Sentiment Analysis of Twitter Discourse in the Context of UN Climate Conferences

James Hughes, Ameer A Saleem, Vignesh Sridhar.
Supervised by Janique Krasnowska and Fernando A Zepeda

March 5, 2023

Abstract

The following document entails a student-led end-to-end data science project, the goal of which being to investigate the changes in perceptions towards climate change over time. The project involved a review of the current literature and available data, sentiment analysis, neural network model training and evaluation of the developed model, as well as projections for the future.

Contents

1 Introduction

Since their inception, social media platforms have been analysed to produce insights into public opinion using increasingly complex natural language processing (NLP) approaches. Earlier models such as VADER [4] relied on rule-based linguistic techniques to capture how grammatical features such as adverbs and conjunctions affected the intensity and negation of sentiment within a text. This approach matched the performance of individual humans on consensus-based sentiment labelling in a variety of online media contexts, and has been used to analyse the sentiment of tweets expressing opinions about climate change[9]. In the early 2010s, word-embedding vectorisation became a widespread technique for semantic modelling, with the release of Word2Vec [5] and GloVe [7] data. This enables the study of lexical semantic relationships on a large scale; in 2017 researchers used this approach to systematically explore sub-topics within climate change discourse on Twitter[6].

In this way, machine learning represents a novel and emerging paradigm in climate change discourse analysis. It has brought new understanding to the shifting landscape of public discussion around climate change by assisting with topic classification [1], sentiment analysis and user-community clustering [2], especially through analysing social media platforms such as Twitter. The insights produced by these tasks can complement social science research. Indeed, Stede and Patz [10] highlight the increasing overlap of the two fields in the analysis of climate change discourse and emphasise the need for considerations around representativeness and validity of modelling when NLP is used in this way.

We discussed many different avenues for research and project focus initially. One of our ideas revolved around analysing pledges made by political leaders and conduct some data-backed investigation into how much they keep to their word. There is plenty of data available, which can be investigated in many different scopes (e.g. national vs global, a few select politics vs advocates for a particular party, etc). However, the language used by such politicians is constrained, which could make it difficult to train a model. Also, it would be difficult to associate a metric with keeping promises, as this could be subjective and it is unlikely that a database encapsulating this with objective labels exists. The scope of individuals to investigate is also somewhat constrained, regardless of the aforementioned scope. Another idea was to investigate the changes of attitudes of the public towards climate change following different climate summits. This was a genuinely interesting question for us, the implications of which could be used to forecast how attention on climate change could evolve through the years. However, this was a very ambitious scope for a student-led project- the title required refining in order to give us a well-defined problem to investigate.

After considering the benefits and drawbacks of each of the project ideas

we eventually settled on the key questions outlined below.

Twitter is a widely-used social media platform which can be used to sample the public's reaction to international events. Our study used historical data of Tweets occurring across the courses of multiple COP conferences from COP24 to COP26. We used a human-labelled dataset [8] to construct a model to classify the sentiment of a Tweet, limited to *negative*, *neutral*, or *positive*, which we then used to label our original dataset. Additionally, we manually categorised the Twitter accounts in our original data - limited only to the top 100 users ordered by the maximum likes of any of their tweets in the dataset. Our analysis was then centered around the following key questions:

1. Who are the most influential stakeholders in the Twitter discourse that surrounds the COP events?

2. Is there a relationship between the sentiment of a tweet and how much exposure it receives?

3. How does this relationship vary across types of users and over time?

From here, we conducted a literature review to ascertain which of these avenues would be the most fruitful to pursue.

2 Preliminary Information

2.1 Labelled Sentiment Data

In our study we used data from the 2017 International Workshop on Semantic Evaluation [8]. The annual workshop releases manually labelled data and welcomes a range of teams from around the world to create accurate classifier models, and compares their results. The data we used for our model training was from English Subtask A, which received submissions from 38 teams. The workshop demonstrates some of the widespread NLP techniques used for sentiment analysis, including deep learning methods such as convolutional and recurrent neural networks, as well as common machine learning approaches such as logistic regression, decision trees, and support vector machines.

2.2 Scraping Tweets

While Twitter has an API for scraping tweets, we settled on the use of snscrape, which appeared easier to implement.
Twitter is a prevalent platform with immense usage, meaning that scraping within an arbitrary timeframe will yield a lot of irrelevant data. It makes reasonable sense to focus on time intervals before, during and after COP

summits; this is the time when most 'notable figures' would be discussing climate change and voicing their respective opinions on it the most. This also extends to the public, providing us a greater density of climate-related data.

2.3 COP

COP, standing for 'Conference of the Parties', is the main decision-making body regarding any major decisions regarding climate change. COP is organised by the United Nations and involves almost every country in the world. The goal of such summits is to discuss what major governmental authorities can achieve in the fight against climate change. COP usually take place every year and aims to review the global progress made in mitigating the impact of climate change. The first COP summit took place in Berlin in March 1995. At the time of writing, the most recent COP summit, COP 27, took place in Egypt in November 2022.

