Using Natural Language Processing to Predict Returns and Risk in the Oil Market

Nida Cakir Melek, Charles W. Calomiris and Harry Mamaysky

September 17, 2019

Why Focus on Forecasting?

- The improve predictions and reduce uncertainty about the future.
- To understand what news is most important for the market. Is it about shifts in supply and demand in the traditional sense or about changes in risk, and if so, what risks?
- To understand the functioning of the market. Does the market incorporate news comprehensively and quickly? Are there time-varying risk premia in the market? Is news relevant for risk priced in the market?

Using NLP with Thomson Reuters

- NLP of TR maps the important news flow from a major news platform that is targeted at sophisticated market participants, oil market experts, and governmental agencies.
- We will see if a parsimonious mapping of this news flow, using techniques that have proven valuable elsewhere, provides incremental explanatory power beyond what one can achieve using a "kitchen sink" of other (traditional) forecasting variables.
- We will examine forecasting results for oil returns (using both spot and futures definitions of prices to compute returns), volatility (realized from Bloomberg), oil companies' stock returns (BP, Exxon, Shell), oil inventories, and oil production.
- By looking at these variables together, we may learn more from common or opposing results about the nature of the news flow.

Literature Review

- Many studies have identified potentially useful forecasting variables (Hong and Yogo 2012, Acharya et al. 2013, Gorton et al. 2013, Yang 2013), as well as evidence of the importance of macroeconomic news and the oil market (Elder et al. 2013, Baumeister and Hamilton 2019), time variation in oil risk premium (Baumeister and Kilian 2017), and the potential usefulness of NLP (Loughran et al. 2019 use simple NLP measures to forecast high-frequency oil returns).
- Ours is the first study to apply NPL to longer horizons (4 or 8 weeks), to use a comprehensive NPL mapping, and to look at many endogenous variables simultaneously. This follows several recent related studies of other markets (equity and FX) where such long-horizon effects have been identified (Calomiris and Mamaysky 2019a, 2019b).

Our Baseline ("Kitchen Sink") Model

- Lagged returns, lagged volatility, change in lagged volatility.
- Lagged change in oil production, change in inventories.
- Lagged VIX, change in VIX.
- Lagged 10 Year Treasury yield, trade weighted dollar return, S&P return.
- Lagged futures basis (ratio of 3 mo. Futures Price/1 mo. Futures Price)
- Lagged Change in BH WIPI (yoy).
- Time trend.

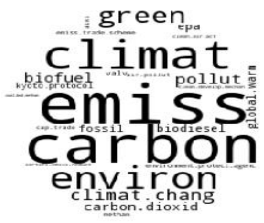
Our Augmented (with NLP Variables) Model

- All variables are lagged, averaged from prior four weeks.
- *artcount, entropy, f[topic], s[topic]*. This permits sentiment and frequency to have different implications (e.g., if a topic is more supply-related then sentiment or frequency relating to it may have opposite implications for price; Calomiris and Mamaysky 2019a found this was relevant for equity returns).
- Seven topics are identified by applying the Louvain method to a list of energy words (441 tokens – including bigrams and trigrams – constructed using a starter list of 387 from industry glossaries, then adding other words based on co-occurrence in the corpus and high frequency in the corpus).
- Topics that emerge, with labels we attach to them: *Co* (companies), *Gom* (global oil market), *Env* (environment), *Epg* (energy/power generation), *Bbl* (crude oil physical), *Rpc* (refining and petrochemical), *Ep* (exploration and production)
- TR Corpus is available from 1998 to 2019. Over a million articles (tagged as energy). We use Loughran-McDonald sentiment dictionary.

