Capstone 3:
Sentiment Analysis on Remote Work

Content:

1. Introduction

Problem Statement

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. However, 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.
I plan to classify tweets according to their sentiments (positive, negative, or neutral) to answer the first question. I would also like to implement simple topic identification on the negative tweets to get some ideas on the problems people have to face in relation to remote working.

Data

I will scrape Twitter using twint. There is no limit in the amount of data entries that I can extract. I plan to use 100,000 entries.

	date	tweet	username
0	2021-05-13 06:59:59	Recruiter .com's April 2021 recruiter index® h...	RecruiterDotCom
1	2021-05-13 06:58:35	@TimSackett @lruettimann @FrankZupan @Lars +1....	rucsb
2	2021-05-13 06:58:02	Congratulations to Dr. Michael Starrett on suc...	ArneEkstrom1
3	2021-05-13 06:55:20	@AlvarezGibson Concur. My company was 100% "yo...	thirdnline
4	2021-05-13 06:55:01	@mattreign Why not ask if we really need that ...	toofarnorth49
...	...	...	...
10027	2021-05-10 02:01:34	Banks don't like bitcoin. Taxis don't like Ube...	jxtphan
10028	2021-05-10 02:01:28	During the pandemic, many people found that th...	LatitudeGuides
10029	2021-05-10 02:00:56	Management Strategies for Developing an Effect...	KenCollinsMktng
10030	2021-05-10 02:00:29	Peter Weckesser, Chief Digital Officer at Schn...	LizHendersonSE
10031	2021-05-10 02:00:19	.@GoDaddy is looking for a #PublicRelations Ma...	RemotePRJobs

10032 rows × 3 columns

2. Data Wrangling

1. Remove non english tweets

First we will remove non-english tweets. As this is an unsupervised problem, I will need to be able to evaluate the result myself. Since I am not fluent with other languages, I will stick to english tweets. I used the *langdetect* library for this. Langdetect cannot detect some languages and throws out errors. So I will need to remove that manually.

2. Remove ads

There are many tweets that include http links. Since this does not have any effect on the sentiment, I will remove the link from the tweets.

There are also many job ads that include remote work hashtags:

- *We're #hiring a P/T In House #Editor! Come work with us and dream forward positive futures! #job #freelance #remotework @WritersofColor https://t.co/m2JTEpMIxo*
- *Green Man Gaming are on the lookout for an EVP Performance Marketing Fully remote! Based in 🇬🇧 UK https://t.co/aNRXbQt8T6 #remote #job #remotework*

These tweets almost always include #job, #hiring, #remotejob hashtags which makes it relatively easy to remove them. There are also other hiring tweets that include the words 'seeking', 'looking for' , ' job vacancy'. I decided to remove all of them by using all of these keywords.

3. Remove username with 'remote'

There are also tweets that will not be giving us any insights such as :

- *<RemoteWorkNews> Weight loss app Lose It mulls 100% remote work in exchange for salary cut.*
- *<remote_wander> @Mr_Arizona2424 @MikeGroat1 @secupp It literally does work like that and it's worked like that since the end of the Second World War. Why do you think the US dollar is the worlds reserve currency? We have pretty beaches?*

These are captured because the tweets are posted by users that have the word 'remote' as its username, but their tweets are usually ads, and a reply to other users with not much relevance to remote working. So I removed these too.

4. Remove unwanted characters and words

We shouldn't simply remove emojis in sentiment analysis as smileys can give huge cues in the sentiment. So I use emoji library's demozie method to convert emoji to words. For example:
😎 will be converted to *:smiling_face_with_sunglasses:*

5. Other cleaning

Other general cleaning includes removing mention @some_user, expanding contractions, and removing duplicates.

3. VADER Training

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains. Since this is a rule based NLP, I need to do further cleaning since the model cannot really understand context.

The further cleaning includes:

1. Removing hashtags that appears at the end of the sentence since these are not a part of a sentence
2. Remove phrases like read more, learn more, find out more, as they usually come with links where to read them after.
3. Lemmatizing
4. Remove special characters other than period and question mark.

There are hashtags that appear in the middle of a sentence. These are usually a part of a sentence:

- *Shifting to a **#remotework** environment created challenges for many businesses &; government institutions.*
- *Study reveals growing **#cybersecurity** **#risks** driven by **#remotework** Report shows that changing #work styles and behaviors are creating new **#vulnerabilities** for companies, individuals, and their data. #workfromhome #insiderthreats #cyberrisk #cyberthreat.*

I extracted these words and split them so that our sentences have proper words instead of concatenated hashtags. Since I'm not sure which would give a better result, I made two pipelines for the data preprocessing, one with split hashtags words, and one with the hashtags in the middle of the sentence intact.