2.4 Word Embeddings

In order to get a computer to understand the words in each tweet it is necessary to first convert them into vectors in higher dimensional space. The idea is to give words with similar meaning/context a similar vector value. For example, we may want to give the words 'lion' and 'leopard' similar vector representations in a bank of words containing the names of all the animals in the world. There are a few choices as to how this can be achieved, such as Word2Vec or GloVe's pre-trained word embeddings. We went with the latter for this project. The two methods of word embedding would most likely yield very similar results for us, though they work in different ways. Word2Vec utilises a feedforward neural network. GloVe uses a matrix-based approach: it starts by constructing a matrix whose rows are words and columns are contexts it has identified in the input word bank. This matrix can then be decomposed into smaller matrices that are easier to work with (e.g. from word x context to word x feature and feature x context).

Once the words of a dataset have been word embedded, we can implement the clustering algorithm to have the computer try and identify patterns in the words for us. For the sake of this project, we investigated clustering with different numbers of clusters in the EDA portion of the project.

We encountered another potential limitation when manually categorising the data in our scraped tweets. A lot of tweets were indeed related to climate change, but sometimes they discussed an individual rather than the topic of climate change itself, for example praising Greta Thunberg for her efforts in this space. This is not directly related to our goal of attributing sentiment labels to these tweets *with regards to stance on climate change.*

2.5 Clustering Techniques

We used Centroid-based clustering. These types of algorithm work by assigning the optimal location for the center of each cluster, known as the *cluster centroid*). The position of each centroid is updated by finding the average of the vectors that belong to that cluster, terminating once a prespecified tolerance has been reached. The K-Means algorithm is the most popular centroid-based clustering technique.

3 The Main Project

3.1 Main Twitter Datasets

We scraped historical tweet data, with the queries detailed in Table 2. For the datasets relating to each conference we limited the query to English Tweets containing the corresponding string "COP2_", which is case insensitive and includes tweets with the substring "#COP24", for instance. Each tweet was collected along with important meta-data such as information about the author account, and metrics such as number of likes and replies.

Next we compiled the users authoring tweets among the most liked tweets in the dataset, and manually categorised them according to nine stakeholder categories show in Table 1. We then filtered each of the original datasets to contain only tweets from these categorised users. Lastly, we added the sentiment predictions from our trained model.

3.2 Implementation

We began by focusing our attention on dates close to/during COP summits. We scraped tweets containing the string 'COP–' within time frames of a few weeks at a time during periods of high COP discussion and activity, yielding tens of thousands of tweets in each scrape. Our goal at this stage was to explore patterns at both tweet and word level.
We started by accumulating all the tweet data in each scrape and clustering at the word level, with both K-means clustering and agglomerative clustering. The most interesting insights were as follows:

- The most frequently used words in each scrape were what are called *stop words*. In the context of NLP, these are words that carry very little information. Some examples are shown in Figures 1a and 1b. This was something we would have to take care of during data preprocessing. Note that we didn't remove phrases such as "#COP..." since their abundance in the data was a consequence of our filtering criteria.

- For K=__, the K-Means clustering algorithm was able to independently group together tweets pertaining to different categories of *people*, i.e.

Category	Description	Examples
Activist	Accounts belonging to activists or charitable organisations working to help climate change mitigation or adaption.	@GretaThunberg @SumakHelena
Business	Accounts related to a business or personal accounts of business leaders.	@BNPParibas @BoschGlobal
Celebrity	Personal accounts for which no other category applies and with more than 10,000 followers.	@LeoDiCaprio
International Organisation	Official accounts for non-commercial international organisations.	@GreenpeaceNZ @UNFCCC
Journalist	Personal accounts of news reporters and commentators.	@katieworth @LeoHickman
News	Official accounts for large news and media outlets, including digital news websites.	@Reuters @AJEnglish
Politician	Official accounts for individuals in (or formerly in) government positions or international political organisations.	@NicolaSturgeon @Justin Trudeau
Scientist	Accounts related to research organisations or personal accounts for people whose primary occupation is research or public science discourse.	@Astro_Alex @Peters_Glen
Misc	None of the categories above apply.	@Damo_Mullen @mamaloe66

Table 1: The nine user categories and their descriptions.

separate clusters related to politics, climate change activism and environmental science. This gave us the idea to categorise the tweets by the occupation of the user (or their relation to the topic of climate change).

	Word	Count
10	the	5138
66	to	3545
23	#cop24	2869
34	and	2464
12	of	2058
21	at	1690
42	in	1573
47	a	1491
56	climate	1301
26	for	1270

(a) COP24

	Word	Count
20	the	3959
7	to	2510
51	#cop25	2208
11	and	2054
75	of	1612
22	in	1380
59	at	1142
126	a	1062
0	for	1030
54	is	1012

(b) COP25

Figure 1: The top 10 most common words in the COP24 and COP25 tweet scrapes.

3.3 The Model

We implemented a neural network model in Tensorflow to learn the patterns of label associations with the tweets.

In order to train the model, we required a dataset with sentiment labels attributed to each tweet. This could be done by a computer with clustering, but defeats the purpose of using the dataset for training, since we would have no feasible way of validating so many rows of labelled data.

