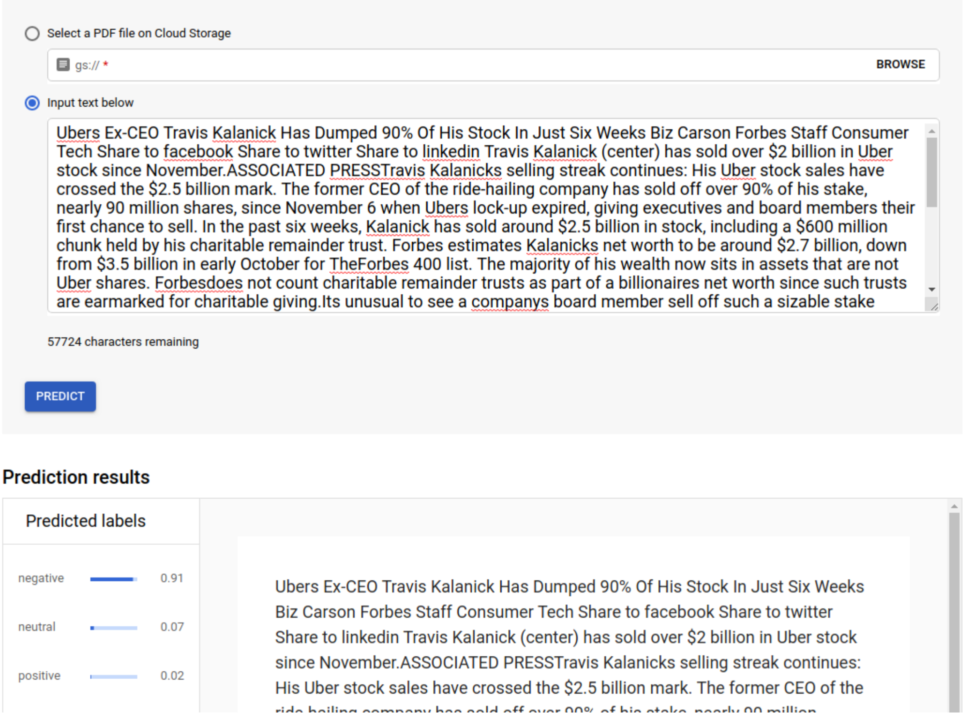
Location: European Union

Name	ID	Location	Objective	Dataset	Training items	AuPrc	Precision	Recall
sentiment_analysi_20191219080838	TST8228850575391850496	European Union	Sentiment Analysis	sentiment_analysis	88	--	72.73%	72.73%
untitled_15767197_20191219015421	TCN4184618110013145088	European Union	Single-label Classification	untitled_1576719703856	88	0.82	81.82%	81.82%

Scrape the latest financial news about top mentioned stocks

- Use `get_stock_news(top_stocks, today=True)` to fetch latest news. (Or use any news if the user would like to as an alternative input).
  - e.g:

```
top stocks: ['GOOGL', 'FB', 'UBER', 'MSFT', 'NVDA', 'INTC', 'T', 'CRM',
'AMD', 'AAPL']
UBER has 2 news
        >> url:
https://www.forbes.com/sites/elanagross/2019/12/19/uber-settles-with-
the-eeoc-to-the-tune-of-44-million/
        >> url: https://www.forbes.com/sites/bizcarson/2019/12/18/uber-
travis-kalanick-sells-most-of-his-shares/
        completed scapping for UBER
```

- Go to `Test & Use`
  - **AutoML Text & Document Classification**:
    - 

Ubers Ex-CEO Travis Kalanick Has Dumped 90% Of His Stock In Just Six Weeks Biz Carson Forbes Staff Consumer Tech Share to facebook Share to twitter Share to linkedin Travis Kalanick (center) has sold over $2 billion in Uber stock since November.ASSOCIATED PRESSTravis Kalanicks selling streak continues: His Uber stock sales have crossed the $2.5 billion mark. The former CEO of the ride-hailing company has sold off over 90% of his stake, nearly 90 million shares, since November 6 when Ubers lock-up expired, giving executives and board members their first chance to sell. In the past six weeks, Kalanick has sold around $2.5 billion in stock, including a $600 million chunk held by his charitable remainder trust. Forbes estimates Kalanicks net worth to be around $2.7 billion, down from $3.5 billion in early October for TheForbes 400 list. The majority of his wealth now sits in assets that are not Uber shares. Forbesdoes not count charitable remainder trusts as part of a billionaires net worth since such trusts are earmarked for charitable giving.Its unusual to see a companys board member sell off such a sizable stake