They both turn out to give the same prediction. There are 3 times more positive tweets than negative. I won't be evaluating them at this stage, since I want to compare it with the result using transformers.

4. BERT Training

DistilBERT is a small, fast, cheap and light Transformer model trained by distilling BERT base. It has 40% less parameters than bert-base-uncased, runs 60% faster while preserving over 95% of BERT's performances as measured on the GLUE language understanding benchmark.

There are two methods that I want to get the sentiment label:
Method 1: Extract all unique words from our tweets and cluster these words into two clusters. Which would give us the positive and negative clusters. Get the average of these two word lists and assign labels to our tweet by computing the cosine distances between each tweet and these two embeddings.

Method 2: Generate our own positive and negative words. Get the mean embedding for these two groups. And get cosine distance between each tweet and these two sentiment embeddings.

1. Prepare Embedding
Convert tweets into tokens , convert tokens into its ids, and create attention mask. The result is shown in the dataframe below:

	tweet	original_tweet	tokens	token_ids	attention_mask
0	Real Estate Market would crash if there is no ...	Real Estate Market would crash if there is no ...	[[CLS], real, estate, market, would, crash, if...	[101, 2613, 3776, 3006, 2052, 5823, 2065, 2045...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
1	Concur My company was 100 you MUST work in the...	Concur. My company was 100% "you MUST work in ...	[[CLS], con, ##cu, ##r, my, company, was, 100,...	[101, 9530, 10841, 2099, 2026, 2194, 2001, 253...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
2	Why not ask if we really need that thing I thi...	Why not ask if we really need that thing? I th...	[[CLS], why, not, ask, if, we, really, need, t...	[101, 2339, 2025, 3198, 2065, 2057, 2428, 2342...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
3	Dear Line Managers Appraisal your subordinate ...	Dear Line Managers, Appraisal your subordinate...	[[CLS], dear, line, managers, app, ##rai, ##sa...	[101, 6203, 2240, 10489, 10439, 14995, 12002, ...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
4	I have had more opportunities to work cross fu...	I have had more opportunities to work cross-fu...	[[CLS], i, have, had, more, opportunities, to,...	[101, 1045, 2031, 2018, 2062, 6695, 2000, 2147...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
...	...	...	...	...	...
2990	I am available if a remote work opportunity ar...	I am available if a remote work opportunity ar...	[[CLS], i, am, available, if, a, remote, work,...	[101, 1045, 2572, 2800, 2065, 1037, 6556, 2147...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, ...
2991	It made almost everyone work with limited reso...	It made almost everyone work with limited reso...	[[CLS], it, made, almost, everyone, work, with...	[101, 2009, 2081, 2471, 3071, 2147, 2007, 3132...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
2992	I started moving things into my new office ups...	I started moving things into my new office ups...	[[CLS], i, started, moving, things, into, my, ...	[101, 1045, 2318, 3048, 2477, 2046, 2026, 2047...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
2993	DYK that we have spent more than 5 5 billion m...	DYK that we have spent more than 5.5 billion m...	[[CLS], d, ##yk, that, we, have, spent, more, ...	[101, 1040, 15922, 2008, 2057, 2031, 2985, 206...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
2994	Are finance leaders finally loosening the rein...	Are finance leaders finally loosening the rein...	[[CLS], are, finance, leaders, finally, loosen...	[101, 2024, 5446, 4177, 2633, 29476, 2075, 199...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...

2. Get Embedding for all the words in the corpus

:

	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 × 3 columns

3. Clustering words from original tweet

I first reduce the dimension using UMAP before clustering using K-Means. These are the samples **from cluster 1:**

Real
market
crash
demand
commercial
space
hybrid
work
works
office
space
##ize
fellow
human
beings
##cu
##r
company
100
work
Office

And from cluster 0:

estate
would

remote
work
design
decades
space
worked
social
con
must
said
leaders
hired
really
hard
put
thing
anyone
privileged
businesses
got
coming
feels
teacher
learn
##rs
##ive
managers
subordinate

There is a mix of positive and negative words from each cluster. *'fever'* is negative while '*happy*' is positive for example, but they are both present in cluster 0. It's not very clear which one is negative and which one is the positive cluster. This could be investigated by computing the top 15 words that are closest to each of the centroids. But I decided to skip this part and assume that the clusters are correctly grouped by sentiments.