We decided to use the publicly available data from the 2017 International Workshop on Semantic Evaluation, in particular from English Subtask A which included roughly 60,000 human-labelled tweets, with 3 labels corresponding to *negative*, *neutral*, or *positive* [8]. The labels were produced based on annotations from at least 5 judges, and strict quality control measures excluded annotations from judges who failed a certain threshold number of hidden tests. We decided to retain 2,000 tweets for testing purposes after the hyper-parameter tuning.

Among the tweets used for model-training, there was a clear class imbalance, with 46.0% of tweets being labelled neutral, 34.9% as positive, and 19.1% as negative. Therefore we proposed a suitable baseline accuracy as 46.0%, which is that of a dummy modal-class-predictor model.

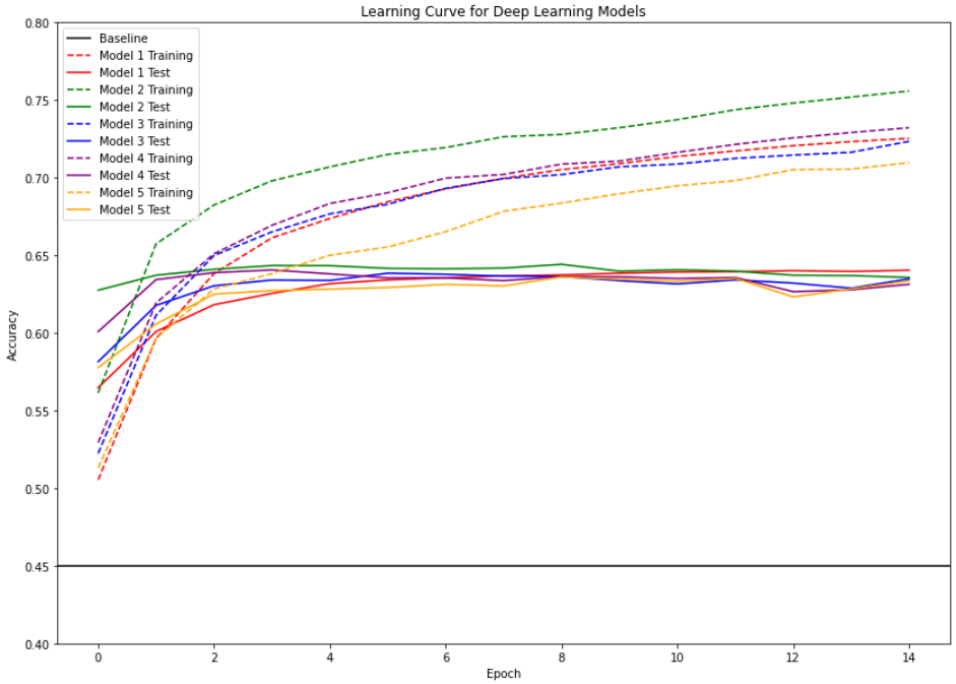

Figure 2: A plot of the training and testing accuracies of the five neural network models.

```
# Compile model.
embedding_dim = 16

model_1 = tf.keras.Sequential([
    tf.keras.layers.Embedding(vocab_size, embedding_dim),
    tf.keras.layers.GlobalAveragePooling1D(),
    tf.keras.layers.Dense(3, activation="softmax")
])

model_1.compile(optimizer='RMSProp', loss='categorical_crossentropy', metrics='accuracy')
```

Figure 3: Model 1 architecure.

We implemented and trained five neural network models of various complexity using TensorFlow on the English tweets sentiment dataset. The testing and training accuracies are shown in Figure 2. From this, we opted with our first model, consisting of architecture shown in Figure 3. Our final model, the simplest in the set of five, is comprised of the following layers:

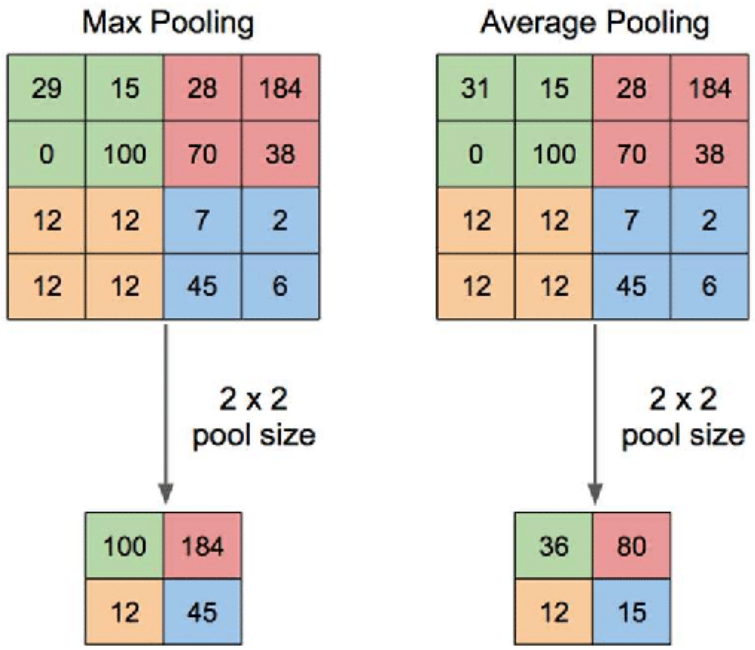

Figure 4: Illustrations of common pooling methods. [Source: ResearchGate.]

