

Sentiment Analysis of Tweets about Remote Work

Sherly Hartono

Motivation

- Given the changing landscape of the workplace from office to home, people have different opinions regarding these large scale transitions. Many say this trend would likely continue after the pandemic.
- There will be many challenges that companies will need to overcome in order for remote work to stay. But the question still remains,
- **How do remote employees actually feel about remote work? and what are the challenges that they are facing daily?**
- This insight will help businesses to plan and prioritize their next step in this paradigm shift.

Approach

Collect Data

- Using Twint

Data Wrangling

- Remove irrelevant tweets
- Clean each tweets for Rule-Based and Transformers modelling

Rule Based NLP

- User VADER Package

Transformer: BERT

- Create two types of word clusters
- Get Cosine differences between tweets and the two clusters

Topic Modelling

- Choose results from the three models
- Clustering on negative tweets
- Generate Topics using TF-IDF

Collecting Data

- Twint: An advanced Twitter scraping & OSINT tool written in Python
- Doesn't use Twitter's API
- Scrape a user's followers, following, Tweets and more while evading most API limitations.
- Data can be saved in json file or converted to a pandas dataframe to store selected information

```
# Configure
config = twint.Config()

config.Search = "remote work"
config.Lang = "en"
config.Since = "2020-08-01"
config.Until = "2021-05-13"
config.Limit = 10000
config.Pandas = True
config.Filter_retweets = True
```

```
# Run
twint.run.Search(config)
```


```
1392630666471297029 2021-05-13 06:59:59 +0700 <RecruiterDotCom> Recruiter .com'
s April 2021 recruiter index® has found that the demand for in-person jobs is o
utpacing that of remote work. Hit the link below for the full Recruiter Index
®. https://t.co/vh2wufdTp0 #recruiterindex #recruiters #recruitment #jobmarke
t #labormarket #hiringtrends
1392630316494295042 2021-05-13 06:58:35 +0700 <rucs> @TimSackett @lruettimann
@FrankZupan @Lars +1. Real Estate Market would crash if there is no demand for
commercial space. Hybrid work / Remote work works . If we design for it. For
decades, Office space worked as space to socialize with fellow human beings.
1392630178359042054 2021-05-13 06:58:02 +0700 <ArneEkstrom1> Congratulations to
Dr. Michael Starrett on successfully defending his dissertation! Mike worked o
n everything from immersive VR with a treadmill to remote VR testing! Great wo
```

Data Wrangling

Remove tweets:

- Remove non english tweets using langdetect()
- Remove duplicate tweets
- Remove job opening and advertisement tweets
- Remove username with 'remote'
-

Clean tweets:

