REUTERS/Toru Hanai

Mining insights from text

Improving Predictability of Oil via Reuters News Text

Sameena Shah[1], Armineh Nourbakhsh[2]

[1]Director, Research and Head of R&D NY Labs, Thomson Reuters

[2]Research Engineer, R&D, Thomson Reuters

THOMSON REUTERS

Do words predict market movement ?

- What companies say about themselves (SEC filings)
- What journalists say about companies (news)
- What the crowd says about companies (social media)
- How do words impact commodities markets ?

Does News Impact Oil ?

- Traditionally, models focused on responses to quantitative or ‘hard’ measures like demand supply numbers.
- How about a certain pipeline being attacked ? How do we quantify “soft” measures and their impact ?
- Our approach: Measure Public Reaction – as reflected in price change or volatility
- If yes, then can we turn the affirmation into a predictive model

What is it about news that impacts oil prices/volatility

- Time – inventory numbers released periodically
- Events – captured by news
- Text and Narrative – what does it really mean ?
- Previous News and Market state

How do we capture what a story conveys ?

- Sentiment
  - Map certain words to “negative”, “fear”, “optimism”, “positive” (Mood words)
  - TRNA

- Bag of words
  - Instead of coming up with a mapping, work at the word level
  - Easy to implement and often good enough solutions.

Previous Work – Phase 1 (Sisk & Killian 2013)

- Represent a story as a vector where each component corresponds with a word in a filtered vocabulary. If a word is in a story, that component is 1, otherwise 0.
- Example: Vocabulary = [bag, of, words, oil]

  story = “This oil story contains words.”

  vector = [0, 0, 1, 1]

Lessons

- It didn’t work. No better than baseline.
- Regression is bad in high dimensions.
- Not all stories are meaningful.
- Solution: Filter stories better and find a compact representation of text – decrease its dimensionality in a “meaningful” way.

Dictionary terms (courtesy Jonathan Leff)

fire
explosion
leak
shut
halt
unplanned
maintenance
outage
strike
attack
bombs

tsunami
delayed
postponed
flare
start up
deferred
disruption
threat
sanctions
hurricane
on schedule

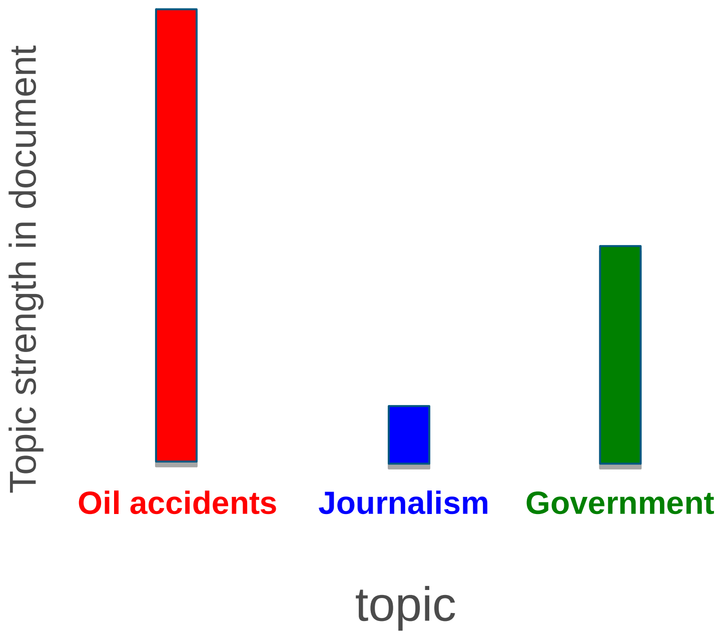
shutting-in
saudi production
libya
nsea
gulf
cushing
whiting
eubridge
seaway
keystone

Our Approach to Reduce Dimensionality: Latent Dirichlet Allocation (LDA)

- Input: large collection of text
- Automatically finds “topics”
- Output: mixture of topics
- Meaningful dimensionality reduction, easier for regression to grab a hold of.

