REUTERS/Toru Hanai

Improving Predictability of Oil via Reuters News Text

Andrew Nystrom, R&D
Sameena Shah, R&D
Jacob Sisk, TRGR
Isabelle Moulinier, R&D

THOMSON REUTERS

Lessons from previous work

- Vector space model 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.

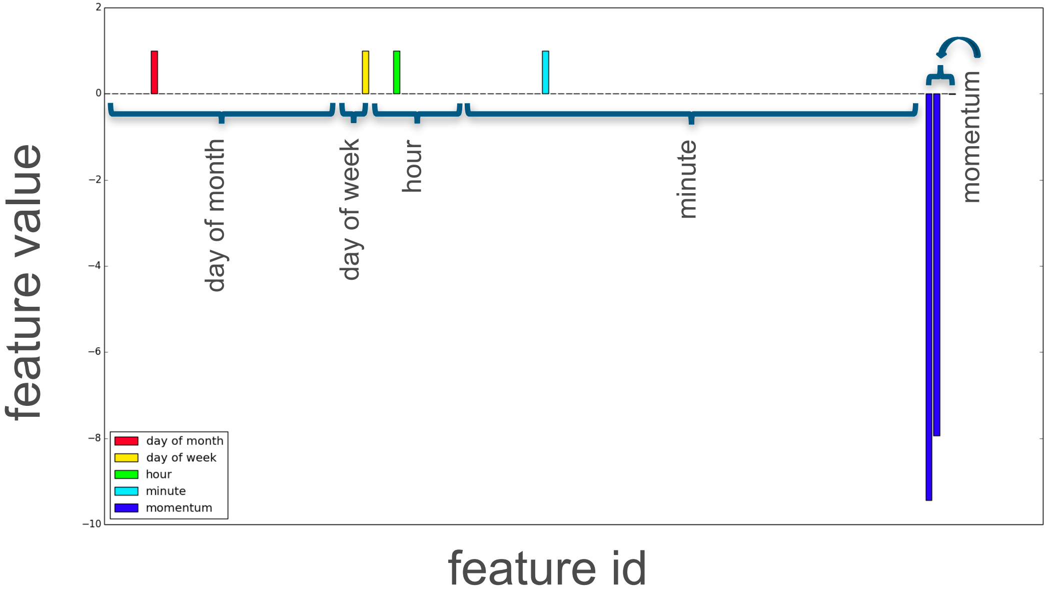
Experimental Setup

- 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)

