Scraping SeekingAlpha :

a crowd-sourced site for insights on stocks and more

Fangye Shi
8/9/2018

Motivation:

- Gather opinions on stocks (long or short) and the reasonings behind the opinions.
- With enough data, hope to train a model to do sentiment analysis.

Difficulty

- **The website deploys heavy anti-scraping measures.**

To overcome this, I used a pool of user agents and a pool of proxies.

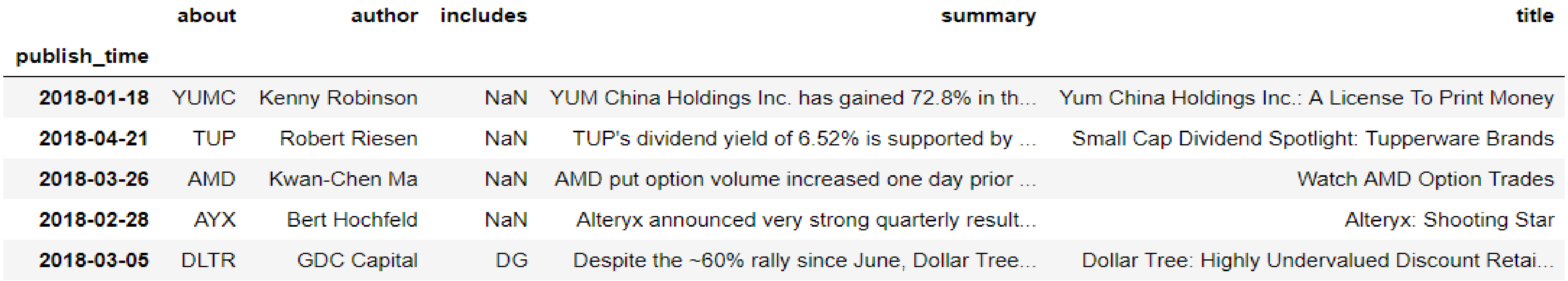
In the end, I was able to scrape around 8500 items on long ideas and 10300 items on short ideas.

- **Not enough data/time to train a model from scratch.**

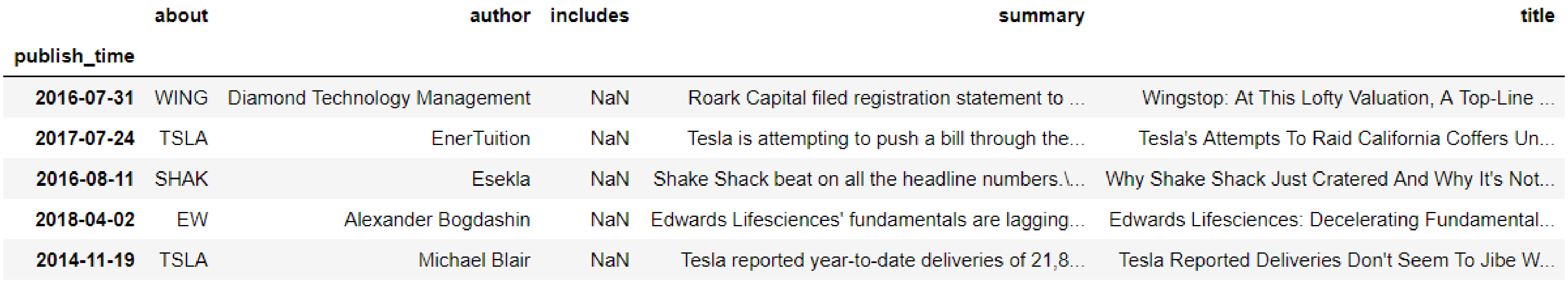
To overcome this, I did a transfer learning based on a pretrained model on text

embedding.