Our Approach to Reduce Dimensionality: Topic Modeling

BP said its containment cap system at the site of a Gulf of Mexico oil leak captured about 7,920 barrels (332,640 U.S. gallons/1.26 million liters) of oil in the first 12 hours of Wednesday. If that rate continues, BP could capture nearly 15,900 barrels (667,800 gallons/2.53 million liters) for the 24-hour period -- the highest per-day amount since the system was installed last week. The total amount collected since June 4 reached 64,444 barrels (2.7 million gallons/10.25 million liters) with Wednesday's half-day tally, according to BP figures. The top U.S. official overseeing the operation said earlier on Wednesday that as the capture rate ramps up, BP is working to nearly double the capacity to handle it at the surface. U.S. Coast Guard Admiral Thad Allen said at a news conference in Washington that BP is working to increase processing capacity at a drillship and a service rig at the water's surface to 28,000 barrels (1.18 million gallons/4.45 million liters) a day to handle the load as the company ramps up the collection rate from the seven-week-old leak.



From tens of words to 3 action labels

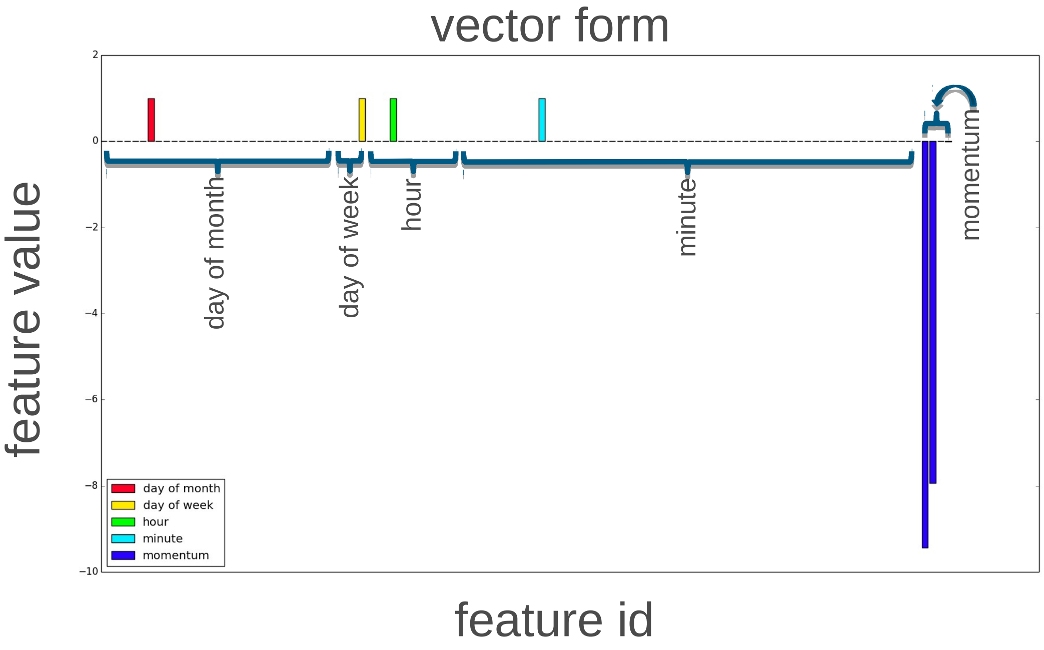
Modeling

Experimental Setup

- Predict 90 minutes in future
- Calendar effects
  - Minute, hour, day of week, day of month
- Momentum
  - Cumulative return for previous 5 and 60 minutes
  - Log of volatility for previous 5 and 60 minutes
- Baseline: calendar effects and momentum
- Our model: calendar effects, momentum, news features