57724 characters remaining

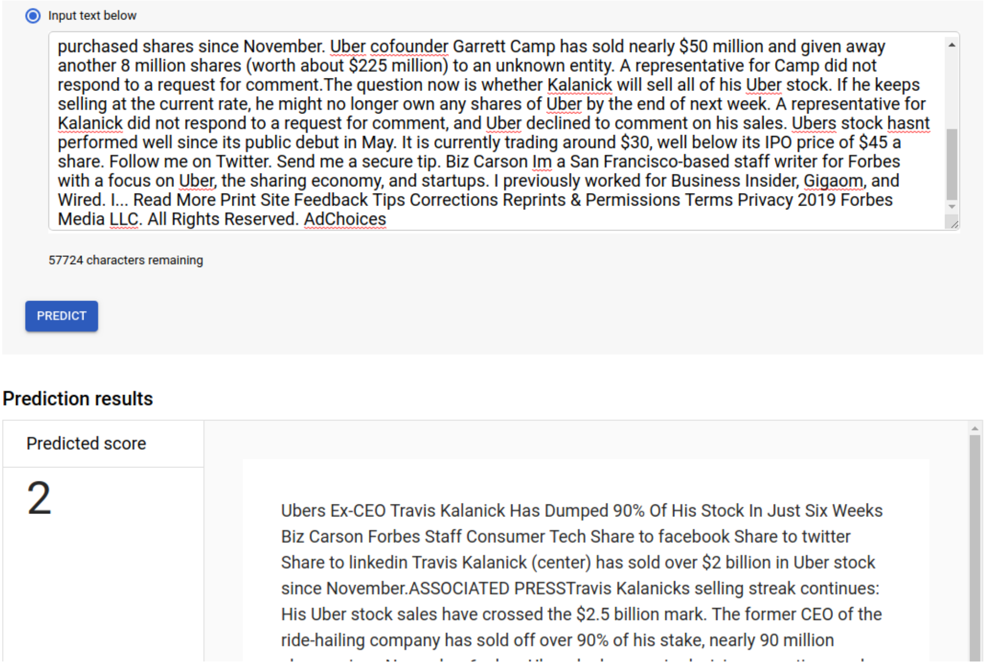
Prediction results

Predicted labels

negative	0.91
neutral	0.07
positive	0.02

Ubers Ex-CEO Travis Kalanick Has Dumped 90% Of His Stock In Just Six Weeks Biz Carson Forbes Staff Consumer Tech Share to facebook Share to twitter Share to linkedin Travis Kalanick (center) has sold over $2 billion in Uber stock since November.ASSOCIATED PRESSTravis Kalanicks selling streak continues: His Uber stock sales have crossed the $2.5 billion mark. The former CEO of the ride-hailing company has sold off over 90% of his stake, nearly 90 million

  - **AutoML Sentiment Analysis**:



○ Input text below

purchased shares since November. Uber cofounder Garrett Camp has sold nearly $50 million and given away another 8 million shares (worth about $225 million) to an unknown entity. A representative for Camp did not respond to a request for comment.The question now is whether Kalanick will sell all of his Uber stock. If he keeps selling at the current rate, he might no longer own any shares of Uber by the end of next week. A representative for Kalanick did not respond to a request for comment, and Uber declined to comment on his sales. Ubers stock hasnt performed well since its public debut in May. It is currently trading around $30, well below its IPO price of $45 a share. Follow me on Twitter. Send me a secure tip. Biz Carson Im a San Francisco-based staff writer for Forbes with a focus on Uber, the sharing economy, and startups. I previously worked for Business Insider, Gigaom, and Wired. I... Read More Print Site Feedback Tips Corrections Reprints & Permissions Terms Privacy 2019 Forbes Media LLC. All Rights Reserved. AdChoices

57724 characters remaining

PREDICT

Prediction results

Predicted score
2

Ubers Ex-CEO Travis Kalanick Has Dumped 90% Of His Stock In Just Six Weeks Biz Carson Forbes Staff Consumer Tech Share to facebook Share to twitter Share to linkedin Travis Kalanick (center) has sold over $2 billion in Uber stock since November.ASSOCIATED PRESSTravis Kalanicks selling streak continues: His Uber stock sales have crossed the $2.5 billion mark. The former CEO of the ride-hailing company has sold off over 90% of his stake, nearly 90 million

- As we can see, both the classification and sentiment analysis suggested the same prediction for the sample news as "negative" opinion.

Software Tutorial

Setup

Scrapper

Install new packages

```
# Ensure your pip is up to date
pip install --upgrade pip

# install Beautiful Soup
pip install bs4

# install requests
pip install requests
```

Required packages

```
import time # required to sleep for n seconds if encounter web service throttling
import requests # required to make get request to API
import datetime import datetime, timedelta # required to calculate time frame
```

