



```

1 . cd "C:\Users\eliven\Dropbox\ELLW_2026\code"
  C:\Users\eliven\Dropbox\ELLW_2026\code

2 . doedit "04_dataset_cleaning.do"

3 . do "C:\Users\eliven\Dropbox\ELLW_2026\code\04_dataset_cleaning.do"

4 . *****
  > *****
5 . *****
  > *****
6 . *****
  > *****
7 .
8 . * Clean datasets and prepare controls used in main paper analyses
9 .
10 . *****
  > *****
11 .
12 . * Set directory
13 . global dataset "datasets"

14 .
15 .
16 . *****
  > *****
17 . * Import daily aggregate dataset
18 . use "$dataset\daily", clear

19 .
20 . * Merge industry classification
21 . merge m:1 stkcd using "$dataset\industry", keep(1 3) nogen

```

Result	Number of obs
Not matched	125,820
from master	125,820
from using	0
Matched	2,637,581

```

22 .
23 .
24 . * Categorize earnings announcements (EA) based on SUE
25 . gen neg_ea = event_ea & SUE_netearn < 0

26 . gen pos_ea = event_ea & SUE_netearn > 0
  >

27 .
28 . * Merge firm date datasets about query volume provided by the platform
29 . merge 1:1 stkcd date using "$dataset\firm_date"
    (variable date was int, now float to accommodate using data's values)

```

Result	Number of obs	
Not matched	1,999,065	
from master	1,903,755	(_merge==1)
from using	95,310	(_merge==2)
Matched	859,646	(_merge==3)

```

30 . keep if _merge == 3
    (1,999,065 observations deleted)

31 . drop _merge

32 .
33 . * Merge management guidance dataset
34 . merge m:1 stkc date using "$dataset\guidance_formerge", keep(1 3) nogen

```

Result	Number of obs
Not matched	857,587
from master	857,587
from using	0
Matched	2,059

```

35 .
36 . * Replace missing values with 0 for query counts, guidance attributes, and other daily
    > controls
37 . mvencode *_topics specific_q_* total_sentence reason_sentence_ratio total_word SVI_All
    > annual_report quarter_report *_q ques_hudong ind* *_reason mef_text *_c*, mv(0) overrid
    > e
content_to~s: 857587 missing values recoded
reason_top~s: 857587 missing values recoded
specif~0ge_1: 720416 missing values recoded
specifi~0g_1: 724009 missing values recoded
specifi~0g_0: 687042 missing values recoded
specif~0ge_0: 688568 missing values recoded
specif~0ge_0: 757730 missing values recoded
specifi~0g_0: 753421 missing values recoded
specifi~0g_1: 805713 missing values recoded
specif~0ge_1: 803135 missing values recoded
total_sent~e: 857587 missing values recoded
reason_sen~o: 857587 missing values recoded
total_word: 857587 missing values recoded
SVI_All: 148135 missing values recoded
annual_rep~t: 854536 missing values recoded
quarter_re~t: 854536 missing values recoded
gen_task_q: 651684 missing values recoded
edu_task_q: 651684 missing values recoded
info_proc_q: 651684 missing values recoded
info_sense_q: 651684 missing values recoded
stock_rec_q: 651684 missing values recoded
event_mon_q: 651684 missing values recoded
other_sens~q: 651684 missing values recoded
info_searc~q: 651684 missing values recoded
cur_info_q: 651684 missing values recoded
hist_info_q: 651684 missing values recoded
search_q~l_q: 651684 missing values recoded
search_q~t_q: 651684 missing values recoded
src_report_q: 651684 missing values recoded
src_announ~q: 651684 missing values recoded
src_invest~q: 651684 missing values recoded
src_otherc~q: 651684 missing values recoded
src_mediar~q: 651684 missing values recoded
src_regula~q: 651684 missing values recoded
src_otherp~q: 651684 missing values recoded
info_verif~q: 651684 missing values recoded
verif_repo~q: 651684 missing values recoded
verif_anno~q: 651684 missing values recoded
verif_inve~q: 651684 missing values recoded
verif_ot~p_q: 651684 missing values recoded
verif_medi~q: 651684 missing values recoded
verif_regu~q: 651684 missing values recoded
verif_ot~t_q: 651684 missing values recoded
cause_anal_q: 651684 missing values recoded
summary_q: 651684 missing values recoded
compare_q: 651684 missing values recoded
sameco_tim~q: 651684 missing values recoded
trend_pred_q: 651684 missing values recoded