Baseline Features

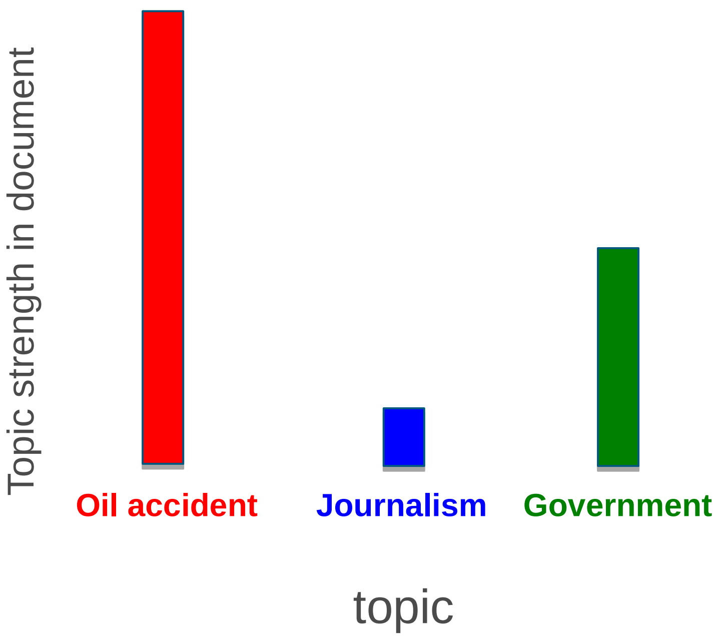
11:11 AM, 7th of January, 2011 (Friday)