API: Stock News API

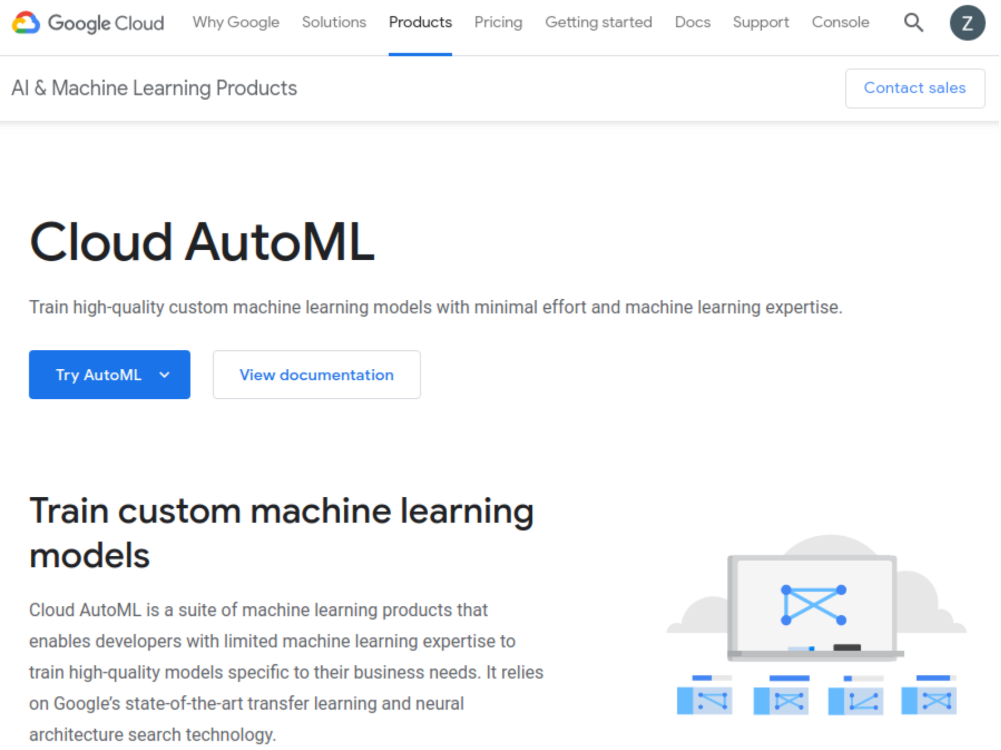
- API Documentation
- Available News Sources:
    - 24/7 Wall Street, Benzinga, Bloomberg Markets and Finance, Bloomberg Technology, Business Insider, CNBC, CNBC International TV, CNBC Tlevision, CNET, CNN, CNN Business, Digital Trends, Deadline, Engadget, ETF Trends, Fast Company, **Forbs**, Fox Business, GeekWire, Globe News Wire, Investopedia, Investors Business Daily, Market Watch, New York Post, New York Times, Reuters, **Seeking Alpha**, TechCrunch, The Motley Fool, The Street, Wall Street Journal, Yahoo Finance

Tasks

1. `get_top_mentioned_stocks_last30days(sector=All)`
    - Retrieves the top mentioned company tickers from the aggregated news sources.
    - Different `sectors` are available: default is set to `All`
        - For the purpose of this project, a familiar industry is chosen, namely `Technology`
    - A default `30days` time frame was chosen arbitrarily. It could be extended much longer.
2. `get_stock_news (stocks, today=False)`
    - Retrieves historical news for the given stocks in a `list` form.

    - `today` is set to `False` for retrieving news from the `INCLUDE_SRC` news source for the past `60days`.
        - historical news time frame usually should be longer than the top mentioned stock time frame.
        - certain news source could be added or removed.
3. `scrape_page_[news source](url)`
    - Write custom scrapper for each news source.
        - For the purpose of this project, **Seeking Alpha** and **Forbs** were chosen arbitrarily.
4. Different output files are written folders with specific file hierarchy
    - Will be explained further in details under Google AutoML Section

Google AutoML

- API Documentation
- Registration
    - Google AutoML Home Page:
        - 

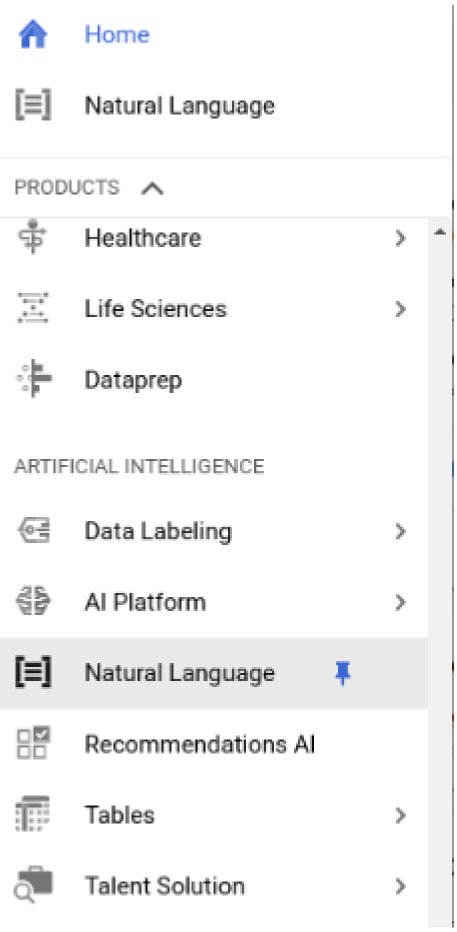
Cloud AutoML

Train high-quality custom machine learning models with minimal effort and machine learning expertise.