vector form

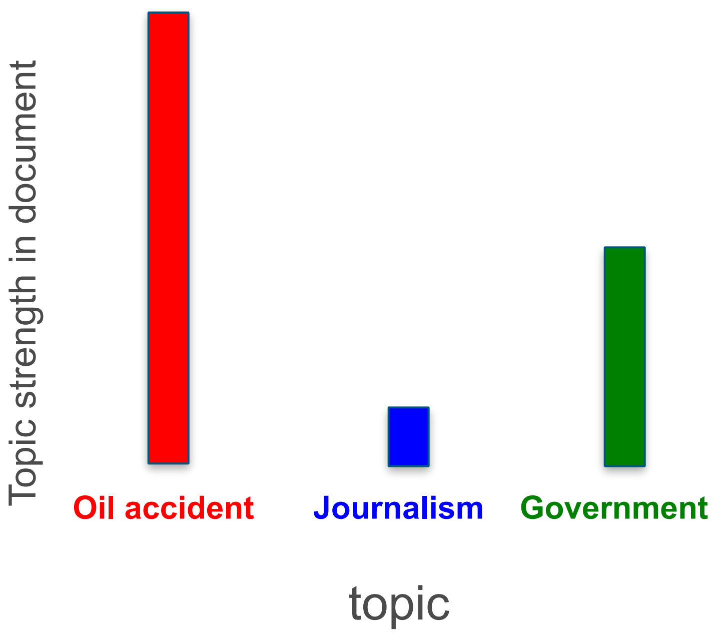


THOMSON REUTERS

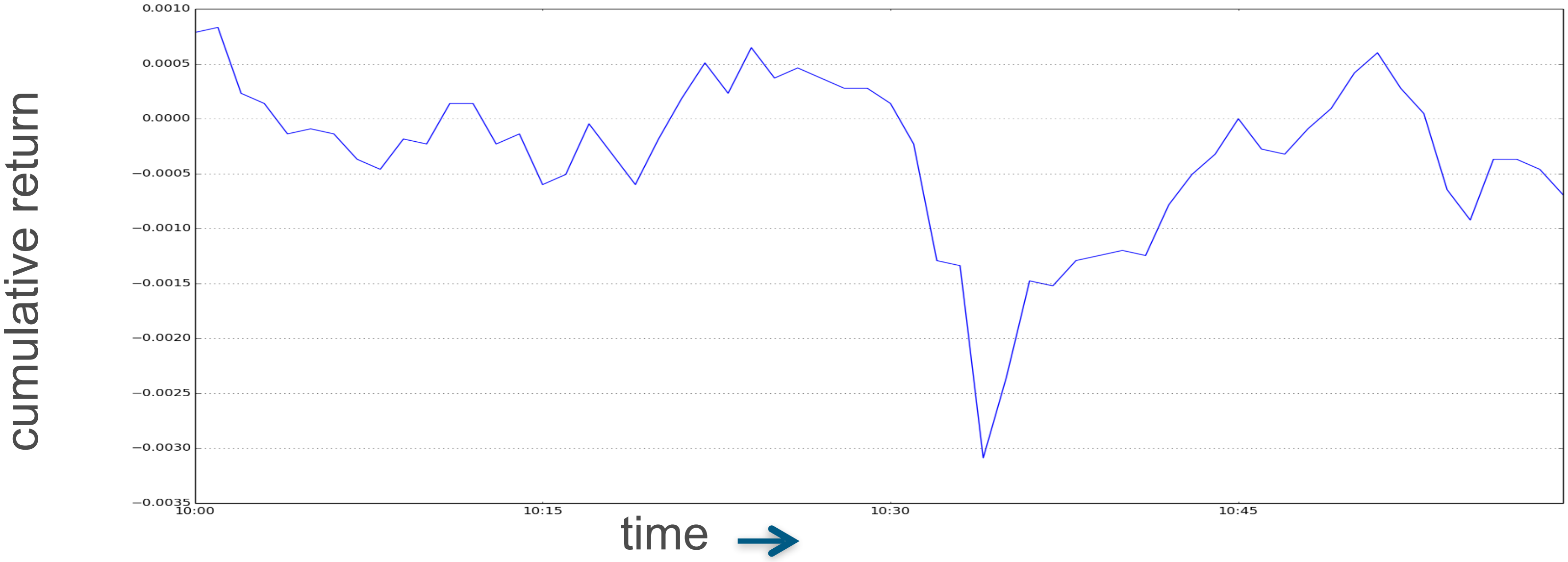
Topic features

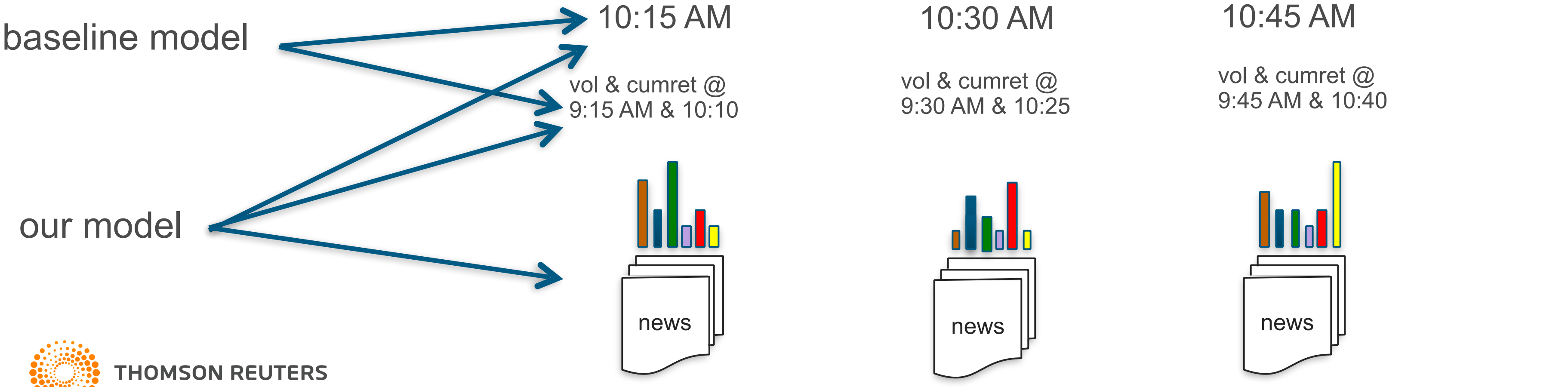
- Trained LDA model outputs topic distributions