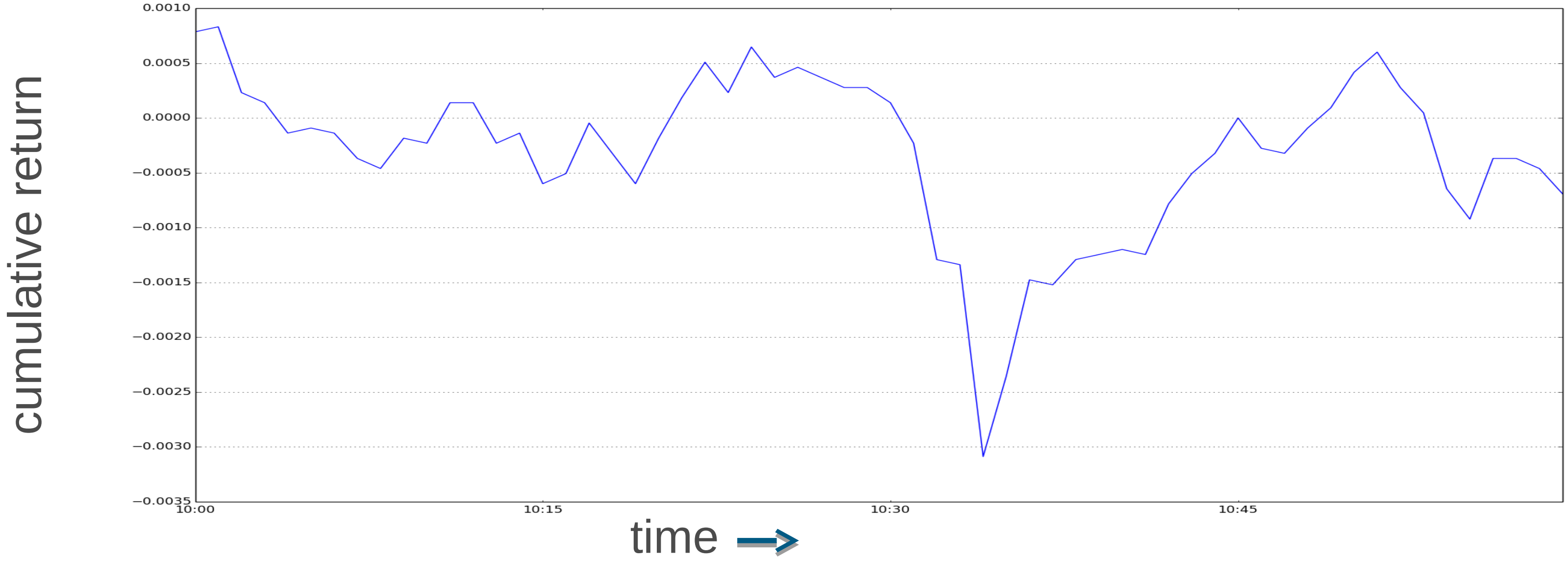
Topic features

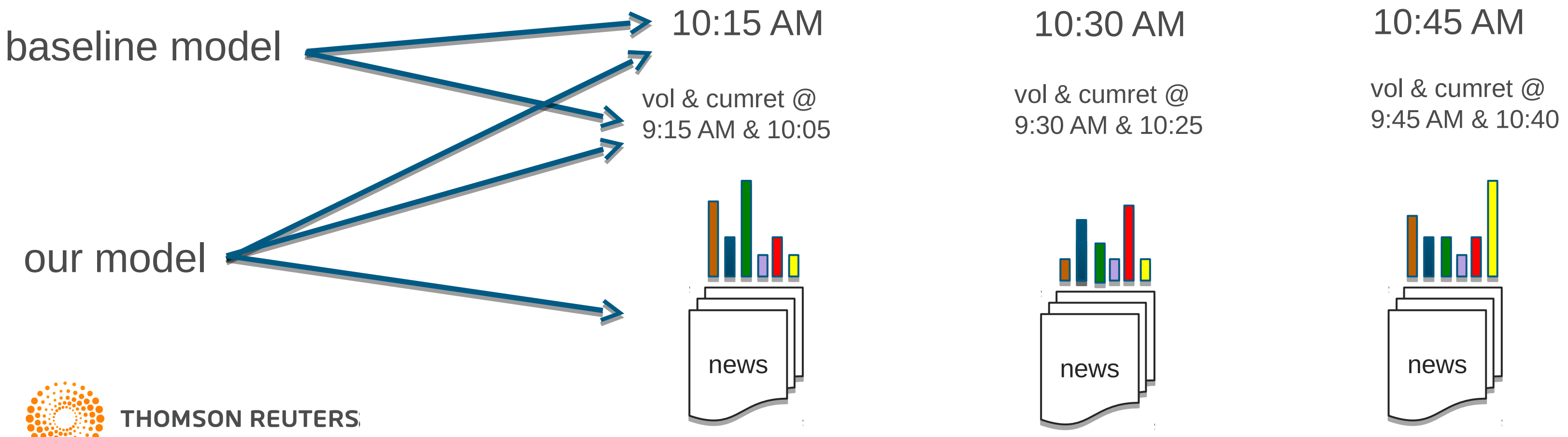
- Trained LDA model outputs topic distributions

BP said its containment cap system at the site of a Gulf of Mexico oil leak captured about 7,920 barrels (332,640 U.S. gallons/1.26 million liters) of oil in the first 12 hours of Wednesday. If that rate continues, BP could capture nearly 15,900 barrels (667,800 gallons/2.53 million liters) for the 24-hour period -- the highest per-day amount since the system was installed last week. The total amount collected since June 4 reached 64,444 barrels (2.7 million gallons/10.25 million liters) with Wednesday's half-day tally, according to BP figures. The top U.S. official overseeing the operation said earlier on Wednesday that as the capture rate ramps up, BP is working to nearly double the capacity to handle it at the surface. U.S. Coast Guard Admiral Thad Allen said at a news conference in Washington that BP is working to increase processing capacity at a drillship and a service rig at the water's surface to 28,000 barrels (1.18 million gallons/4.45 million liters) a day to handle the load as the company ramps up the collection rate from the seven-week-old leak.