Dataset



	about	author	includes	summary	title
publish_time					
2018-01-18	YUMC	Kenny Robinson	NaN	YUM China Holdings Inc. has gained 72.8% in th...	Yum China Holdings Inc.: A License To Print Money
2018-04-21	TUP	Robert Riesen	NaN	TUP's dividend yield of 6.52% is supported by ...	Small Cap Dividend Spotlight: Tupperware Brands
2018-03-26	AMD	Kwan-Chen Ma	NaN	AMD put option volume increased one day prior ...	Watch AMD Option Trades
2018-02-28	AYX	Bert Hochfeld	NaN	Alteryx announced very strong quarterly result...	Alteryx: Shooting Star
2018-03-05	DLTR	GDC Capital	DG	Despite the ~60% rally since June, Dollar Tree...	Dollar Tree: Highly Undervalued Discount Retai...



	about	author	includes	summary	title
publish_time					
2016-07-31	WING	Diamond Technology Management	NaN	Roark Capital filed registration statement to ...	Wingstop: At This Lofty Valuation, A Top-Line ...
2017-07-24	TSLA	EnerTuition	NaN	Tesla is attempting to push a bill through the...	Tesla's Attempts To Raid California Coffers Un...
2016-08-11	SHAK	Eseklа	NaN	Shake Shack beat on all the headline numbers.\...	Why Shake Shack Just Cratered And Why It's Not...
2018-04-02	EW	Alexander Bogdashin	NaN	Edwards Lifesciences' fundamentals are lagging...	Edwards Lifesciences: Decelerating Fundamental...
2014-11-19	TSLA	Michael Blair	NaN	Tesla reported year-to-date deliveries of 21,8...	Tesla Reported Deliveries Don't Seem To Jibe W...