- Emojis are converted to words using emoji.demojized()
 → :smiling_face_with_sunglasses:
- Remove special characters

```
my_functions.py 4 X
1 from __future__ import division
2 import pandas as pd
3 import spacy
4 import string
5 import re, nltk
6 import nltk.corpus
7 from nltk.corpus import stopwords
8
9 from collections import Counter
10
11 > def print_tweet(the_df, end_index):--
12
13
15
16 > def print_empty_tweet(the_df):--
17
18
19 > def delete_empty_tweet(the_df):--
20
21
22
23
24
25
26
27
28
29 > def print_sentiment_tweet(the_df, sentiment = 'NEGATIVE'):-
30
31
32
33
34
35
36
37
38
39
40 # For VADER and BERT: Read learn more
41 # Remove all characters after the phrase read more, learn more, find ou
42 > def remove_regex(the_regex, the_df):-
43
44
45
46
47 > def remove_read_more(the_df):-
48
49
50
51
52
53
54
55
56
57 # 1. Remove End Hashtag
58 > def remove_end_hashtag(the_df): --
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82 #2. Lemmatizing
83 > def get_lemmatized_text(the_text):-
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103 > def lemmatized_df(the_df):-
104
105
106
107
108
109
110
111
112
113
114 #3. Remove Special Char
115 > def remove_char_from_text(the_text):-
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143 # OPTIONAL
144 # 1. Split Hashtags
145 WORDS = nltk.corpus.brown.words()
146 COUNTS = Counter(WORDS)
147
148 > def pdist(counter):-
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171 > def segment(text):-
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196 > def get_split_word(the_word):-
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240 > def split_hashtag(the_df):-
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

Rule Based Training : VADER

1. CLEANING

- Attuned to sentiments expressed in social media
- Create functions for this to create different pipeline (split and non split hashtags)
- Hashtags are removed but only the ones at the end of the tweets.
- Lemmatizing
- Remove phrases like read more, learn more, find out more, as they usually come with links where to read them after.

2. SCORE AND PREDICTION

	tweet	original_tweet	VaderScore	VaderSentiment
0	real Estate Market would crash demand commerci...	Real Estate Market would crash if there is no ...	-0.4939	NEGATIVE
1	Concur . company must work office say go . imp...	Concur. My company was 100% "you MUST work in ...	-0.0325	NEGATIVE
2	not ask really need thing ? I think would fair...	Why not ask if we really need that thing? I th...	0.4939	POSITIVE
3	dear Line Managers , Appraisal subordinate bas...	Dear Line Managers, Appraisal your subordinate...	0.3818	POSITIVE
4	I opportunity work cross functionally engage c...	I have had more opportunities to work cross-fu...	0.8225	POSITIVE
...

Study reveals growing **#cybersecurity #risks**
driven by **#remotework**



Study reveals growing **cyber security risks**
driven by **remote work**

Transformers : BERT

STEP 1: Create negative and word clusters

Method 1

Extract all unique words from our tweets and create their embeddings

	words	embeddings	sentence_idx
0	real	[0.84305966, 0.26398277, 0.32490465, 0.0248823...	0
1	estate	[0.5303595, -0.2668757, 0.75979805, 0.19747669...	0
2	market	[0.28153154, -0.28520963, 0.5332744, 0.2479845...	0
3	would	[-0.30469832, -0.46680853, -0.14788364, 0.0218...	0
4	crash	[-0.24616344, 0.039229684, 0.06317264, 0.44311...	0
...
64894	responsibility	[-0.008578587, 0.1220773, 0.24025224, -0.05287...	2994
64895	staff	[0.31579545, 0.04895662, 0.29348424, -0.167502...	2994
64896	welfare	[0.20462838, 0.31706288, 0.05414169, -0.091104...	2994
64897	cannot	[-0.12302854, -0.030138453, 0.32204738, -0.080...	2994
64898	avoided	[-0.40538037, -0.13855177, -0.17537802, 0.0789...	2994

64899 rows x 3 columns

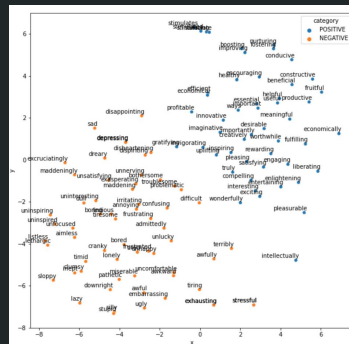
Method 2

Generate our own positive and negative words using gensim and create their embeddings

```
print(dict_clusters["POSITIVE"], '\n')
print(dict_clusters["NEGATIVE"], '\n')

['essential', 'depressing', 'constructive', 'helpful', 'truly', 'creatively', 'healthy', 'boosting', 'invigorating', 'intellectually', 'interesting', 'desirable', 'fostering', 'nurturing', 'exhausting', 'stimulated', 'gratifying', 'encouraging', 'profitable', 'emotionally', 'stimulating', 'efficient', 'fruitful', 'less', 'conductive', 'inspiring', 'uplifting', 'stressful', 'worthwhile', 'useful', 'surprisingly', 'pleasing', 'enormously', 'importantly', 'fulfilling', 'innovative', 'wonderfully', 'economically', 'exciting', 'meaningful', 'important', 'satisfying', 'entertaining', 'compelling', 'beneficial', 'liberating', 'ways', 'stimulate', 'quite', 'enlightening', 'improving', 'enjoyable', 'economical', 'incredibly', 'immensely', 'pleasurable', 'extremely', 'engaging', 'productive', 'imaginative', 'stimulates', 'rewarding']