Predicting 90 minutes in the future





baseline model

our model

10:15 AM

vol & cumret @
9:15 AM & 10:05

news

10:30 AM

vol & cumret @
9:30 AM & 10:25

news

10:45 AM

vol & cumret @
9:45 AM & 10:40

news

THOMSON REUTERS

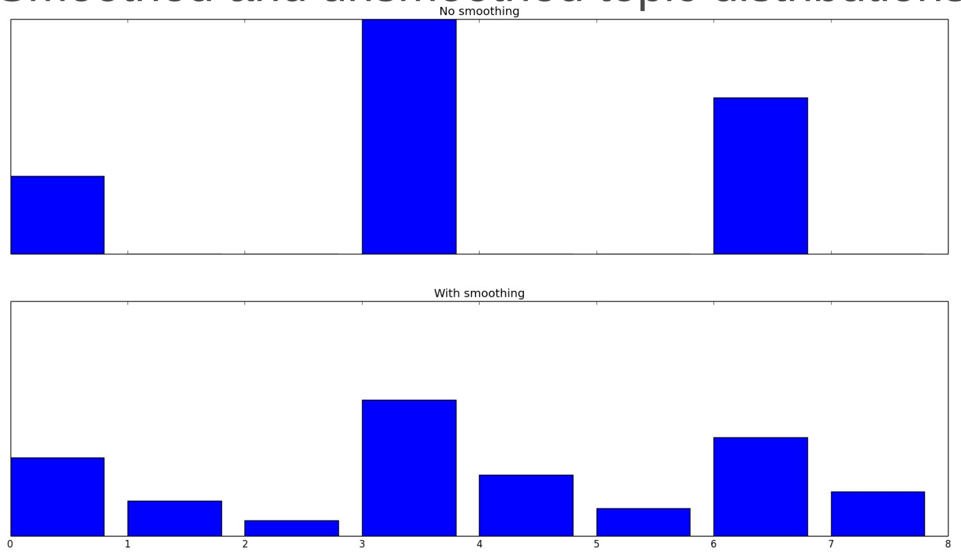
Example topics

Oil Accident	Middle EastPolitical Unrest	Oil Shipping
pipelin 0.09	polic 0.06	oil 0.26
line 0.05	wound 0.05	export 0.10
oper 0.04	bomb 0.04	fuel 0.10
#energy_sector# 0.03	kill 0.04	import 0.09
leak 0.02	sourc 0.03	port 0.04
spill 0.02	two 0.03	sourc 0.03
carri 0.02	car 0.02	termin 0.02
shut 0.01	peopl 0.02	tank 0.02
compani 0.01	km 0.02	farm 0.01
flow 0.01	secur 0.02	storag 0.01

Capturing long-term effects of news

- Smooth topic distributions into the future

Smoothed and unsmoothed topic distributions



Topic Mentions

No smoothing

With smoothing

Time ->

All stories are not made the same

Price updates	Dec. 6 Bonito +$9.20 HLS +$11.15 LLS +$10.50, +$10.45, +$10.90 Mars +$6.50, +$6.90 Jan-Feb box +40 cents Thunder Horse +$8.85 WTI at Midland -65 cents Dec. 5 Bonito +$9 HLS +$10.55 LLS +$10.50 Mars +$6.30 WTS -85 cents WTI at Midland -65 cents Dec. 2 Bonito +$9.20 HLS +11.45 LLS +10.75, +$10.45, +$10.50 Mars +$6.90, +$6.80, +$6.75 Poseidon -80 cents to Mars Dec, 1 Bonito +$9.40
Top headlines	Iran warns U.S. over Strait of Hormuz [nL6E7NT2WZ] > Oil falls below $107, US stocks and Iran in focus [nL6E7NT245] > Saudi Arabia to donate fuel to troubled Yemen [nL6E7NT258] > Thailand, Cambodia aim for oil development [nL3E7NT4TZ] > Saudi Arabia to cut February crude OSPs in Asia [nL3E7NS3R5] > Ghana latest in Africa to cut fuel subsidies [nL6E7NT2F0] > Kazakh Atyrau oil refinery inks upgrade
Non-English	歐洲部分 葡萄牙周一新發行了35億歐元5年期基準國債，互換利率中價為+360基點，相當于2016年2月份到 期的Bobl 159債券402.3基點。一大型銀行知情人士向MNI透露。再招標價格為99.762，票息利率為6.40%
Good story	BP said its containment cap system at the site of a Gulf of Mexico oil leak captured about 7,920 barrels (332,640 U.S. gallons/1.26 million liters) of oil in the first 12 hours of Wednesday. If that rate continues, BP could capture nearly 15,900 barrels (667,800 gallons/2.53 million liters) for the 24-hour period -- the highest per-day amount since the system was installed last week.