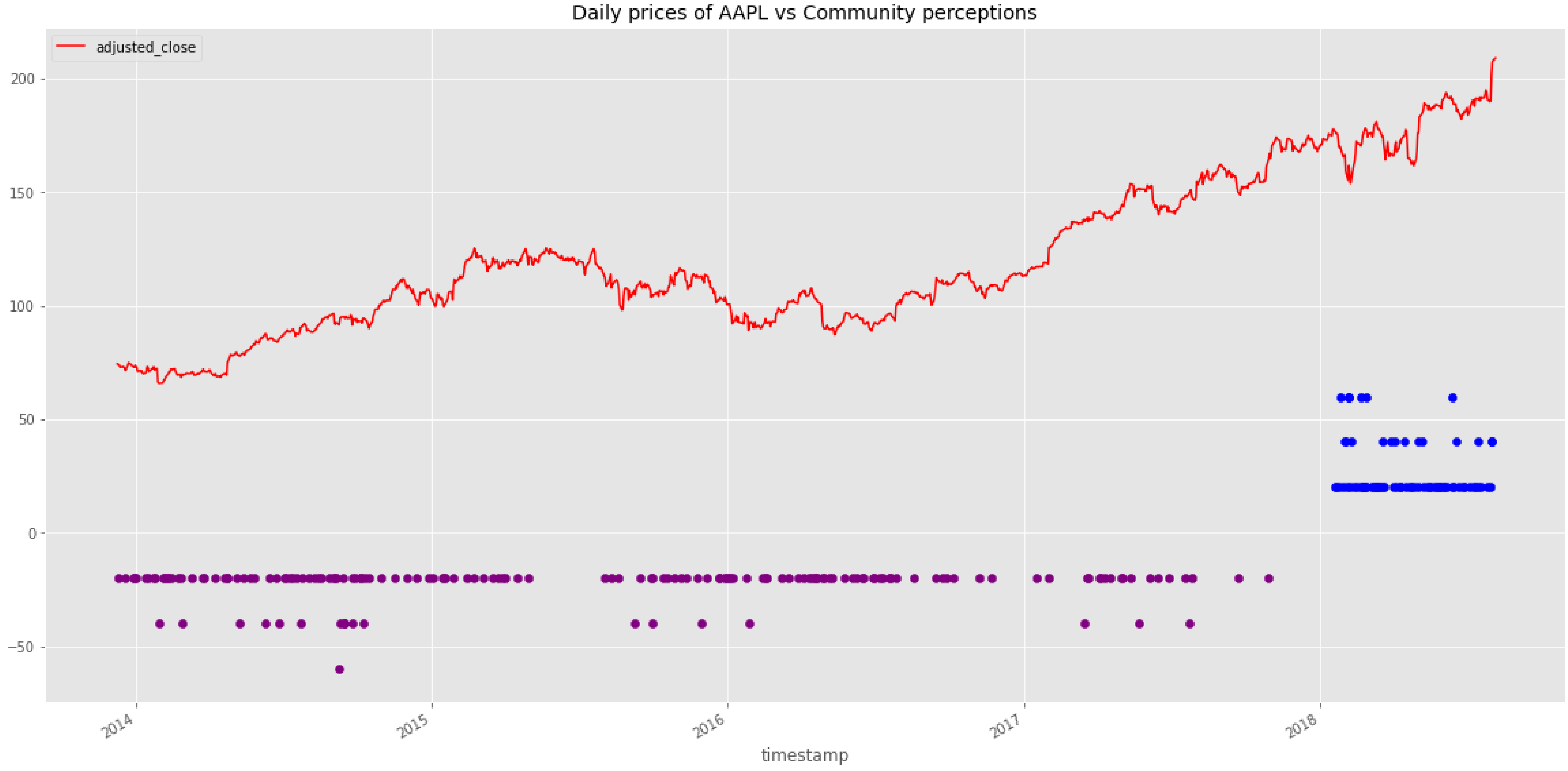
Trivia

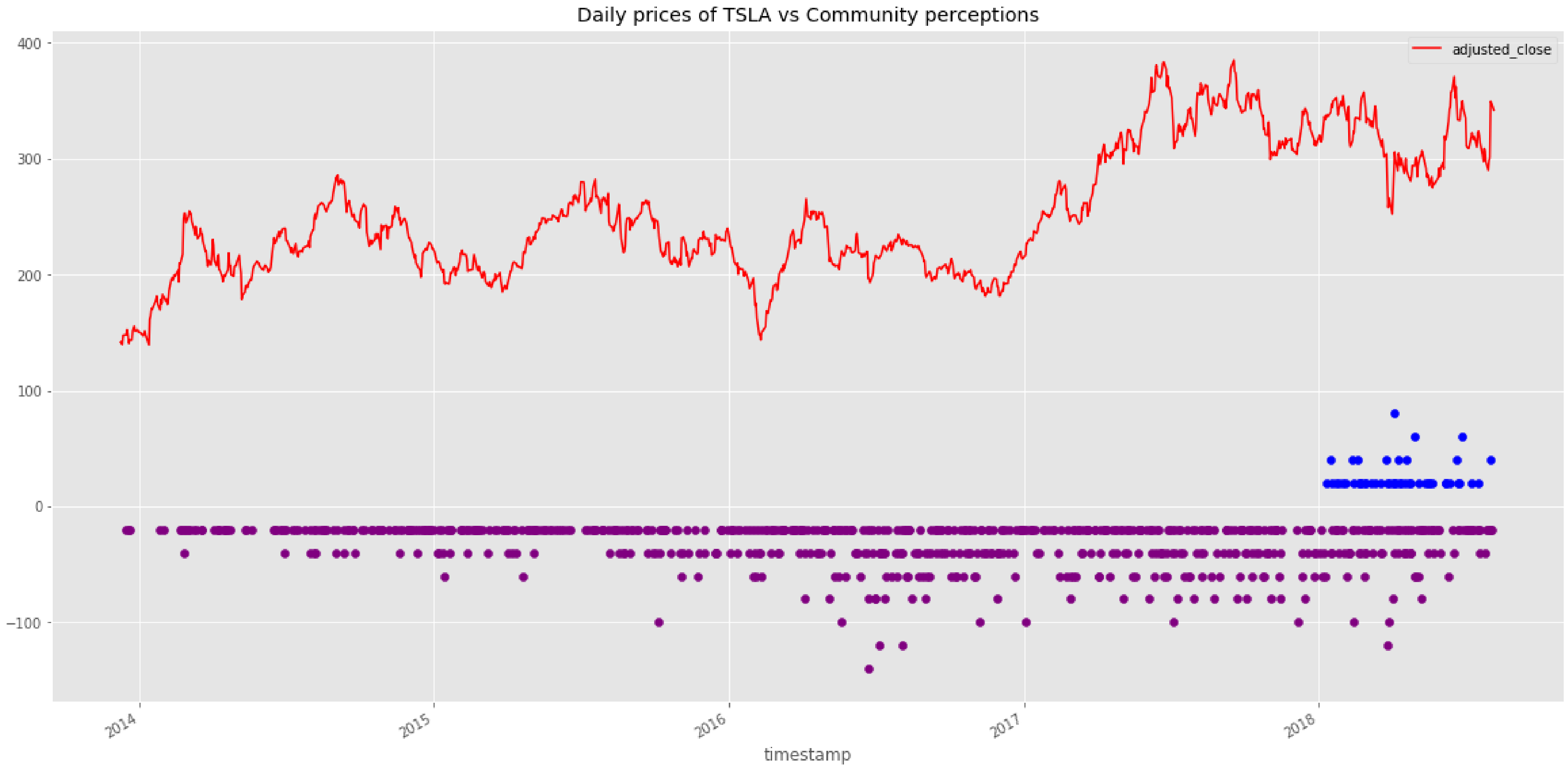
- Over 7000 entries on long ideas are from this year! In contrast, only 700 entries on short ideas are from this year.
- Top 10 most frequently mentioned stocks to long:

AAPL (130),FB (109), TSLA (92), GE (87), MU (82), AMD (78), AMZN (53),BABA (51), MSFT (43) NVDA (40)

- Top 10 most frequently mentioned stocks to short:

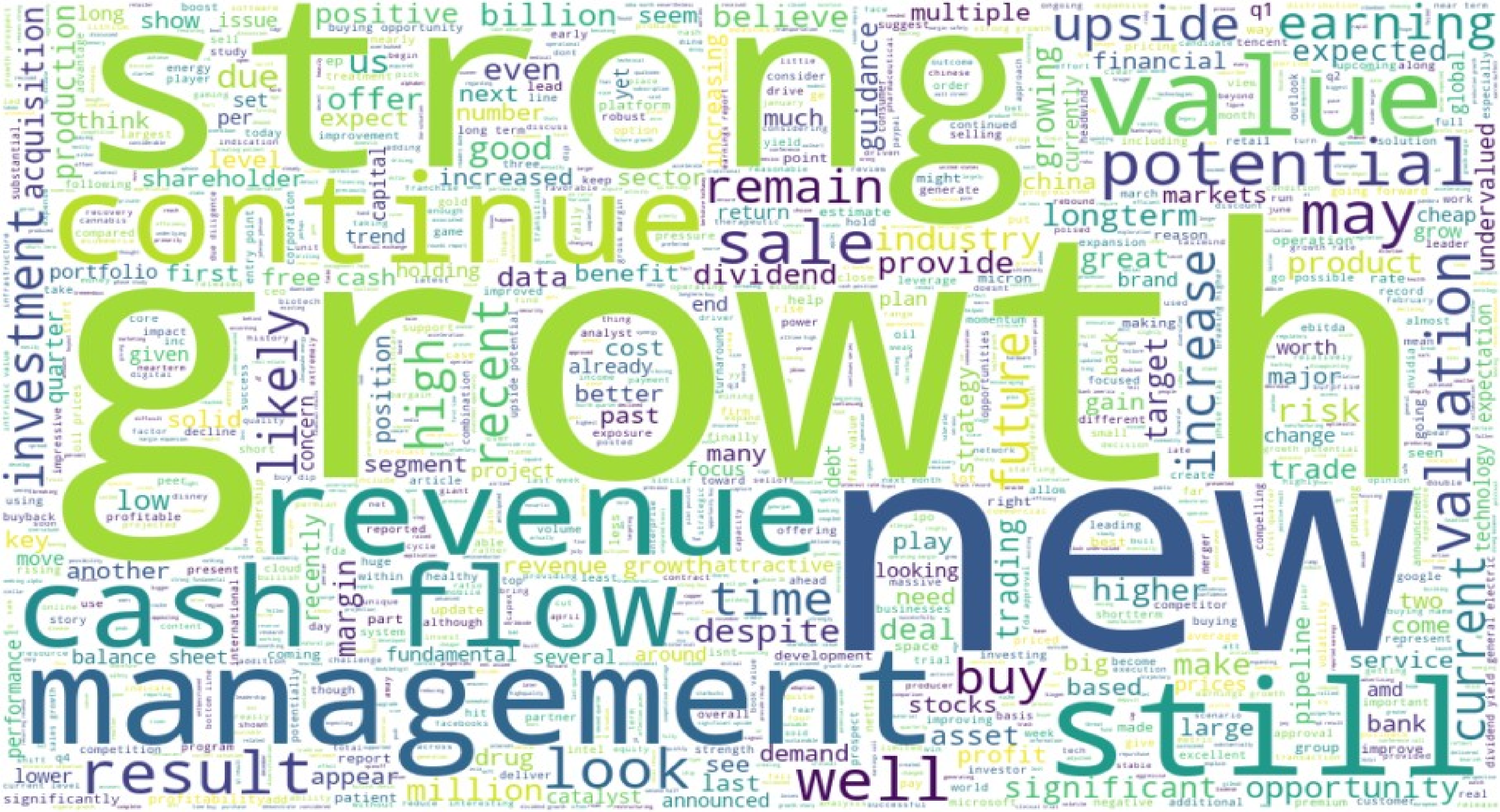
TSLA (1085), HLF (336), AAPL (182), AMZN (174), NFLX (172), BB (129), TWTR (97), BHC (77), AMD (76), DRYS (74)