Train custom machine learning models

Cloud AutoML is a suite of machine learning products that enables developers with limited machine learning expertise to train high-quality models specific to their business needs. It relies on Google's state-of-the-art transfer learning and neural architecture search technology.

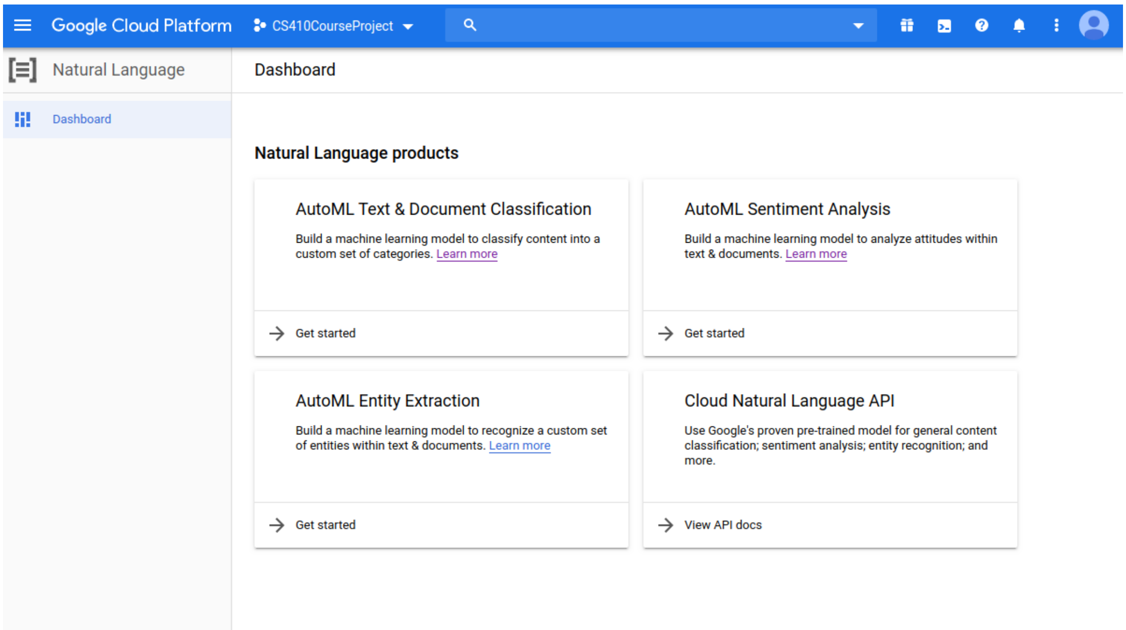
    - Browse to `Console` tag on the top-right-hand corner. Choose `Natural Language` under `ARTIFICIAL INTELLIGENCE` tag from the top-left-hand hamburger icon.

- 

Document Preparation & Uploads

Available Products

Four types of services are available:

- 

Natural Language products

AutoML Text & Document Classification

Build a machine learning model to classify content into a custom set of categories. Learn more

Get started

AutoML Sentiment Analysis

Build a machine learning model to analyze attitudes within text & documents. Learn more

Get started

AutoML Entity Extraction

Build a machine learning model to recognize a custom set of entities within text & documents. Learn more

Get started

Cloud Natural Language API

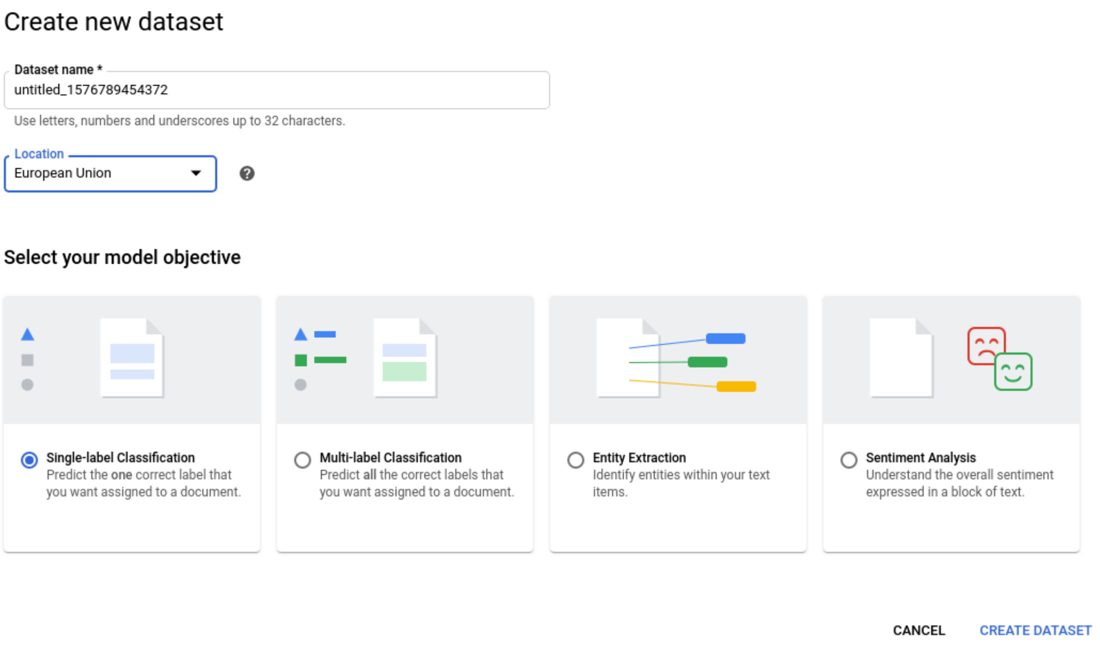
Use Google's proven pre-trained model for general content classification; sentiment analysis; entity recognition; and more.