Solution: Apply story filtering

All text is not made the same

- Don’t associate authors & news sources with topics
  - Remove boilerplate text
- Improve topic quality
  - Stemming
  - dictionary check
  - stopword removal
- Topics should be generalizable
  - Company name replaced with sector
  - Remove names of people and locations

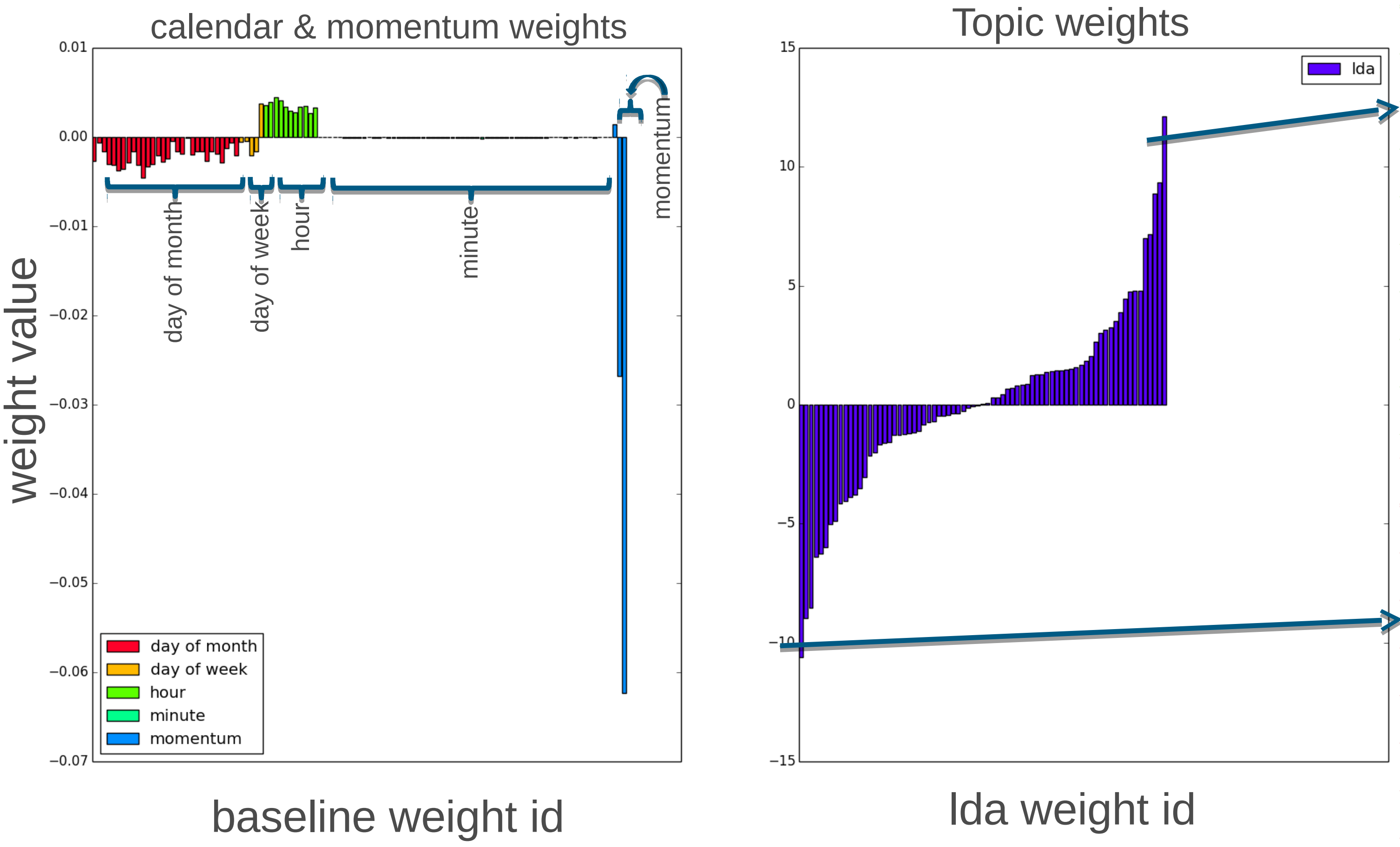
Prediction model

- Features for time t to predict target at t+90 mins
- Fit ordinary least squares
- $y \approx (\text{calendar effects})\cdot\vec{w}_1+(\text{momentum})\cdot\vec{w}_2+(\text{topic features})\cdot\vec{w}_3$

Not all instances are made the same

- Only accept those we’re most sure of.
- Big |prediction| ⇒ more certain
- To find threshold for what “big”, check predictions 90 minutes ago:
  - Reject an instance that should have been accepted, lower the threshold.
  - Accept an instance that should have been rejected, increase the threshold.

Weight Analysis



price
market
rise
high
higher
increas
level
rais
consum
cost

oil
crude
product
output
export
produc
oilfield
boost
major
heavi

Experiment: 2011 data

Model	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
News-based Model	60.73	52.9	47.83	54.94	51.4	60.04	57.31	59.51	51.23	51.84	55.1	47.34
Baseline	53.31	51.01	49.83	57.53	51.63	57.84	43.11	50.38	44.36	34.02	57.47	49.86

Average Accuracy (baseline): 50.01%