Daily prices of AAPL vs Community perceptions
adjusted_close
200
150
100
50
0
−50
2014
2015
2016
2017
2018
timestamp

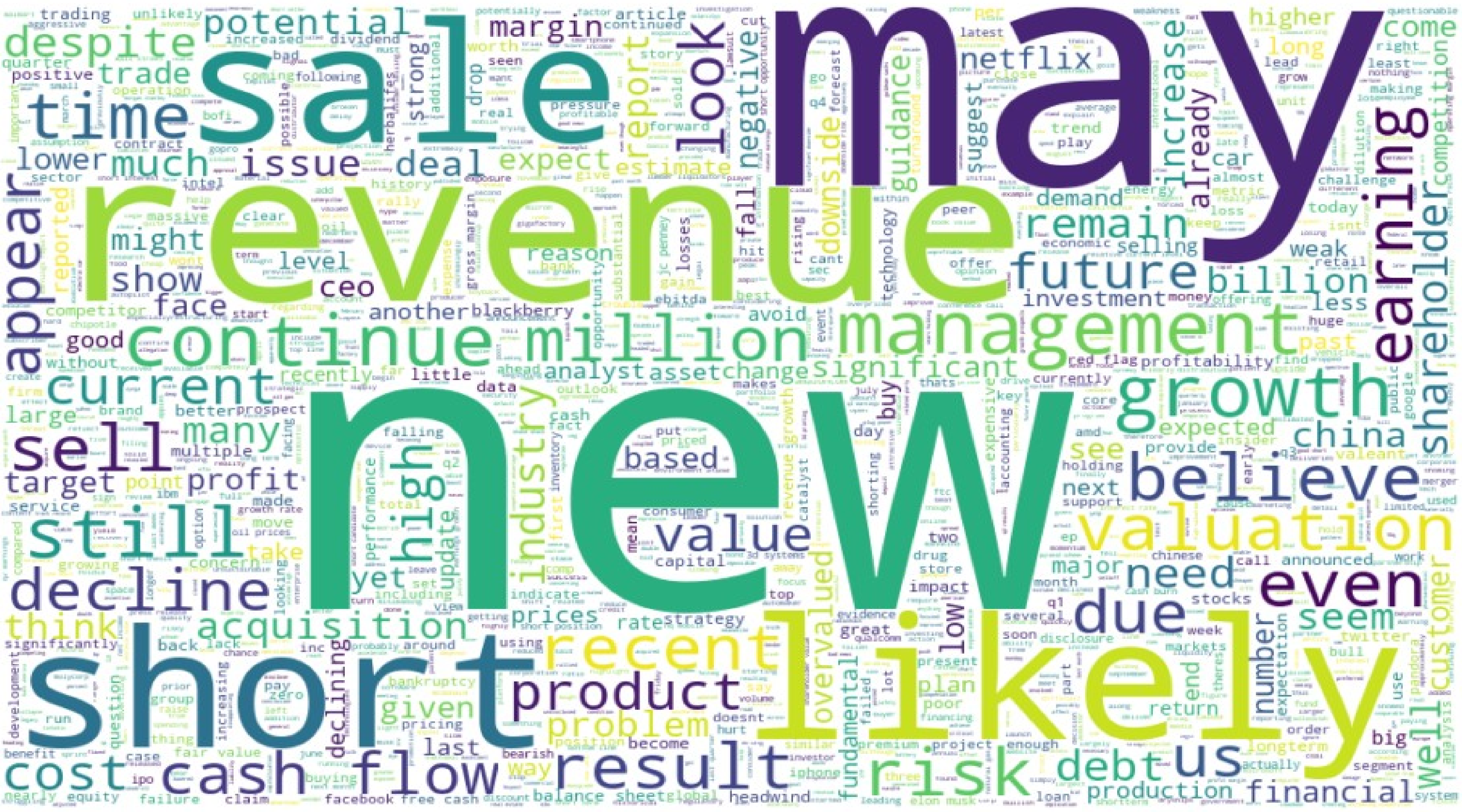


Daily prices of TSLA vs Community perceptions

Word Cloud for long ideas



Word Cloud for short ideas

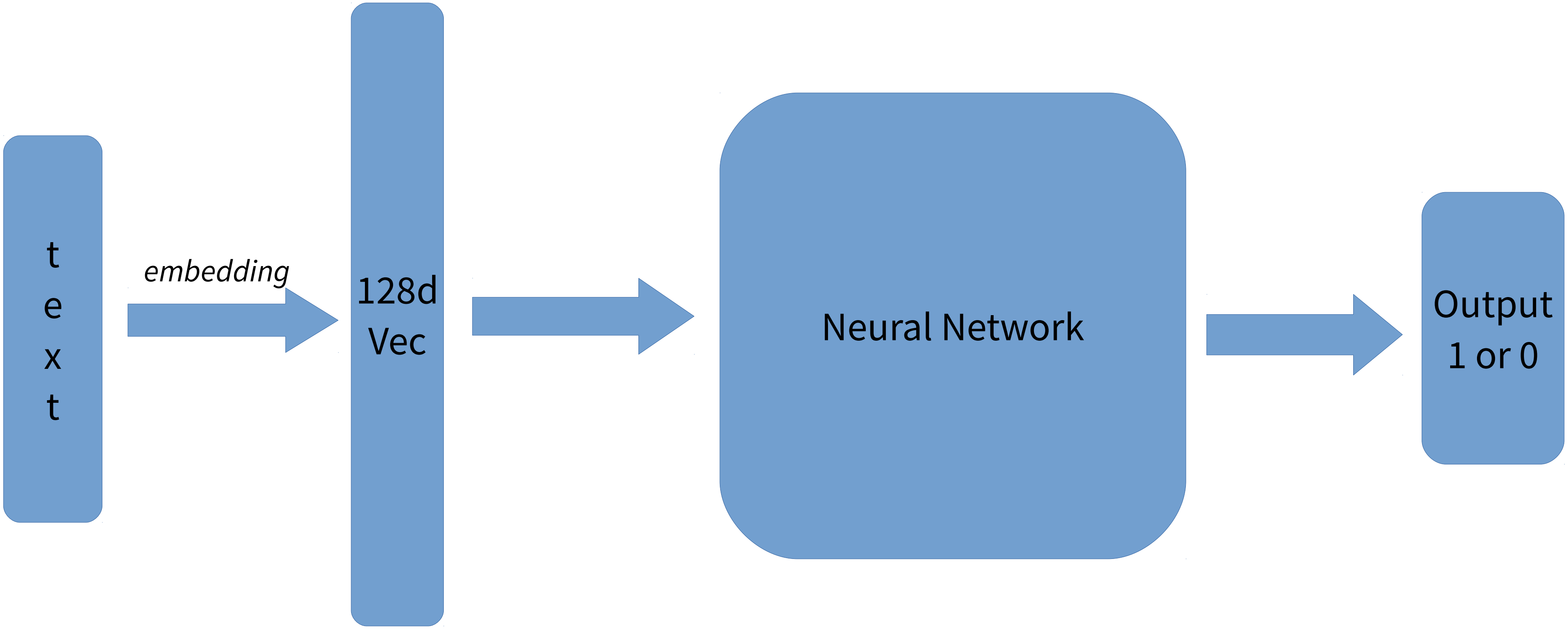


Transfer learning

- Based on a pretrained sentence embedding module

Map each word to a vector in 128-dimensional space and map each sentence to sum(word_vector)/sqrt(length)

- After this, train a DNN classifier with two hidden layers.



Accuracies

- **Comparison of 4 different models.**

First one use the approach described in the last slide. Note that we do not change the parameters in word embedding.