- **Embedding layer**: similar to the GloVe package, this Keras layer takes our set of words and embeds them as vectors in n-space. We have specified the input dimension size (the total number of words in our dataset) and the output dimension size (the dimension of space we want to embed our word vectors into). After some experimentation, we found the output space dimension of 16 to be a good fit.

- **Pooling layer**: pooling methods aim to reduce the size of the input space by replacing groups of points with single points formed by aggregating the groups' values. The two most common types of pooling methods: max pooling, in which the maximum value of a group of points is taken as the value of the new aggregated point; and average pooling, where the average value is taken instead. In general, max pooling is able to highlight stark contrasts in a dataset, whereas average pooling is better for 'smooothening out' the values in the entirety of the dataset. We opted with average pooling for our chosen model.

The other four models consisted of the above layers as well as the following:

- **Convolution layer**: a convolution is a filter applied to an input, with multiple filters constituting a feature map. Convolution layers are often used in tandem with pooling layers: convolutions create multiple feature channels in parallel; and the pooling layer reduces the dimension of each of these channels.
- **Dropout layer**: this layer randomly sets activation values to 0, temporarily 'dropping out' these nodes from the network. The main purpose of dropout layers is to help prevent overfitting. The utilisation of this kind of layer is especially useful when training a large neural network on a relatively small dataset.

We then used the test data to measure the performance of the trained model. The unfiltered confusion matrix for our final model is shown in Figure 5. The confusion matrix for a model with 100% accuracy would be a diagonal matrix. Our model outputs a vector $\{p_1, p_2, p_3\}$ with $p_1 + p_2 + p_3 = 1$, analagous to the network's confidence that the input tweet belongs to each of the three classes. We then return the index of the largest p_i, so for example if the output for a particular tweet is $\{0.1, 0.4, 0.5\}$, then we take the sentiment label to be 0 (i.e. negative. This gives us two different ways of evaluating our model: restricting our view to the final decision made by the model, due to selecting the index of the node with largest p_i; or by filtering the results based on the network's confidence in its classification. The confusion matrix for the model's classifications with confidence greater than 80% is shown in Figure 6, i.e. for these 665 tweets, $\max_i p_i > 0.8$. This added filter excludes 66.75% of the tweets from the original test set.

Unfiltered confusion matrix

	Predicted Negative	Predicted Neutral	Predicted Positive	Accuracy	False Signal Rate
Actual Negative	122	147	43	0.391026	0.137821
Actual Neutral	60	632	177	0.727273	0.272727
Actual Positive	28	242	549	0.670330	0.148962

Total sample size: 2000
Total accuracy: 0.6515

Figure 5: Unfiltered confusion matrix

It is worth noting that according to Table 6 in [8], our model would have been ranked at 5th place in terms of accuracy (65.15%) compared to the models produced by the 38 teams at the time of the workshop. However, the teams were constructing models to optimise average recall across the classes, a slightly different performance metric. In addition, adding further complexity to increase the accuracy according to the test dataset may have

Confusion matrix; confidence > 0.8

	Predicted Negative	Predicted Neutral	Predicted Positive	Accuracy	False Signal Rate
Actual Negative	53	22	7	0.646341	0.085366
Actual Neutral	11	189	32	0.814655	0.185345
Actual Positive	3	37	311	0.886040	0.150997

Total sample size: 665
Total accuracy: 0.8315789473684211

Figure 6: Confusion matrix showing tweets classified with at least 80% accuracy.

-	COP24	COP25	COP26
Year	2018	2019	2021
COP Date Range	Sunday 2nd Dec to Friday 14th Dec	Monday 2nd Dec to Friday 13th Dec	Sunday 31st Oct to Friday 12th Nov
Query Date Range	15th Nov to 30th Dec	15th Nov to 27th Dec	14th Oct to 27th Nov
Total Tweets	108405	152662	981982
User-Categorised Tweets	3551	2560	3489

Table 2: The number of tweets scraped across the three conferences.

reduced our model's ability to generalise to the twitter data we wished to study later on.

3.4 Analysis

The first part of our analysis concerns treating the datasets as time series. Since our final data collection spanned three separate COP conferences, we decided to use study the profiles of tweet frequency and number of likes over time by averaging these trends over the different conferences. In order to fairly align the different conference dates (see Table 2) we decided to encode the dates for each conference as increasing integers, standardised so that the end date in Table 2 is encoded as zero in each case. In the charts that follow, we annotate the start date as code -12, and it should be noted that this encodes the official start date of two of the conferences, and represents a Sunday.

To create figure 7 we aggregated the total number of tweets for each date and then divided by the total number of tweets, for each conference. We then took a simple average of these three frequency profiles to produce

Conference	Start Date	End Date
COP24 (‘18)	Sunday 2 December	Friday 14 December
COP25 (‘19)	Monday 2 December	Friday 13 December
COP26 (‘21)	Sunday 31 October	Friday 12 November

Table 3: The official start and end dates for each conference[11][12][13].

the plot show. As expected, we can see that the greatest volume of tweets occurs during the conference, with this volume rapidly decaying in either time direction before and after the event. Additionally we see two spikes in volume corresponding to the start and finish of the conference; the latter being the sharper fluctuation, on average. Figure 8 shows in a similar fashion that the mean likes per tweet is greatest during the conference itself. However there is much more fluctuation in the mean likes, especially as the conference progresses and after it ends.