View API docs

1. **AutoML Text & Document Classification**
2. **AutoML Sentiment Analysis**
3. AutoML Entity Extraction
4. Cloud Natural Language API

- For the purpose of this project, first two were chosen.

AutoML Text & Document Classification

- Choose `Get started` to initialize the project
- Choose `New Dataset`
- Choose `Single-label Classification` for sentiment classification
- 

  Create new dataset

  Dataset name *
  untitled_1576789454372
  Use letters, numbers and underscores up to 32 characters.

  Location
  European Union

  Select your model objective

  ● Single-label Classification
  Predict the **one** correct label that you want assigned to a document.

  ○ Multi-label Classification
  Predict **all** the correct labels that you want assigned to a document.

  ○ Entity Extraction
  Identify entities within your text items.

  ○ Sentiment Analysis
  Understand the overall sentiment expressed in a block of text.

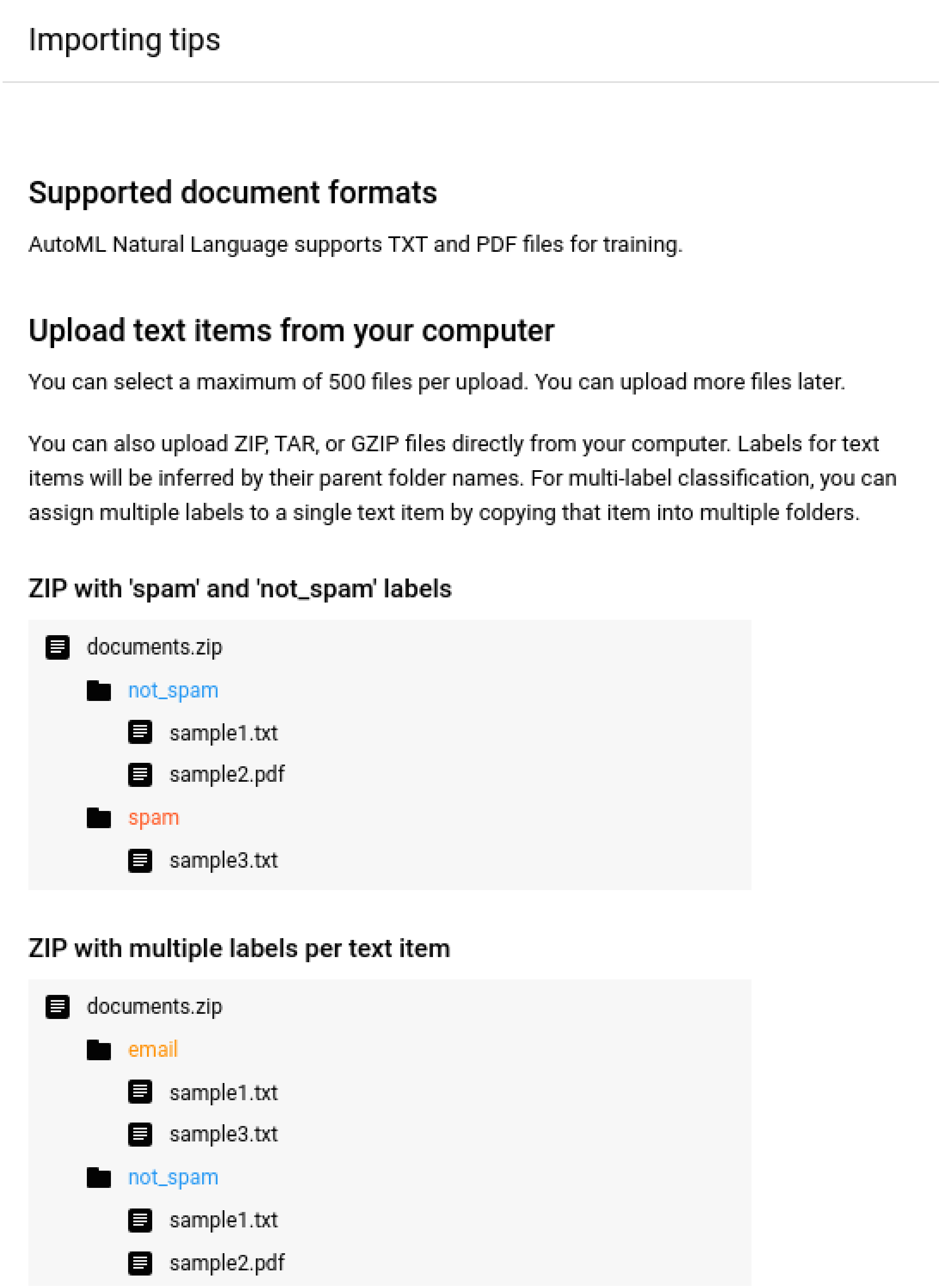
- Worth noticing that:
  - `Dataset name` cannot be modified later.
  - `Location` is important as newly release AutoML service is extremely unstable. Hosting service on regional server may have a better performance and availability than that provided by the default `Global` scope.