```

trend_qual_q: 651684 missing values recoded
 trend_quan~q: 651684 missing values recoded
 co_trend_q: 651684 missing values recoded
 ind_trend_q: 651684 missing values recoded
 impact_ana~q: 651684 missing values recoded
 co_impact_q: 651684 missing values recoded
 ind_impact_q: 651684 missing values recoded
 macro_impa~q: 651684 missing values recoded
 overall_ev~q: 651684 missing values recoded
 fs_anal_q: 651684 missing values recoded
 risk_anal_q: 651684 missing values recoded
 strat_anal_q: 651684 missing values recoded
 other_anal_q: 651684 missing values recoded
 thumbsup_c~q: 651684 missing values recoded
 copy_count_q: 651684 missing values recoded
 editordown~q: 651684 missing values recoded
 share_coun~q: 651684 missing values recoded
 all_positi~q: 651684 missing values recoded
 thumsuporc~q: 651684 missing values recoded
 thumsupore~q: 651684 missing values recoded
 thumsupors~q: 651684 missing values recoded
 copyorshar~q: 651684 missing values recoded
 copyoredit~q: 651684 missing values recoded
 shareoredi~q: 651684 missing values recoded
 thumsupsha~q: 651684 missing values recoded
 info_query_q: 651634 missing values recoded
 info_analy~q: 651634 missing values recoded
 shareholde~q: 651634 missing values recoded
 product_te~q: 651634 missing values recoded
 customer_b~q: 651634 missing values recoded
 macro_env_q: 651634 missing values recoded
 supply_cha~q: 651634 missing values recoded
 regulation~q: 651634 missing values recoded
 restructur~q: 651634 missing values recoded
 invest_dec~q: 651634 missing values recoded
 finance_de~q: 651634 missing values recoded
 production~q: 651634 missing values recoded
 exec_team_q: 651634 missing values recoded
 internal_c~q: 651634 missing values recoded
 talent_mgm~q: 651634 missing values recoded
 csr_q: 651634 missing values recoded
 market_com~q: 651634 missing values recoded
 business_o~q: 651634 missing values recoded
 financial_~q: 651634 missing values recoded
 stock_perf_q: 651634 missing values recoded
 listing_de~q: 651634 missing values recoded
 bank_q: 651634 missing values recoded
 renewable_~q: 651634 missing values recoded
 ai_q: 651634 missing values recoded
 pharma_q: 651634 missing values recoded
 media_ente~q: 651634 missing values recoded
 education_q: 651634 missing values recoded
 specific_q: 651634 missing values recoded
 ques_hudong: 708683 missing values recoded
 industry: 3427 missing values recoded
 industry_c~t: 547494 missing values recoded
 ind_stockr~o: 547494 missing values recoded
 financial_~n: 857587 missing values recoded
 production~n: 857587 missing values recoded
 tech_reason: 857587 missing values recoded
 competitio~n: 857587 missing values recoded
 business_o~n: 857587 missing values recoded
 mef_text: 857587 missing values recoded
 SUE_comearn: 852272 missing values recoded
 cop_pos_c: 662418 missing values recoded
 cop_neg_c: 662418 missing values recoded
 copy_posit~t: 662418 missing values recoded
 copy_negat~t: 662418 missing values recoded
 n_cop_avg_~t: 670649 missing values recoded
 n_copy_ave~e: 708578 missing values recoded
 n_cop_pos_c: 662418 missing values recoded
 n_cop_neg_c: 662418 missing values recoded

```

n_cop_po~g_c: 662418 missing values recoded
n_cop_n~ig_c: 662418 missing values recoded
pos_sent_c: 662418 missing values recoded
neg_sent_c: 662418 missing values recoded
share_posi~t: 662418 missing values recoded
share_nega~t: 662418 missing values recoded
n_pos_sent_c: 662418 missing values recoded
n_neg_sent_c: 662418 missing values recoded
n_share_po~t: 662418 missing values recoded
n_share_ne~t: 662418 missing values recoded
cpsent_c: 662418 missing values recoded
cnsent_c: 662418 missing values recoded
cpsig_c: 662418 missing values recoded
cnsig_c: 662418 missing values recoded
n_cpshare_~t: 670717 missing values recoded
n_cpshare_~g: 708642 missing values recoded
n_cpsent: 662418 missing values recoded
n_cnsent: 662418 missing values recoded
n_cpsig_co: 662418 missing values recoded
n_cnsig_co: 662418 missing values recoded
ts_pos_sen~c: 662418 missing values recoded
ts_neg_sen~c: 662418 missing values recoded
ts_pos_sig_c: 662418 missing values recoded
ts_neg_sig_c: 662418 missing values recoded
n_ts_pos~t_c: 662418 missing values recoded
n_ts_neg~t_c: 662418 missing values recoded
n_ts_pos~g_c: 662418 missing values recoded
n_ts_neg~g_c: 662418 missing values recoded
posi~t_count: 651634 missing values recoded
nega~t_count: 651634 missing values recoded
posi~l_count: 651634 missing values recoded
nega~l_count: 651634 missing values recoded
general_co~t: 547494 missing values recoded
stock_rec_c: 547494 missing values recoded
total_gene~t: 547494 missing values recoded
add_count: 666395 missing values recoded
add_count_no: 692064 missing values recoded
add_count_~s: 808922 missing values recoded