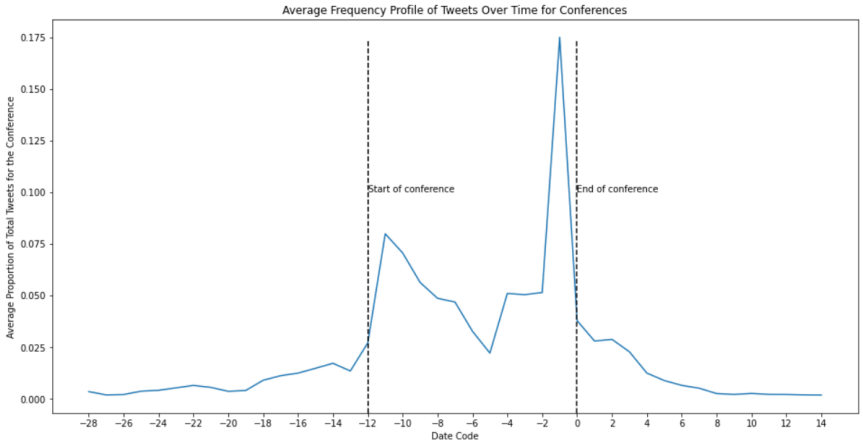

Figure 7: Average standardised frequency profile over time.

In Figure 9 we break the frequency profiles down into sentiment classes. The three datasets are separately standardised to produce a proportion of tweets from each class on each date, and then combined via simple average to create the first plot. The graph appears to show that the proportion of positive and negative tweets remains fairly constant around 12% for before and during most of the conferences. The proportion of positive tweets does not seem to diverge from this significantly following the end of the conference. Negative tweets, however, seem to become more prevalent towards the end of the conferences, spiking rapidly around the final conference date, and then remaining at around 16% in the two weeks following.

Figure 10 seems to show that the mean likes for tweets of all sentiment classes increase steadily leading up to the conferences in general. Remarkably, negative tweets seem to receive many more likes on average around the

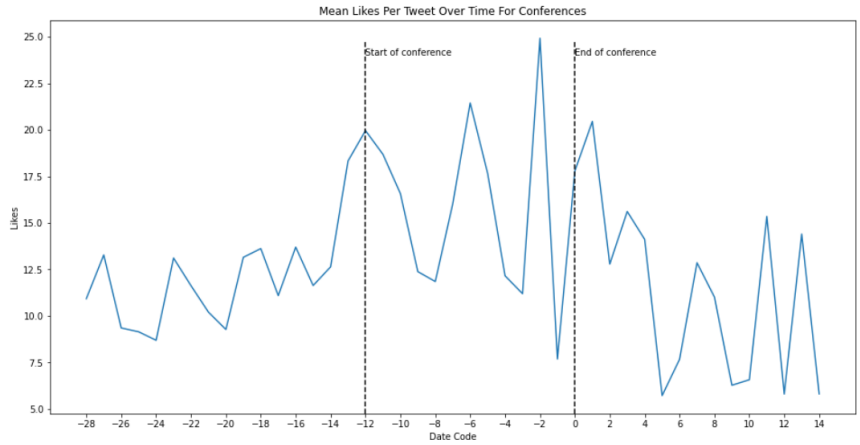

Figure 8: Mean likes per tweet over time for all conferences.

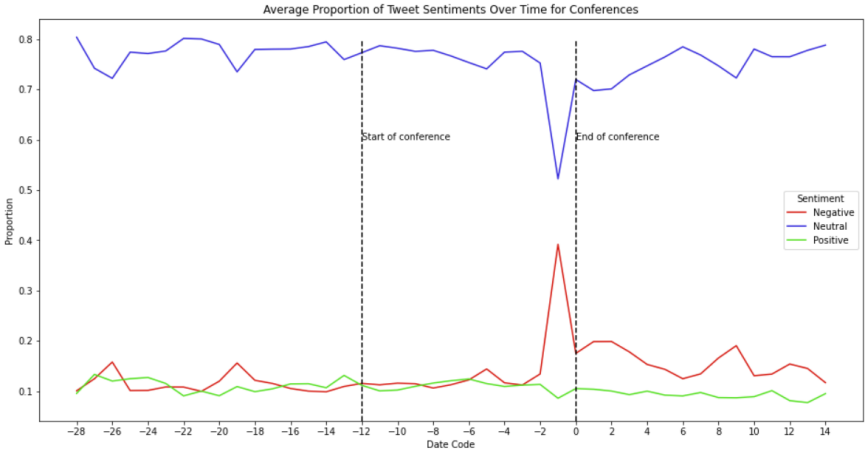

Figure 9: Average relative frequency of tweets in each sentiment class over time.