AutoML Sentiment Analysis

- Similar to the **AutoML Text & Document Classification** step.

Document Uploads

-



Importing tips

Supported document formats

AutoML Natural Language supports TXT and PDF files for training.

Upload text items from your computer

You can select a maximum of 500 files per upload. You can upload more files later.

You can also upload ZIP, TAR, or GZIP files directly from your computer. Labels for text items will be inferred by their parent folder names. For multi-label classification, you can assign multiple labels to a single text item by copying that item into multiple folders.

ZIP with 'spam' and 'not_spam' labels

- documents.zip
  - not_spam
    - sample1.txt
    - sample2.pdf
  - spam
    - sample3.txt

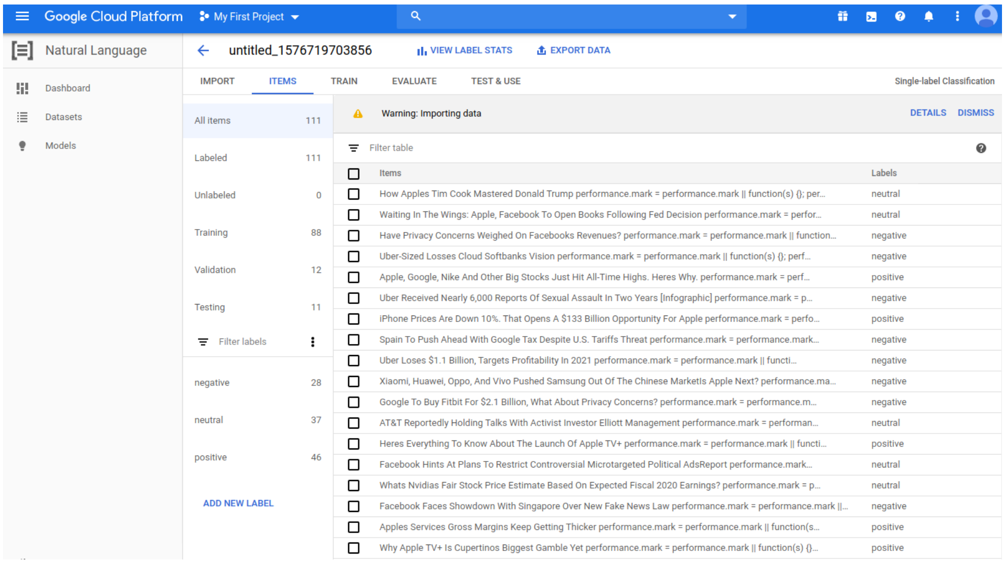
ZIP with multiple labels per text item

- documents.zip
  - email
    - sample1.txt
    - sample3.txt
  - not_spam
    - sample1.txt
    - sample2.pdf

- **AutoML Text & Document Classification**
  - **ONLY USE TEXT LABELS**
    - e.g.

```
+ documents.zip
|---- [text label 1]
|      + sample1.txt
|      + sample2.txt
|---- [text label 2]
|      + sample3.txt
```

  - Make sure each sample document has unique name for only **ONE** of the labels, since we are only doing single label classification, a one-to-one relation classification
  - It will take a moment to upload your documents and import them to the service.