Average Accuracy (news-based model): **54.18%**

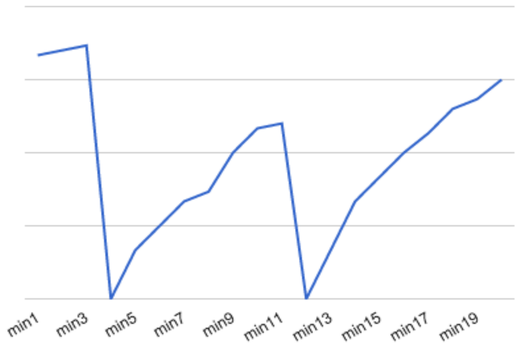
Updated Experiment: 2015 data

Average scores for predicting the direction of cumulative return		#1	#2
		No retraining, 2010-2011 model	Updated model
CLc1	News-based	37.5	**54.9**
	Baseline	31.4	46.8

Thought experiment: 2015 data

Challenges:

- Global oversupply in late 2014-early
- Sharp decline vs. slow recovery



Time →

	2011	2015
News-based model	54.18	54.9
Baseline	50.01	46.8

Conclusion

- We beat the baseline predictability!
- Topics discovered intuitively meaningful
- Lessons learnt along the way
  - Lot of preprocessing
  - How to better model text of news

Acknowledgment

- Andrew Nystrom
- Steve Pomerville
- Armineh Nourbakhsh
- Isabelle Moulinier
- Jacob Sisk

Prediction quality

- Cumulative Return
  - Baseline: 51.1% directional accuracy
  - Our model: 54% directional accuracy
- Volatility
  - Baseline: 0.549 Adjusted R^2
  - Our model: 0.553 Adjusted R^2

Are there different topics before different target variable movements?

- Learn 3 topic models: low, medium, high response
- Idea: force LDA to find better topics
- Topic features are distribution output of all 3
- Yielded similar performance as regular LDA
  - Increases dimensionality
  - Similar topics can be learnt with regular LDA, so let the model decide how to weight them