Company

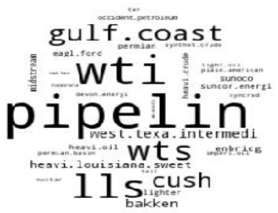
Global oil market

Environment



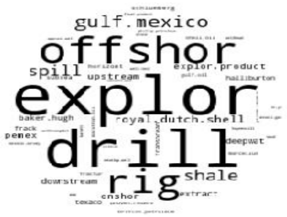
Energy/power generation

Crude oil physical



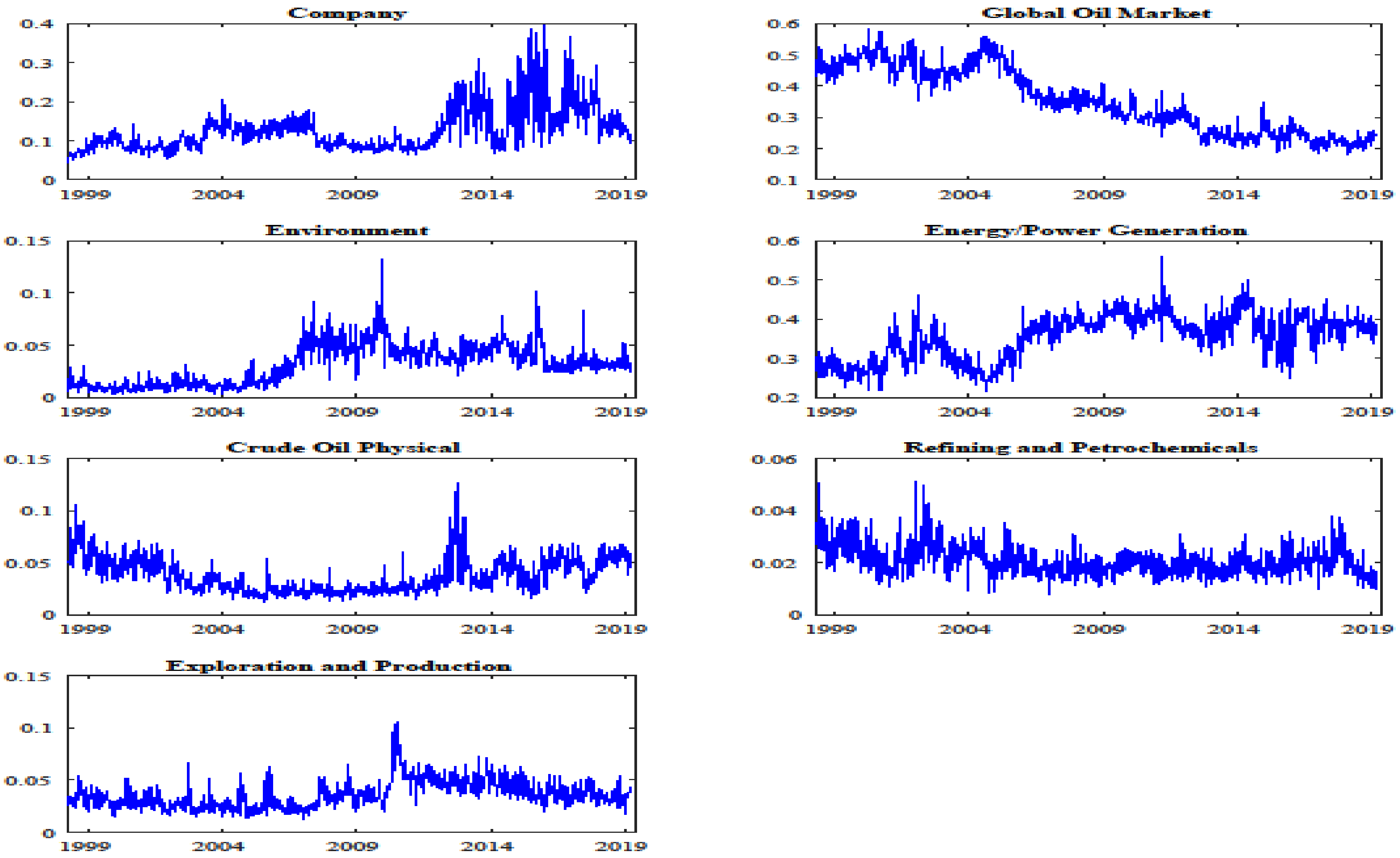
Refining and petrochemicals

Exploration and production

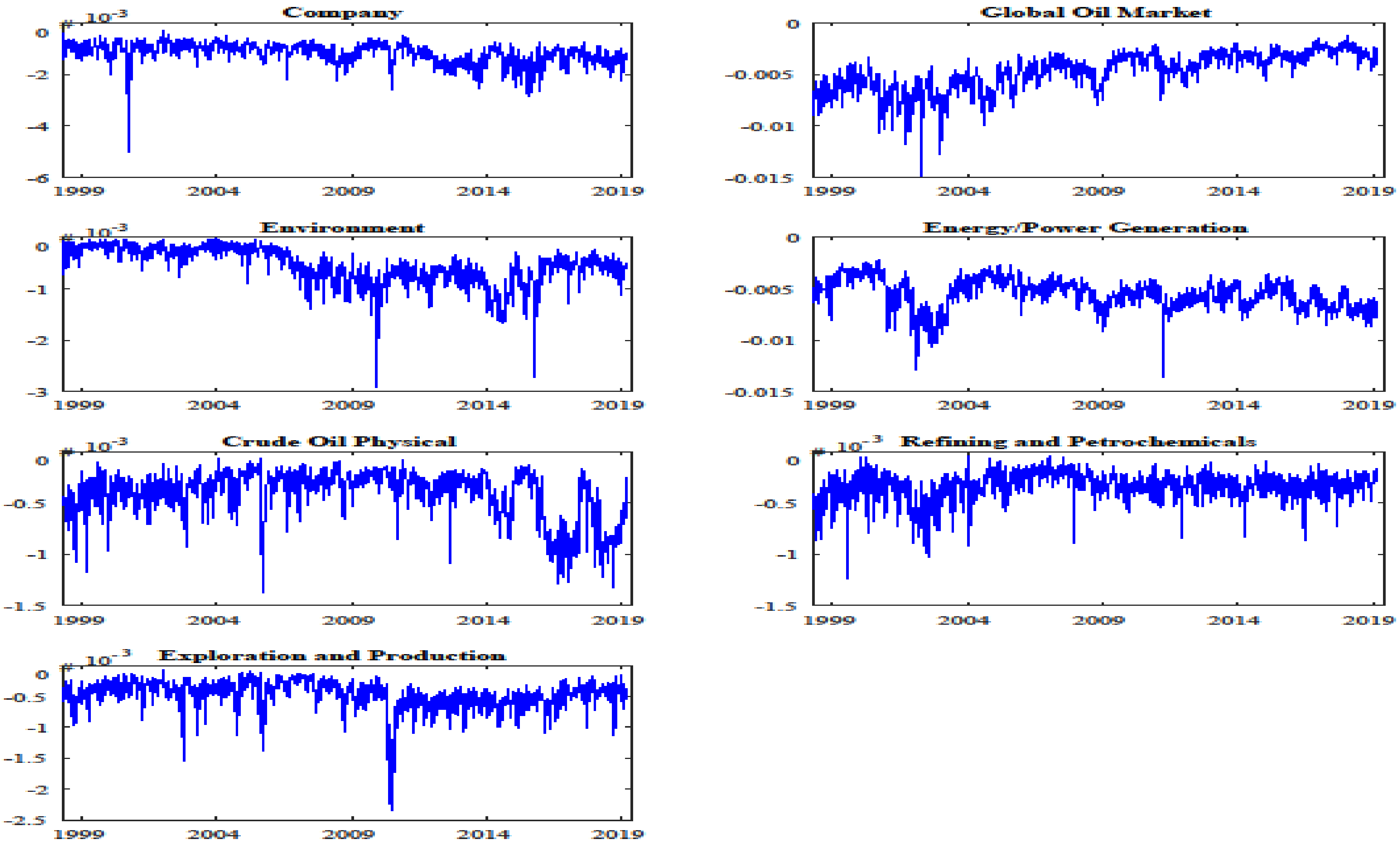


Topic	WordList
global oil market (gom)	oil (4,136,780), barrel (1,226,580), brent (526,719), refin (411,872), crude.oil (409,276), opec (394,754), petroleum (293,525), heat (291,997), diesel (276,319), barg (194,018), ipe (175,841), distil (167,863), tanker (142,160), sulphur (140,039), gallon (136,243), eia (127,622), nwe (70,962), ara (62,293), energi.inform.administr (55,927), bunker (47,736)
energy/power generation (epg)	gas (2,082,748), energi (1,385,165), coal (510,535), outag (409,463), nuclear (381,919), electr (326,305), generat (225,899), equiti (184,324), mine (178,868), lead (165,664), lng (162,184), addit (142,116), reactor (125,164), renew (120,903), solar (101,509), case (91,068), miner (90,722), grid (79,484), hydro (69,220), power.generat (53,787)
company (co)	fuel (1,483,081), bp (369,851), shell (369,655), vitol (237,656), mrpl (220,506), hsfo (158,878), glencor (144,515), exxon (136,651), mop (121,139), hin.leong (113,240), ceypetco (102,883), chevron (96,915), bpcl (95,996), petrochina (93,576), bapco (92,908), essar (90,448), blend (88,597), pertamina (84,403), trafigura (83,198), forti (81,329)