        - 

Google Cloud Platform · My First Project

Natural Language · untitled_1576719703856 · VIEW LABEL STATS · EXPORT DATA

Dashboard · Datasets · Models

IMPORT · ITEMS · TRAIN · EVALUATE · TEST & USE · Single-label Classification

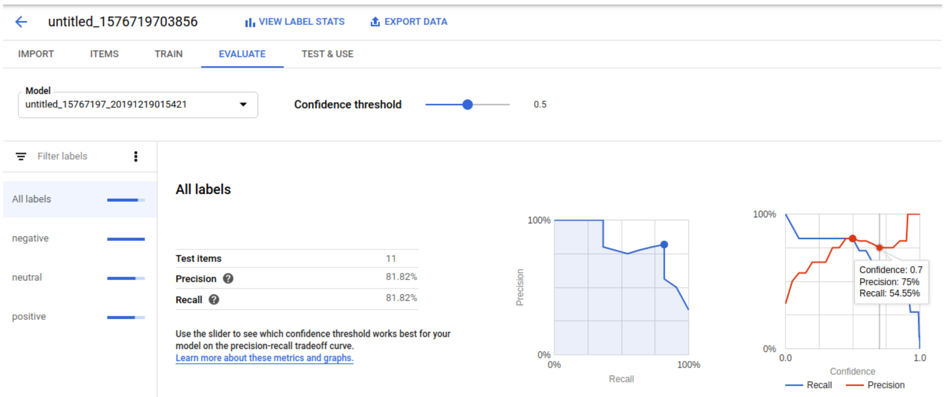
All items	111
Labeled	111
Unlabeled	0
Training	88
Validation	12
Testing	11
negative	28
neutral	37
positive	46

Filter labels · ADD NEW LABEL

Warning: Importing data · DETAILS · DISMISS

Filter table

Items	Labels
How Apples Tim Cook Mastered Donald Trump performance.mark = performance.mark || function(s) {}; per...	neutral
Waiting In The Wings: Apple, Facebook To Open Books Following Fed Decision performance.mark = perfor...	neutral
Have Privacy Concerns Weighed On Facebooks Revenues? performance.mark = performance.mark || function...	negative
Uber-Sized Losses Cloud Softbanks Vision performance.mark = performance.mark || function(s) {}; perf...	negative
Apple, Google, Nike And Other Big Stocks Just Hit All-Time Highs. Heres Why. performance.mark = perf...	positive
Uber Received Nearly 6,000 Reports Of Sexual Assault In Two Years [Infographic] performance.mark = p...	negative
iPhone Prices Are Down 10%. That Opens A $133 Billion Opportunity For Apple performance.mark = perfo...	positive
Spain To Push Ahead With Google Tax Despite U.S. Tariffs Threat performance.mark = performance.mark...	negative
Uber Loses $1.1 Billion, Targets Profitability In 2021 performance.mark = performance.mark || functi...	negative
Xiaomi, Huawei, Oppo, And Vivo Pushed Samsung Out Of The Chinese MarketIs Apple Next? performance.ma...	negative
Google To Buy Fitbit For $2.1 Billion, What About Privacy Concerns? performance.mark = performance.m...	negative
AT&T Reportedly Holding Talks With Activist Investor Elliott Management performance.mark = performan...	neutral
Heres Everything To Know About The Launch Of Apple TV+ performance.mark = performance.mark || functi...	positive
Facebook Hints At Plans To Restrict Controversial Microtargeted Political AdsReport performance.mark...	neutral
Whats Nvidias Fair Stock Price Estimate Based On Expected Fiscal 2020 Earnings? performance.mark = p...	neutral
Facebook Faces Showdown With Singapore Over New Fake News Law performance.mark = performance.mark ||...	negative
Apples Services Gross Margins Keep Getting Thicker performance.mark = performance.mark || function(s...	positive
Why Apple TV+ Is Cupertinos Biggest Gamble Yet performance.mark = performance.mark || function(s) {}...	positive

    - Click on `TRAIN` once imported files to the Google Natural Language service.
        - It takes several hours to train a model
        - Once a model is trained, a overall evaluation of the model will be available.
            - By default, documents are divided into 80% training and 20% testing by default.
            - 

untitled_1576719703856 · VIEW LABEL STATS · EXPORT DATA

IMPORT · ITEMS · TRAIN · EVALUATE · TEST & USE

Model: untitled_15767197_20191219015421

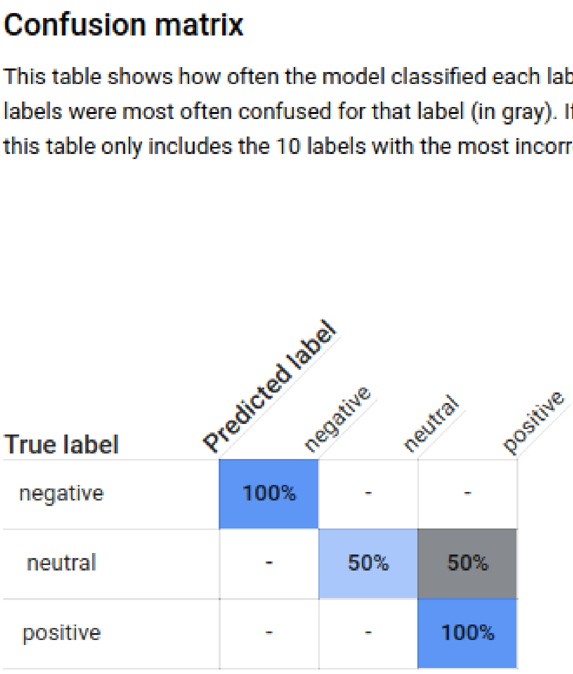
Confidence threshold 0.5

Filter labels: All labels · negative · neutral · positive

All labels

Test items	11
Precision	81.82%
Recall	81.82%

Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.
Learn more about these metrics and graphs.