['dull', 'confusing', 'bored', 'depressing', 'pathetic', 'lazy', 'listless', 'tedious', 'exhausting', 'frustrated', 'unsatisfying', 'bothersome', 'maddeningly', 'unhappy', 'disheartening', 'bit', 'exceedingly', 'lonely', 'stressful', 'exasperating', 'ugly', 'excruciatingly', 'terribly', 'miserable', 'frustrating', 'downright', 'irritating', 'annoying', 'disappointing', 'dreary', 'cranky', 'awful', 'timid', 'silly', 'stupid', 'troublesome', 'slippy', 'difficult', 'dispiriting', 'unlucky', 'uninteresting', 'tiresome', 'uninspired', 'lethargic', 'boring', 'enjoyable', 'incredibly', 'sad', 'inept', 'extremely', 'problematic', 'tired', 'awkward', 'embarrassing', 'tiring', 'unfocused', 'unnerving', 'clumsy', 'uninspiring', 'uncomfortable', 'admittedly', 'pretty', 'maddening', 'awfully', 'aimless']
```



Transformers : BERT

STEP 2: Create Sentence Embedding

STEP 3: Compute Cosine Distance between sentences and the two average clusters

Average Embedding of word clusters

df_own_dict		
<class 'list'>		
	sentiment	embedding
0	positive_own	[-0.23684903980000002, 0.11508918550000001, 0....
1	negative_own	[-0.45195508, 0.0403723456, 0.5418859124000001...

df_data_dict		
	sentiment	embedding
0	positive_data	[-0.6036096215, 0.11336755750000001, 0.6376053...
1	negative_data	[-0.7820361257, 0.20127013330000001, 0.5900069...

Distance
between

Sentence Embedding

df_sentence.head(10)		
	tweet	tweet_embedding
0	Real Estate Market would crash if there is no ...	[0.2257416844, -0.0169709176, 0.6545215249, -1...
1	Concur. My company was 100% "you MUST work in ...	[-0.3543173969, -0.3487285674, 0.2804761231000...
2	Why not ask if we really need that thing ? I t...	[-0.733558774, -0.3607640862, 0.1692087054, -0...
3	Dear Line Managers Appraisal your subordinate ...	[-0.6378751993, 0.2945272923, 0.22799032930000...
4	I have had more opportunities to work cross-fu...	[-0.2694126964, -0.3058058321, 0.7276366353, -...
5	Study reveals growing cybersecurity risks driv...	[-1.0207954645, -0.6177105904, 0.7027013302, -...
6	As a remote employee you may be tempted to che...	[-0.6545557976, 0.033209234500000004, 0.194388...
7	I am lucky mine is moving to a hybrid model. S...	[-0.6591749787, -0.9909458756, 0.4506825209, -...
8	Shifting to a #remotework environment created ...	[-0.4876919985, -0.2119547725, 0.6156088710000...
9	professionals from a range of industries who n...	[-0.7931020856000001, -0.4555655718, 0.8284254...

Transformers : BERT

STEP 5: Get results of prediction by VADER and two clustering results

	tweet	tweet_embedding	data_predicted	own_predicted	VaderSentiment
0	Concur. My company was 100% "you MUST work in ...	[-0.3543173969, -0.3487285674, 0.2804761231000...	POSITIVE	NEGATIVE	NEGATIVE
1	So weird how we all decided to never replace t...	[0.3376812041, -0.1152411997, 0.530261457, 0.3...	POSITIVE	NEGATIVE	NEGATIVE
2	Glad you exposed me to this. What bizarre argu...	[-0.5353716612, 0.3261777461, 0.2838433087, -0...	POSITIVE	NEGATIVE	NEGATIVE
3	Any piece that tries to say why we should ditc...	[-0.2667962313, 0.4768332839, 0.48676985500000...	POSITIVE	NEGATIVE	NEGATIVE
4	So are you telling me I should leave my eight-...	[-0.8893477321000001, -0.361951381, -0.2480771...	NEGATIVE	NEGATIVE	NEGATIVE
...
128	Yeah. A lot of boundaries have dissolved with ...	[0.06808185580000001, -0.4628287554, 0.2636246...	NEGATIVE	NEGATIVE	NEGATIVE
129	Getting very tired of the corporate-sponsored ...	[-0.21660083530000002, -0.0097534209, 0.465162...	NEGATIVE	NEGATIVE	NEGATIVE
130	Counter #2 Most people I know loathe their cow...	[-0.2670706809, 0.0491347052, 0.39799004790000...	NEGATIVE	NEGATIVE	NEGATIVE
131	its funny bc the argument against remote work ...	[-0.0675944909, 0.10974104700000001, 0.4871161...	POSITIVE	NEGATIVE	NEGATIVE
132	The announcing and act of departing a zoom cal...	[-0.6881164908, -0.2324504852, 0.3338147998, -...	NEGATIVE	NEGATIVE	NEGATIVE

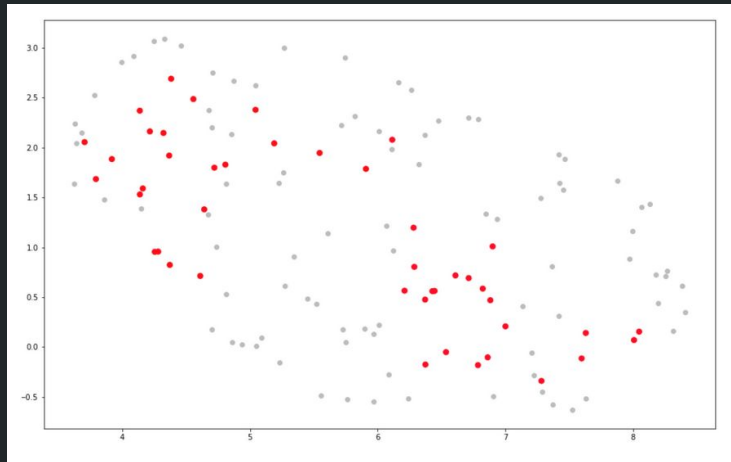
TOPIC MODELLING

STEP 1: Dimensionality reduction to our negative tweets

STEP2: Cluster our negative tweets into topics

The intuition behind the method is as follows. When you apply TF-IDF as usual on a set of documents, what you are basically doing is comparing the importance of words between documents.

What if, we instead treat all documents in a single category (e.g., a cluster) as a single document and then apply TF-IDF? The result would be a very long document per category and the resulting TF-IDF score would demonstrate the important words in a topic.



TOPIC MODELLING

STEP 3: TF-IDF SCORE

Now, we have a single **importance** value for each word in a cluster which can be used to create the topic. If we take the top 10 most important words in each cluster, then we would get a good representation of a cluster, and thereby a topic.