crude oil physical (bbl)	pipelin (409,704), wti (321,512), lls (169,911), wts (117,949), gulf.coast (68,858), cush (53,943), west.texa.intermedi (35,191), bakken (31,987), heavi.louisiana.sweet (22,581), enbridg (18,568), midstream (17,634), permian (13,138), sunoco (12,958), heavi.crude (9,681), lighter (8,541), heavi.oil (8,333), eagl.ford (8,053), suncor.energi (7,419), occident.petroleum (5,411), permian.basin (5,366)
Environment (env)	emiss (189,038), carbon (176,792), climat (105,968), environ (61,429), green (49,666), climat.chang (46,992), pollut (45,532), biofuel (32,514), carbon.dioxid (24,075), epa (22,403), biodiesel (19,407), global.warm (19,067), fossil (18,012), valv (10,182), kyoto.protocol (9,235), environment.protect.agenc (8,036), methan (7,179), emiss.trade.scheme (6,951), alki (6,204), air.pollut (4,723)
exploration & production (ep)	explor (148,206), drill (137,958), offshor (123,543), rig (94,639), shale (58,500), gulf.mexico (52,649), spill (46,891), royal.dutch.shell (37,685), onshor (28,528), pemex (26,894), explor.product (23,701), upstream (23,476), downstream (21,409), baker.hugh (17,968), deepwat (17,860), extract (17,115), halliburton (11,329), texaco (10,093), frack (9,383), transocean (9,373)
refining & petrochemicals (rpc)	reform (110,766), petrochem (88,297), cement (22,637), lpg (20,345), feedstock (18,355), propan (18,259), crude.distil.unit (12,005), netback (7,888), butan (7,407), liquefi.petroleum.gas (6,682), octan (6,045), fluid.catalyt.cracker (5,842), ethan (5,737), visbreak (5,079), olefin (4,370), oxygen (3,418), benzen (2,738), tertiari (2,081), polym (2,075), urea (1,830)