            - 

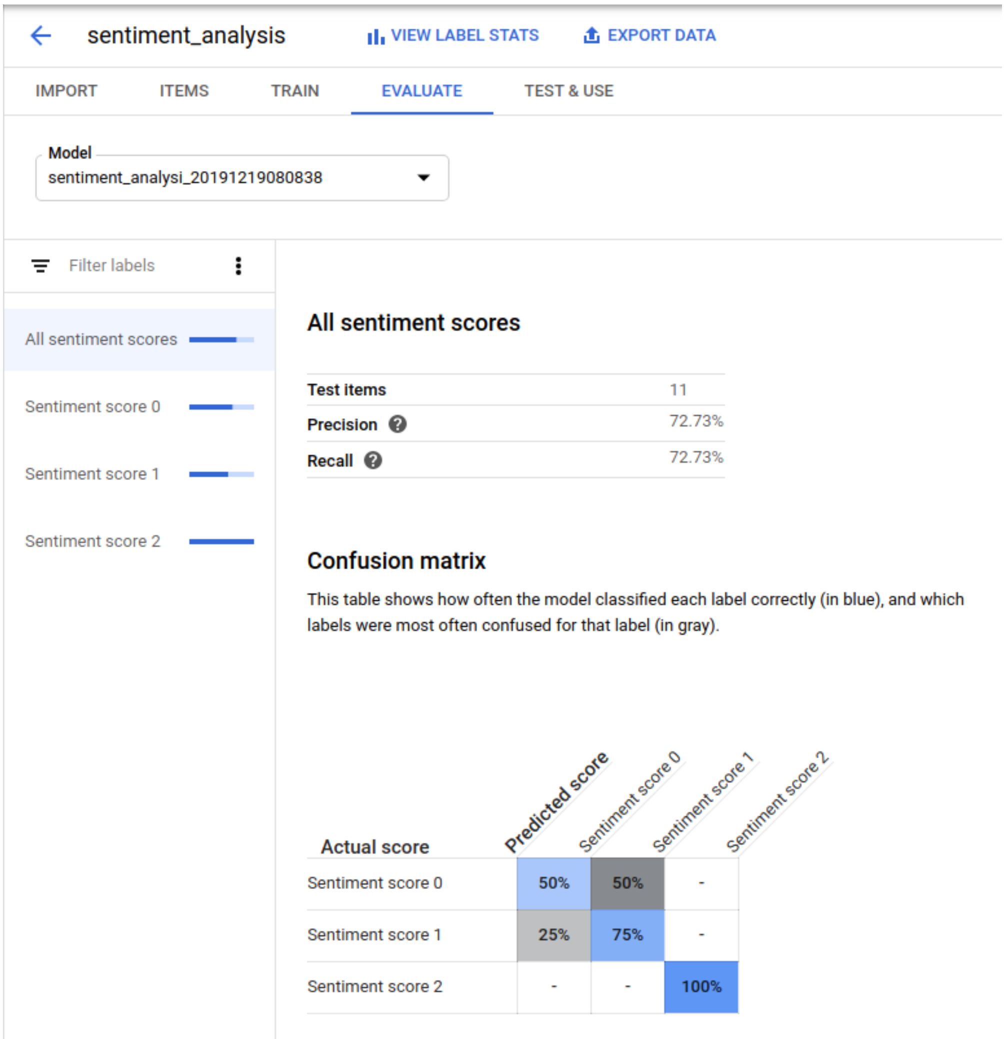
Confusion matrix

This table shows how often the model classified each label correctly (in blue), and which labels were most often confused for that label (in gray). If you have more than 10 labels, this table only includes the 10 labels with the most incorrect predictions.

True label \ Predicted label	negative	neutral	positive
negative	100%	-	-
neutral	-	50%	50%
positive	-	-	100%

- **AutoML Sentiment Analysis**:
    - **ONLY USE INTEGER LABELS**
        - e.g.

```
+ documents.zip
|---- [integer label 1]
|     + sample1.txt
|     + sample2.txt
|---- [integer label 2]
|     + sample3.txt
```

- Make sure each sample document has unique name for only **ONE** of the labels, since we are only doing single label classification, a one-to-one relation classification.
- Similar to the steps in the **AutoML text & document classification**:
    - 

Model
sentiment_analysi_20191219080838

All sentiment scores

Test items	11
Precision	72.73%
Recall	72.73%

Confusion matrix

This table shows how often the model classified each label correctly (in blue), and which labels were most often confused for that label (in gray).

Actual score	Predicted score		
	Sentiment score 0	Sentiment score 1	Sentiment score 2
Sentiment score 0	50%	50%	-
Sentiment score 1	25%	75%	-
Sentiment score 2	-	-	100%