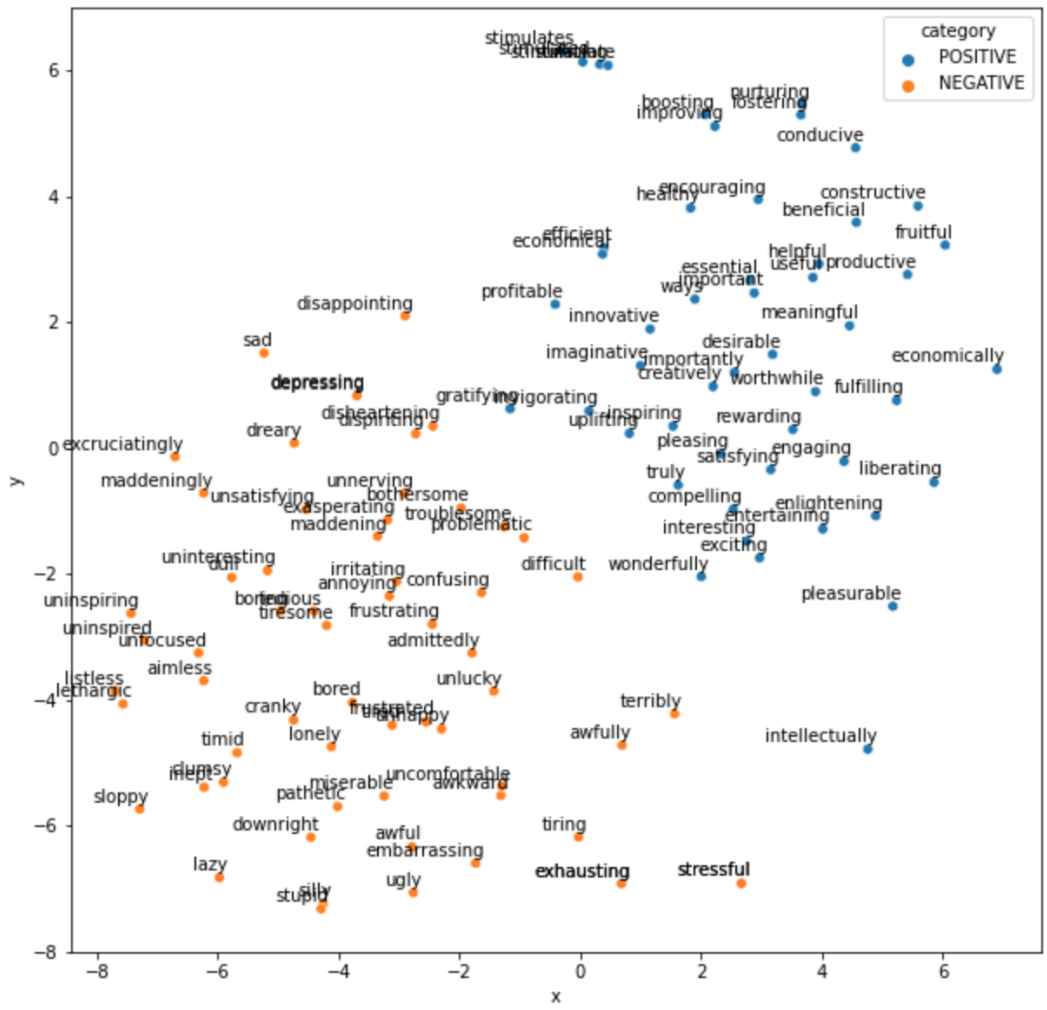
4. Clustering our own generated words

We will use gensim to create around 60 words that will roughly create 2 separate word clusters. I did trial and error methods in choosing the first 3 words for each of the sentiments before finally getting a decent cluster that is reasonably separated from each other:

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

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

['dull', 'confusing', 'bored', 'depressing', 'pathetic', 'lazy', 'listless', 'tedious', 'exhausting', 'frustrated', 'unsatisfyin
g', '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', 'sloppy', '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']
```



5. BERT Prediction

We now have three data frames that contain embedding to compute cosine distance to generate label predictions. We will have two labels for each tweet, one label from our own generated words, and another label using words from the tweet data itself.

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...

```
]: 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...