```
0: [('workers', 0.038508393381364706),  
    ('communication', 0.03385479321030039),  
    ('person', 0.031286203277695206),  
    ('article', 0.031286203277695206),  
    ('disagree', 0.031286203277695206),  
    ('pandemic', 0.03080671470509176),  
    ('companies', 0.02929385014096119),  
    ('culture', 0.02929385014096119),  
    ('company', 0.0262896337425575),  
    ('arguing', 0.024983344926558383),  
    ('documentation', 0.024983344926558383),  
    ('forced', 0.024983344926558383),  
    ('concept', 0.024983344926558383),  
    ('propaganda', 0.024983344926558383),  
    ('shows', 0.024983344926558383),  
    ('ppl', 0.024983344926558383),  
    ('corporate', 0.024983344926558383),  
    ('employee', 0.02256986214020026),  
    ('fear', 0.02256986214020026),  
    ('colleagues', 0.02256986214020026)],
```

```
1: [('home', 0.04191234731448262),  
    ('going', 0.034411755787602506),  
    ('really', 0.03132381892730256),  
    ('day', 0.028650261179633346),  
    ('email', 0.027957883780068058),  
    ('office', 0.027113120920628328),  
    ('lot', 0.02651546654705727),  
    ('life', 0.026177483104688724),  
    ('year', 0.023492864195476915),  
    ('time', 0.023215946033186635),  
    ('energy', 0.022325542274796852),  
    ('use', 0.022325542274796852),  
    ('likely', 0.022325542274796852),  
    ('spent', 0.022325542274796852),  
    ('turning', 0.022325542274796852),  
    ('hired', 0.022325542274796852),  
    ('mother', 0.022325542274796852),  
    ('chance', 0.022325542274796852),  
    ('slack', 0.02016881297634917),  
    ('sick', 0.02016881297634917)]
```

STEP 4: Evaluation

The topic on the left seems to be about disagreement between workers/ employees and companies as predicted. The keywords on the right are not that obvious. But we know from manual evaluation that there are concerns about cybersecurity and isolation.

FUTURE WORK



Due to CPU constraints, I am not able to fully use the 100,000 tweets that I extracted which could be the reason why our K-Means clustering is not doing so well.

Another method that I would like to try is to train the DistilBERT model on labelled data, that is tweets that have already been labelled negative, positive, and neutral but on other topics, then use the trained model to predict my tweets about remote work.

It would also be more insightful to include texts from other sources such as news portals, Quora, and reddit which have longer text and thus will give us a better topic prediction.