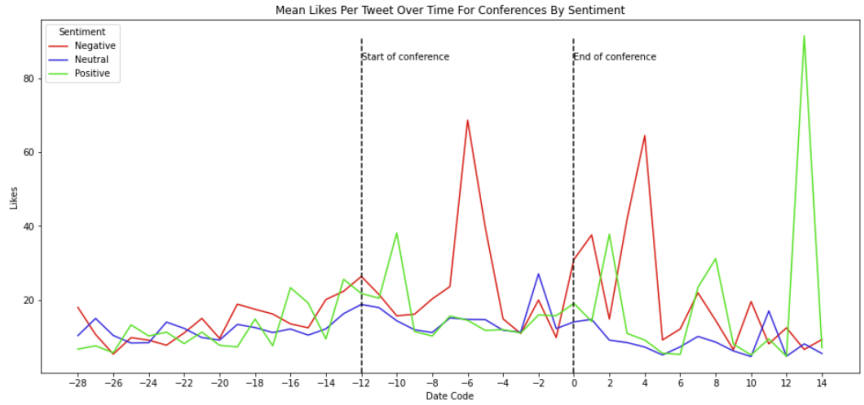

Figure 10: Mean likes per tweet over time for all conferences, by sentiment.

middle of the conference, while the same is not particularly true for positive tweets. In the two weeks following the conference, the mean likes received by positive and negative tweets appears to fluctuate heavily but can be seen to increase substantially on particular days.

We then restricted our focus to user-category labelled data. Compared to the original class imbalance of the training data, Figure 11 indicates a prevailing negative sentiment in the data, which shows that there is a clear overall signal of this sentiment throughout the data, even if some of the predictions are subject to noise. Categories such as scientists, activists, business and news show this trend more strongly. Some of the sentiment compositions are surprising; business accounts appeared to produce far fewer positive tweets than negative. Figure 13, shows a sample of 'business' tweets that were labelled by our model as 'negative'. The sample shows that business-labelled users often publish tweets which are completely unrelated to their own business activity, and instead discuss world events and politics brought up by the COP conferences, as opposed to marketing campaigns or positive stories of their actions in the area of sustainability.

Conversely, politicians showed a stronger signal of positive sentiment. This aligns with our intuition, as politicians will want to establish their agendas in a positive manner, and the Twitter platform is one of the best to advocate this. A small sample of these tweets are shownn in Figure 14.

Figure 15 yields insights into the likes across the various groups in the following ways:

- Politicians, news, and international organisation tweets don't appear to show much variance across the sentiment classes.
- Activists and scientists, showed polarity with negative and positive tweets both receiving more likes than neutral tweets.
- Journalist and business accounts receive more likes for tweets which are negative and less for positive tweets. This may indicate that the audiences for these accounts favour negative content such as those invoking controversy or outrage than in general, when it comes to the topic of climate change.
- Celebrities tweet more negatively than positively, yet their positive tweets receive more likes on average.

4 Conclusion

4.1 Results

We have identified some interesting patterns relating to the most influential stakeholders in the discussion of the COP climate conferences. In our study,