These are the results including our Vader prediction result earlier:

	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	Study reveals growing cybersecurity risks driv...	[-1.0207954645, -0.6177105904, 0.7027013302, -...	NEGATIVE	POSITIVE	NEGATIVE
2	Learn the top business functions to outsource ...	[0.3671964109, 0.4819124937, 0.3259748816, -1....	NEGATIVE	POSITIVE	POSITIVE
3	Wrapping-up remote work for the day... just go...	[-0.4098833501, -0.5858969688, 0.9731462598, -...	NEGATIVE	NEGATIVE	POSITIVE
4	Yup. And seeing an article yesterday trying to...	[0.1640227735, -0.0528694689, 0.1988344491, -0...	NEGATIVE	NEGATIVE	POSITIVE
...	...	...	...	...	...
677	Freelancing is the answer to being a slave to ...	[0.0466130301, -0.9194151759, -0.1806280911, -...	NEGATIVE	NEGATIVE	POSITIVE
678	Effective communication is key when it comes t...	[-0.6447560191, -0.1380660683, 0.3176239431000...	POSITIVE	POSITIVE	POSITIVE
679	in the room we are putting the TV. Samsung wil...	[-0.1216553077, -0.2701974809, 0.0527218617000...	POSITIVE	NEGATIVE	POSITIVE
680	do you ever have to wack the sky remote to get...	[0.0159683842, -0.3788262606, 0.1322553605, -0...	POSITIVE	POSITIVE	NEUTRAL

These are some samples where all of our three models give the same prediction:

So are you telling me I should leave my eight-year-old at home by his se I have no babysitter he is remote learning and I can not go back to work so are you telling me I do not count what I need does not matter what my son need does not matter safety is first what is wrong with you all

I had a supervisor who hated that we had to work from home. He wanted us all to get hired full time by the state so we could have office parties. I reminded him that some of us could never have that happen. I live hours away from "the office " and without remote work I am home

I do not know why our company can not attract more diverse talent." #RemoteWork #WorkFromAnywhere #TalentAcquisition #Recruitment #DEI #HR

if you do not trust your people to get the job done from home just say that. if remote/hybrid work models make it harder for you to ignore that your employees are people with other (dare i say higher) priorities than this job just say that.

Think reading about dickhead bosses insisting on 8 hour surveillance of staff via webcam permanently put me off the idea tbh. If you can not trust your staff to work unsupervised you have got bigger problems than remote working imho.

Commercial real estate will fight remote work trend tooth and nail. They can not afford allowing remote work to become a thing.

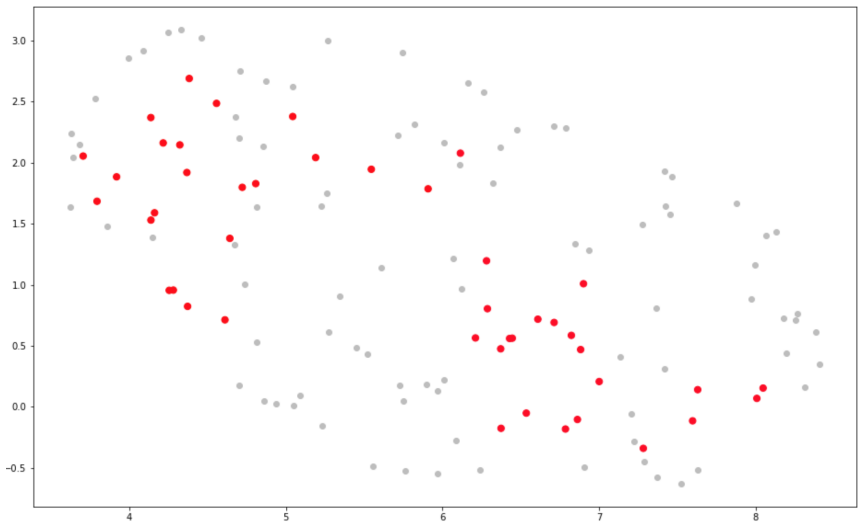
There are many negative tweet about remote work but are actually advocates of remote work. The tweets are negative because they are criticising companies who don't allow them to do remote work

6. Topic Modelling

Using a modified TF-IDF, I use our sentence embedding to extract topics from our corpus. The details of how the TF-IDF can be found here https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6.

I only take the tweets where Vader prediction and own words predictions are negative since i only want to know why do people think negatively about remote work.

Again, we do dimensional reduction to our embeddings and cluster them using HDBSCAN. It gives us 2 clusters. The third (grey) cluster is an outlier.



Here is an excerpt that describes the intuition in using TF-IDF in finding the topics of each cluster:

> *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.*

Here is the resulting topic keywords from the two cluster

```
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)]}
```

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.

7. 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.