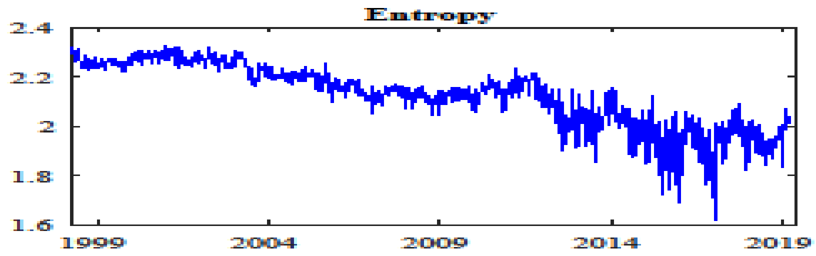
Panel A: Topical Frequency

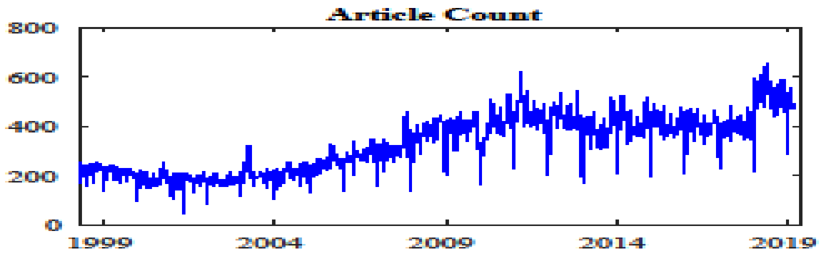


Panel B: Topical Sentiment



Panel C: Article Unusualness and Count





Topic	Sentiment	Headline
company (co)	0.0076	Glencore holds talks with Chinalco over Rio Tinto tie-up –Bloomberg
company (co)	0.0098	UPDATE 1-Asia Jet Fuel-China Aviation issues Q4 tender
company (co)	(0.0716)	RPT-UPDATE 3-U.S. board issues urgent call for BP safety panel
company (co)	(0.0488)	BP appeals Russian court ruling on office search
global oil market (gom)	0.0309	Algeria says oil producers mulling cuts beyound March
global oil market (gom)	0.0280	Oil price not dramatic for German economy-Mueller
global oil market (gom)	(0.0730)	ANALYSIS-Chavez referendum defeat poses possible oil risk
global oil market (gom)	(0.0714)	U.S. crude falls over $2, Brent extends losses

Environment (env)	**0.0179**	**TABLE-EU releases preliminary 2006 C02 emissions data**
Environment (env)	0.0318	UPDATE 1-Obama sees climate deal in Copenhagen -White House
Environment (env)	(0.0726)	German court document names 150 CO2 tax fraud suspects
Environment (env)	(0.0647)	EU's big 3 van makers put brakes on CO2 curbs
energy/power generation (epg)	0.0192	Germany's big four utilities to boost transparency
energy/power generation (epg)	0.0168	RITV-Cheaper Solar Power in Pipeline: Areva - New show available
energy/power generation (epg)	(0.0783)	Moody's cuts Enron, warns of ""low"" recovery rates
energy/power generation (epg)	(0.0894)	NATGAS PIPELINE CRITICAL NOTICE: Southern Natural Gas Revised Fairburn Force Majeure Notice
crude oil physical (bbl)	0.0270	U.S. cash crude price slide linked to futures fall
crude oil physical (bbl)	0.0229	November U.S. cash crudes trade quietly, WTS firm
crude oil physical (bbl)	(0.0500)	U.S. Cash Crude - WTI/Midland firms on cold supply concerns