```

```
38 .
```

```
39 . * Only include firms that have received at least one specific query in the sample perio
> d
```

```
40 . bys stkcd: egen sumspecific = sum(specific_q)
```

```
41 . drop if sumspecific == 0
(701 observations deleted)
```

```
42 .
```

```
43 . * Construct dummy variable for analyst report day and media coverage day
```

```
44 . gen AnaReport_D = AnaReport>0
```

```
45 . gen media_D = media>0
```

```
46 .
```

```
47 . * Generate postive feedback ratio per firm-day
```

```
48 . gen posratio = thumbsupsharecopy_count_q/specific_q
(650,983 missing values generated)
```

```
49 .
```

```
50 . * log(1+x) to retain observations with zero values
```

```

51 . replace AnaReport = log(1+AnaReport)
    variable AnaReport was byte now float
    (15,921 real changes made)

52 . replace media = log(1+media)
    variable media was int now float
    (104,284 real changes made)

53 . gen logwords = log(1+total_word)

54 . foreach v of varlist specific_q SVI ques_hudong {
    2. capture drop log`v'
    3. gen log`v' = log(1+`v')
    4. }

55 .
56 . * Generate "More Topics" and "Longer" variables for cross sectional tests
57 . * More topics - forecasts in which a larger number of topics are discussed than in the
    > sample median forecast
58 . * Longmef - forecasts with a word count above the sample median
59 . * Step 1: Calculate medians for variables (only for values greater than 0)
60 . * (1) reason_topics:
61 . summarize reason_topics if mef_text > 0, detail

```

Count of _reason variables > 0

Percentiles		Smallest		
1%	2	0		
5%	3	0		
10%	4	0	Obs	2,059
25%	5	0	Sum of wgt.	2,059
50%	7		Mean	6.881982
		Largest	Std. dev.	2.653483
75%	9	15		
90%	10	15	Variance	7.040973
95%	11	16	Skewness	.3098494
99%	14	18	Kurtosis	2.978281

```

62 . local median_reason_topics = r(p50)

63 . * (2) logwords:
64 . summarize logwords if mef_text > 0, detail

```

logwords

Percentiles		Smallest		
1%	4.890349	3.89182		
5%	5.062595	4.060443		
10%	5.17615	4.382027	Obs	2,059
25%	5.407172	4.382027	Sum of wgt.	2,059
50%	5.700444		Mean	5.742055
		Largest	Std. dev.	.4618508
75%	6.030685	7.221836		
90%	6.369901	7.261927	Variance	.2133062
95%	6.558198	7.365813	Skewness	.4254746
99%	6.993015	7.488853	Kurtosis	3.243411

```

65 . local median_logwords = r(p50)