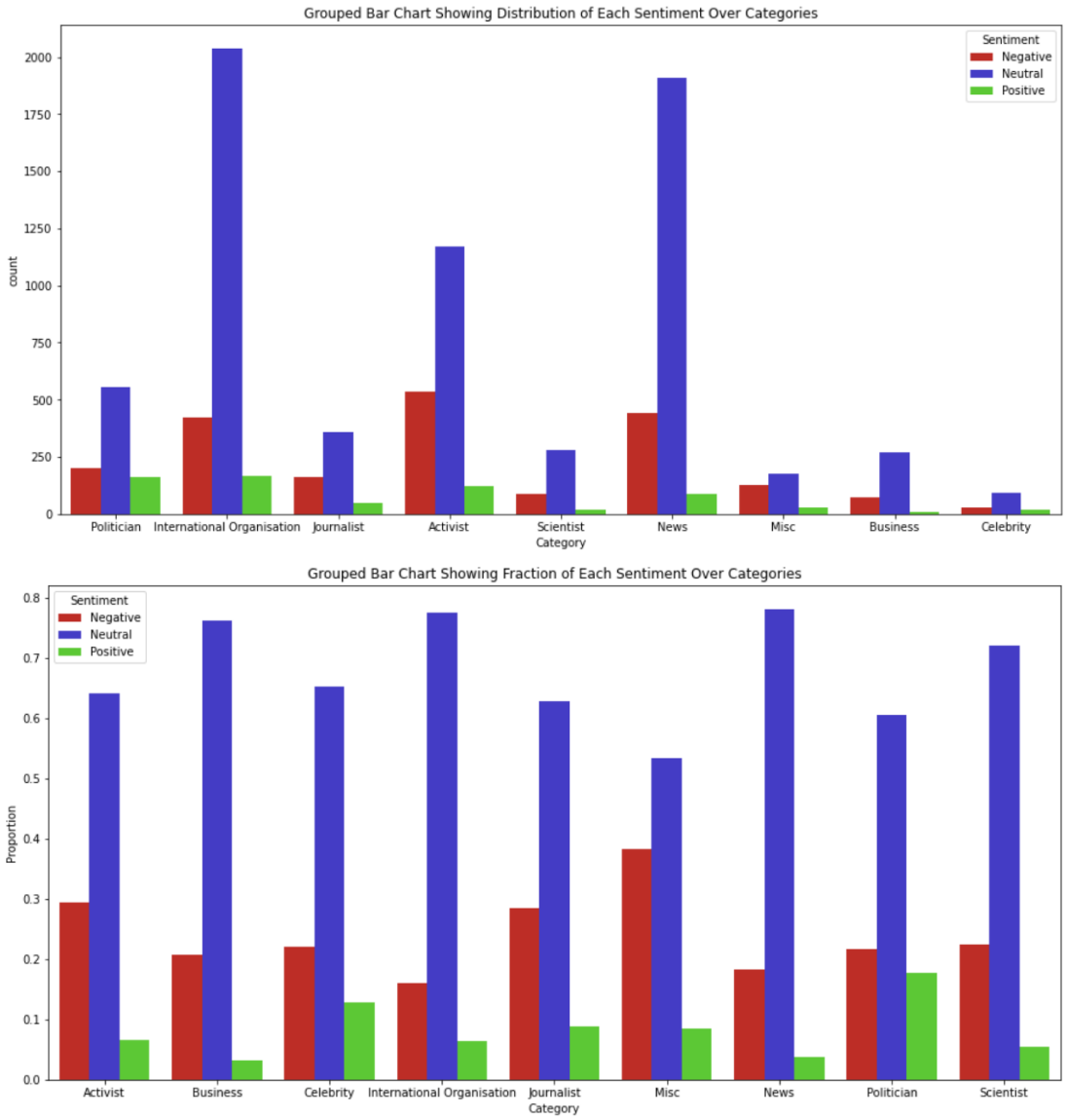

Figure 11: Bar charts depicting the proportions of tweet sentiment per category across the three COP conferences.

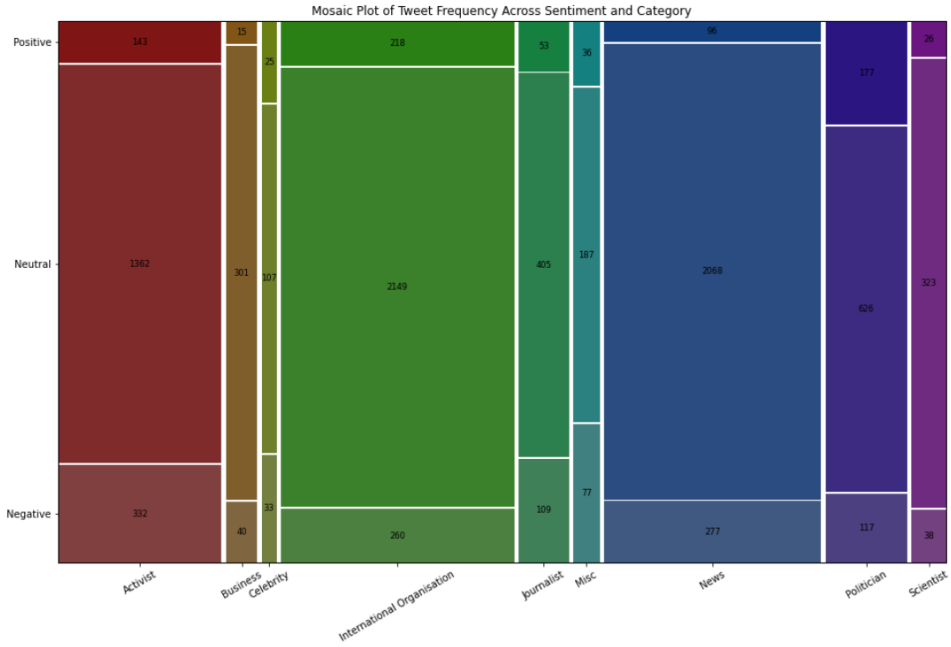

Figure 12: Mosaic Plot of Frequency by Sentiment and User Category.

Content	Category	Sentiment
"CO2 emissions aren't reducing, they are in fact increasing," @GretaThunberg said at #COP25. "We have achieved nothing" https://t.co/ilVonuTLeW	Business	Negative
We're cutting our #CarbonFootprint by 50% by 2030. Our plan is backed by #science & endorsed by @sciencetargets. Find out more: https://t.co/1TymAOkYkv #SYNnovation #ClimateChange #COP25 https://t.co/Y8yEbc46Eu	Business	Negative
At Cop25, a few nations - Brazil, the US, Saudi Arabia and Australia in particular - were emboldened as never before to stand against the world and nakedly try to weaken efforts to tackle climate change to benefit their short term interests.	Business	Negative
"Indigenous blood, not one more drop."\n\nBrazil's indigenous people rallied outside #COP25 in Madrid calling for accountability over the killing of 2 Guajajara leaders. The Guajajara are known for its forest guardians who protect land against illegal deforestation https://t.co/ktrRwJLwmK	Business	Negative
Brazil's indigenous people rallied outside #COP25 in Madrid calling for accountability over the killing of 2 Guajajara leaders. The Guajajara are known for its forest guardians who protect land against illegal deforestation https://t.co/nK4wTxF6pb	Business	Negative

Figure 13: A sample of tweets that were labelled as 'negative' by our model in the 'business' category.