refining & petrochemicals (rpc)	**0.0169**	**Union Carbide <UK.N> seeks E.Europe petchem deals**
refining & petrochemicals (rpc)	0.0432	India's Reliance, GAIL sign petrochemicals deal
refining & petrochemicals (rpc)	(0.0765)	TEXT-S&P cuts LyondellBasell Industries rating to 'B-'
refining & petrochemicals (rpc)	(0.1075)	UPDATE 1-Brazil's political crisis halts labor reform bill
exploration & production (ep)	0.0556	Mexico says implementing measures to boost Pemex finances
exploration & production (ep)	0.0667	BRIEF-SSE in offshore wind pact with Siemens, Subsea 7, Atkins
exploration & production (ep)	(0.0718)	UPDTAE 1-Goldman removes Halliburton from conviction buy list
exploration & production (ep)	(0.0702)	BRIEF-Halliburton says in case of deal termination it would have to pay $1.5 bln as fees to Baker Hughes

Empirical Findings

- Many NLP measures (of all types) are statistically significant (t values) and add substantially to the R-squared of the forecasting regressions (for futures returns, 8 week adjusted R-squared is 24% in Baselin, 43% in augmented; for oil inventories, 8 week adjusted R-squared rises from 11% to 25%).
- 8-week forecasting works “better” (bigger incremental contribution to R-squared, and larger – often more than double – coefficients). Forecasting effects are highly persistent.
- The NLP measures are useful for the two oil returns variables, oil volatility, the three oil companies stock returns, and inventories, not so much for production.
- The significant NLP measures *always* have the same sign for the five returns measures.
- NLP measures, when significant, tend to affect five returns measures with the same sign, and opposite sign for volatility. (This is also visible for many but not all Baseline variables, such as VIX and WIPI, OilVol).

Table 13: Summary of NLP Significant Predictors Using the coefficient estimates derived for the augmented models in Tables 5-12, if an NLP forecasting variable is statistically significant at the 10% level for one or both of the models, the sign of the coefficient appears below. Note that the signs for the 4-week and 8-week models never conflict.

Dependent Variables	Futures Oil return	Spot Oil return	Oil volatil.	Exxon return	BP return	Shell return	Oil Inven-tories	Oil Prod.
Forecasting Variables								
artcount	-	-	+		-		-	
entropy	+	+	-		+	+	-	
sCo				+	+	+		
fCo	-	-						
sGom		+	-					
fGom	-		+	-	-	-		
sEnv	+	+			+			
fEnv								
sEpg	+	+		+	+	+		
fEpg	-							
sBbl	+							
fBbl								
sRpc	-							
fRpc							-	
sEp				-	-	-	-	
fEp				-	-	-		

Empirical Findings (Cont'd)

- That opposite sign finding may either reflect non-priced news, which may makes sense for variables that the market has not been tracking (but which is strange for the Baseline models, which the market knows well), or a negative risk premium. In future work, we will explore this further.
- Sentiment tends to have a positive effect on returns (and negative on volatility), but not in all cases. As in Calomiris and Mamaysky (2019a), context matters for the directional effect of sentiment. In the oil market setting, we believe this reflects the difference between supply and demand. Positive supply shocks should have negative effects on price; the opposite for demand shocks. We believe that *sEp* and *sRpc* are naturally thought of as positive sentiment about supply, while other effects are more dominated by demand-side stories.
- Negative sign on *entropy* in volatility may seem strange, but it has same sign for inventory (as does OilVol)

Conclusions

- Many NLP measures are useful to add to more traditional forecasting variables for the oil market. They are significant individually, they add a lot to R-squared, and they matter persistently (effects are not diminishing from the fourth to the eighth week).
- It is useful to examine their usefulness for a variety of oil-related dependent variables. When one does so, there is remarkable consistency regarding which NLP measures matter for the five returns variables, and their opposite effects for volatility.
- Opposite effects indicate either a lack of risk pricing of these measures, or negative risk premia.
- Future work will examine the usefulness of NLP measures in models of time-varying risk, which will also inform the question of whether the NLP news is priced risk (with negative risk premia). We also will provide out-of-sample lasso estimation, and Monte Carlo simulations to correct for any bias in R-squared from the use of overlapping observations (but this is not likely a problem, as non-overlapping regressions give similar results).