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 & 10:10

10:30 AM

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

10:45 AM

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

THOMSON REUTERS

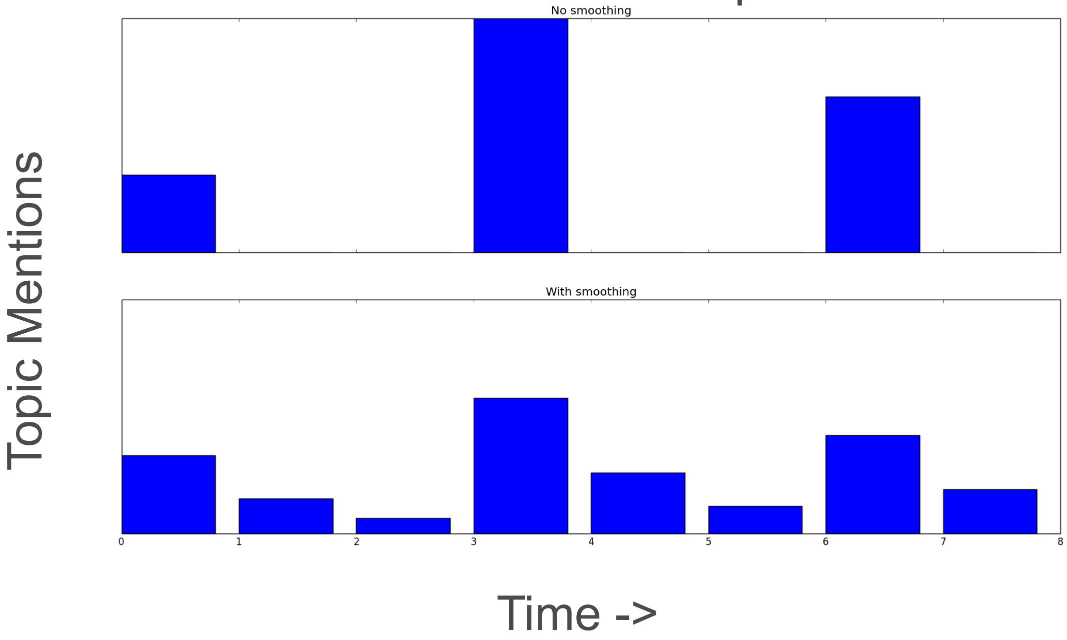
Example topics

Oil Accident	Civil 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



No smoothing

With smoothing

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

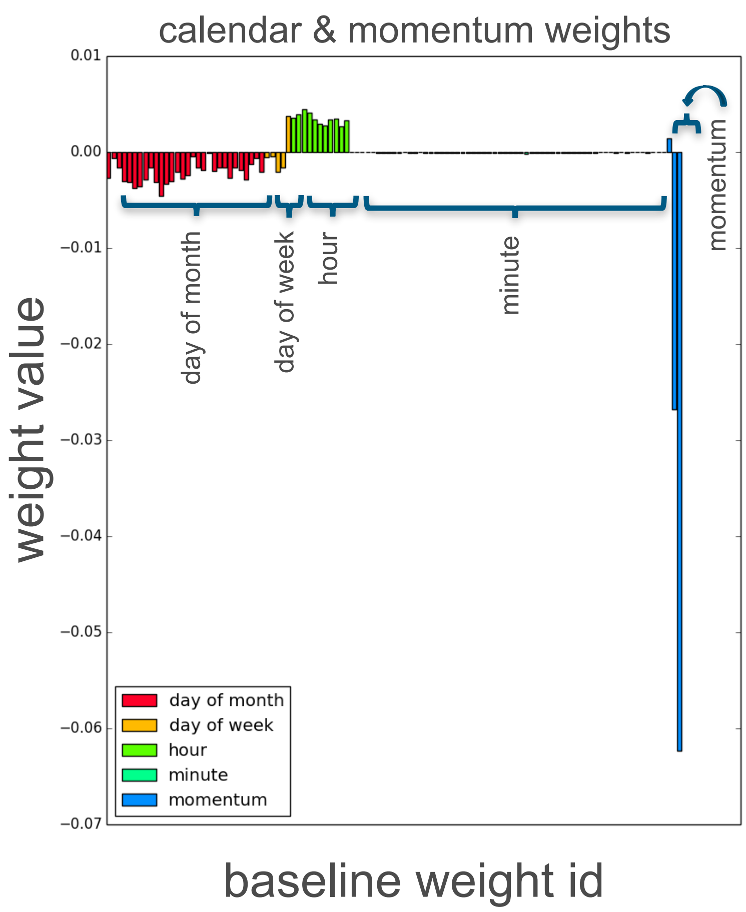
All text is not made the same

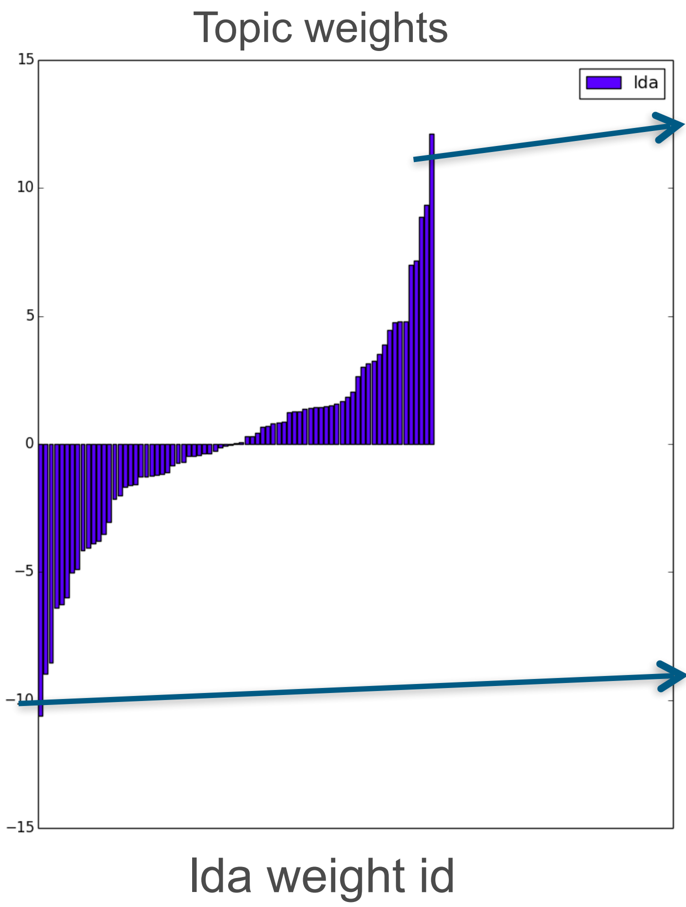
- Don’t associate authors & news sources with topics
  - Remove boilerplate text
- Improve topic quality
  - Stemming
  - dictionary check
  - stopword removal
- Topics should be generalizable
  - Company name replaced with sector
  - Remove names of people and locations

Not all instances are made the same

- Use regression to predict y
- Only accept guesses we're most sure of.
- Big |prediction| ⇒ more certain
- To find rejection threshold by looking 90 mins back:
  - 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

Summary

- Clean & filter text
- LDA
- Stem, entity translation and removal
- Regression to predict sign of cumulative return & volatility
- Learn acceptance threshold
- Success!