```

```

66 . * Step 2: Generate group variables based on comparison with respective medians
67 . gen moretopics = reason_topics > `median_reason_topics'

68 . gen longmef = logwords > `median_logwords'

69 . label variable moretopics "Group: reason_topics > median"

70 . label variable longmef "Group: logwords > median"

71 .
72 .
73 . * Scale liquidity and volatility metrics
74 . foreach v of varlist spread esp_amount illiq AbnSpread20to1 AbnVol AbnVol20to1 vpin {
    2. replace `v' = `v' * 100
    3. }
    (595,590 real changes made)
    (594,681 real changes made)
    (583,652 real changes made)
    (594,700 real changes made)
    (594,844 real changes made)
    (595,553 real changes made)
    (594,548 real changes made)

75 .
76 . save "$dataset\firm_day_dataset",replace
    (file datasets\firm_day_dataset.dta not found)
    file datasets\firm_day_dataset.dta saved

77 .
78 .
79 . *****
    > *****
80 . * Import user interaction dataset
81 . use "$dataset\user_interaction",clear

82 .
83 . * Convert string date variable 'dt' into Stata's numeric date format (YMD)
84 . gen date = date(dt,"YMD")

85 .
86 . * log(1+x) to retain observations with zero values
87 . gen logwords_q = log(total_word_q+1)

88 . gen logwords_a = log(total_word_a+1)
    (39 missing values generated)

89 . gen logref = log(total_ref +1)
    (10,911 missing values generated)

90 .
91 . *
92 . gen pos_fb=1 if thumbsup==1|share==1|copy==1
    (1,558,637 missing values generated)

93 . replace pos_fb=0 if pos_fb==.
    (1,558,637 real changes made)

94 . foreach x of varlist pos_fb thumbsdown{
    2.         replace `x' = `x' *100
    3.         }
    (187,438 real changes made)
    (1,731 real changes made)

```

```

95 .
96 .
97 . * Convert topics into dummy variables
98 . local vars financial_perf_q business_outlook_q production_ops_q stock_perf_q product_te
   > ch_innov_q market_comp_q

99 .
100 . foreach v of local vars {
      2.     gen `v'_dummy = .
      3.     replace `v'_dummy = 1 if `v' == "是"
      4.     replace `v'_dummy = 0 if `v' == "否"
      5. }
(1,746,075 missing values generated)
(406,132 real changes made)
(1,043,327 real changes made)
(1,746,075 missing values generated)
(459,577 real changes made)
(989,882 real changes made)
(1,746,075 missing values generated)
(446,328 real changes made)
(1,003,131 real changes made)
(1,746,075 missing values generated)
(342,198 real changes made)
(1,107,261 real changes made)
(1,746,075 missing values generated)
(388,702 real changes made)
(1,060,757 real changes made)
(1,746,075 missing values generated)
(315,039 real changes made)
(1,134,420 real changes made)

101 .
102 . * Convert tasks into dummy variables
103 . local vars info_sense_q info_search_q cause_anal_q summary_q  compare_q trend_pred_q im
   > pact_anal_q overall_eval_q fs_anal_q

104 . foreach v of local vars {
      2.     gen `v'_dummy = .
      3.     replace `v'_dummy = 1 if `v' == "是"
      4.     replace `v'_dummy = 0 if `v' == "否"
      5. }
(1,746,075 missing values generated)
(677,665 real changes made)
(1,068,410 real changes made)
(1,746,075 missing values generated)
(570,735 real changes made)
(1,175,340 real changes made)
(1,746,075 missing values generated)
(291,768 real changes made)
(1,454,307 real changes made)
(1,746,075 missing values generated)
(58,452 real changes made)
(1,687,623 real changes made)
(1,746,075 missing values generated)
(235,552 real changes made)
(1,510,523 real changes made)
(1,746,075 missing values generated)
(238,112 real changes made)
(1,507,963 real changes made)
(1,746,075 missing values generated)
(326,034 real changes made)
(1,420,041 real changes made)
(1,746,075 missing values generated)
(277,820 real changes made)
(1,468,255 real changes made)
(1,746,075 missing values generated)
(255,969 real changes made)
(1,490,106 real changes made)

```

```

105 .
106 . save "$dataset\user_interaction",replace
    file datasets\user_interaction.dta saved

107 .
108 . *****
    > *****
109 . * Import user dataset
110 . use "$dataset\user_dataset",clear

111 .
112 . * log(1+x) to retain observations with zero values
113 . gen logquery=log(1+post_total_queries)

114 . gen logword=log(1+post_word_q)

115 .
116 . * Make these indicators dummy variables
117 . replace post_roe_q=1 if post_roe_q>1
    (1,316 real changes made)

118 . replace post_solvency_q=1 if post_solvency_q>1
    (1,111 real changes made)

119 . replace post_ar_q=1 if post_ar_q>1
    (1,136 real changes made)

120 .
121 .
122 . save "$dataset\user_dataset",replace
    file datasets\user_dataset.dta saved

123 .
    end of do-file

124 .

```