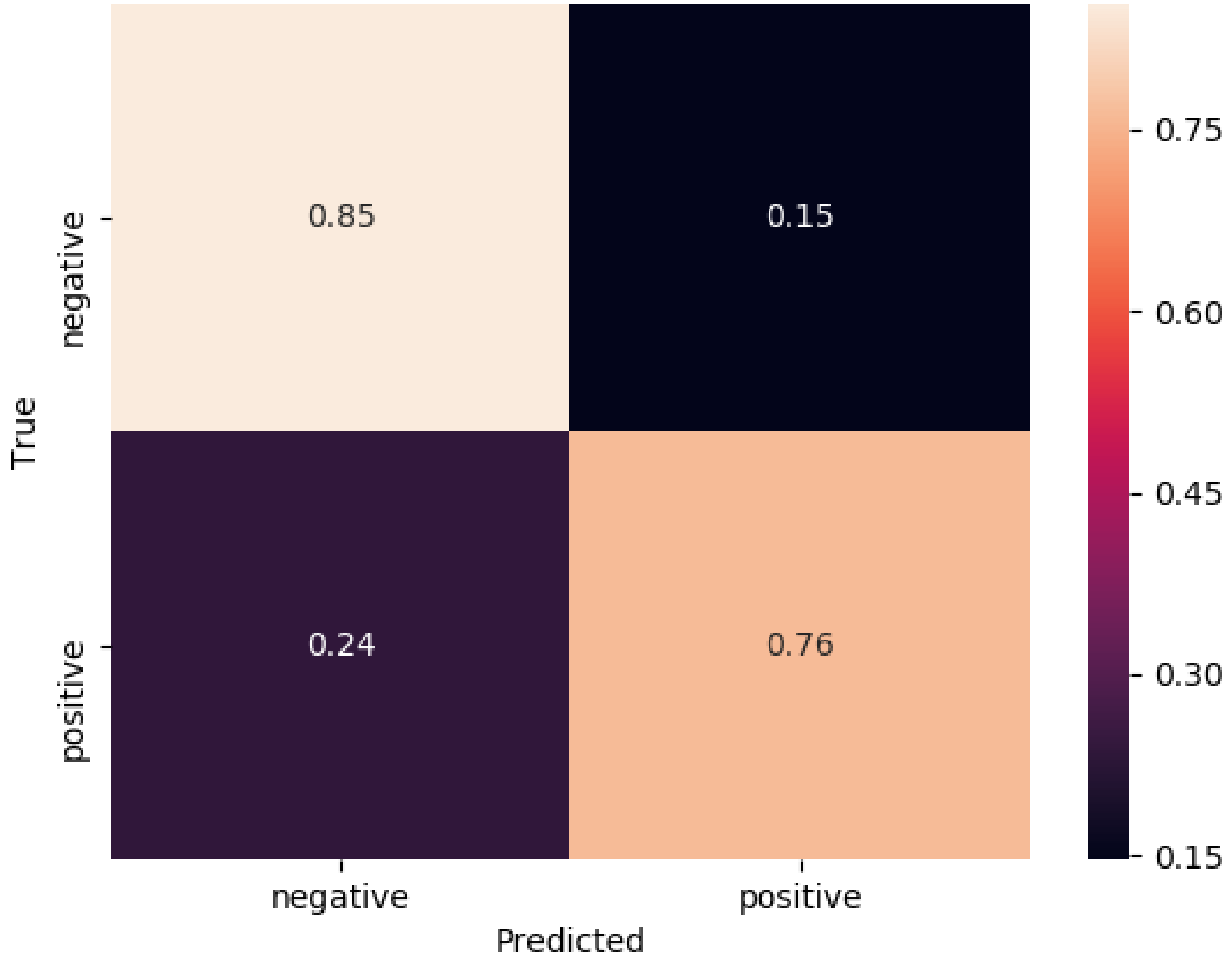
Second one is similar to the first one but this time we also allow our model to tune the parameters in word embedding.

For the last two, we use random parameters in word embedding.

- **Note how quickly the second model is overfitting the training set.**

	Train Accuracy	Test Accuracy
nnlm-en-dim128	0.81	0.80
nnlm-en-dim128-with-module-training	0.98	0.89
random-nnlm-en-dim128	0.80	0.72
random-nnlm-en-dim128-with-module-training	0.84	0.76

Confusion matrix


True
negative
positive
Predicted
negative
positive