Content	Category	Sentiment
Beware greenwash from #COP26. \n\nThere has been some progress but this piece shows there is spin and reality. Halving global emissions this decade is what really matters to keep 1.5 alive.\n\nhttps://t.co/ZibqOboad4	Politician	Positive
Huge thanks @LeoDiCaprio for creating awareness on @UNEP's #EmissionsGap Report. We appreciate your voice & #environmental leadership. At #COP25 we must all commit to #ClimateActionNow. https://t.co/8TGLQfFVS1	Politician	Positive
Canada and the European Union are incredible partners, and we know we can accomplish more when we work together. Thanks for the invite, @MAC_Europa. Awesome chatting with you, @ElliKoestinger, and everyone who attended. #COP24 https://t.co/YtwPged6XY	Politician	Positive
Day three has begun, and the #COP24 negotiations are continuing! Today, I'm having more meetings with our Umbrella Group and other major countries. Together, we'll talk about moving forward and landing the #ParisAgreement rulebook. https://t.co/sVe7ahzxe5	Politician	Positive
Ahead of @COP26 next week, I touched base with Premier @JJHorgan to speak about BC's new and ambitious climate plan. The province continues to be an important partner in the fight against climate change - and we'll keep making progress together. https://t.co/y8tMvknBQC	Politician	Positive

Figure 14: A sample of tweets that were labelled as 'positive' by our model in the 'politician' category.

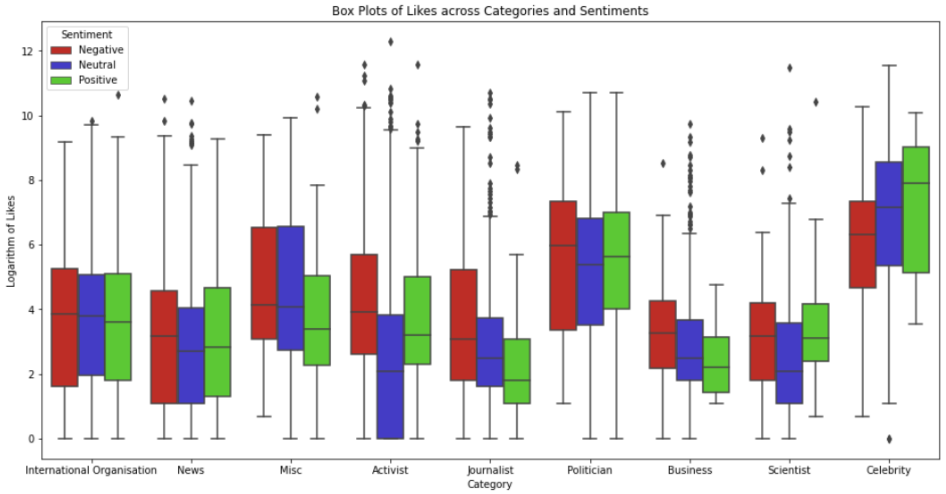

Figure 15: Box plots depicting the (logarithm of) likes, grouped by category and split by sentiment label.

we used the number of likes as a metric of the exposure attained by a tweet, and therefore in some sense an indicator for the influence the tweet had in virtual discussions of the COP conferences.

Using this methodology we managed to examine the most influential stakeholders in the discourse by looking at the accounts publishing the tweets which received the greatest number of likes within the whole dataset. We compiled the maximum number of likes garnered by each account's tweets and then put those accounts in the top 100 by this measure into nine distinct categories by manual labelling. We then examined the volume of tweets produced by each category and their distributions of number of likes. This part of the analysis – and the later analysis involving the categories – could be improved by further manual labelling so that the distribution of tweets and likes across the categories is more reliable.

Our two-way factor analysis of the effects of user-category and sentiment on the number of likes showed how the influence of a tweet may depend on these two factors. Of particular interest was the interaction between these two effects; the degree to which sentiment has a strong effect appears to depend on the category of the user, and the nature of this interaction depends on the category as well. We saw that multiple categories such as activists and scientists had increased likes for tweets which had positive or negative sentiment, as opposed to neutral, which may show an inclination towards emotionally-charged discussions in these circles.

The later part of the analysis showed evidence that the volume and influence of negative sentiment tweets may have an upward trend over the course of the conferences. Further modelling beyond the simple linear regression could be used to investigate these trends further, as well as more data. Col-

lecting data from a larger number of COP conferences would enable us to reliably answer the third question of our investigation, namely studying the long-term changes across the annual conferences, rather than just within each conference.

4.2 Representativeness of Data

Our data collection method was very straightforward with a search query of tweets containing the string "COP2_" with the last character corresponding to each of the conference titles. We collected Tweets before, during, and after the conference period. This approach mirrors that adopted by some other researchers in this area, for instance in a recent Nature article about focusing on polarization [3]. As Falkenberg et al. point out, this obscures Tweets which formed the wider discourse surrounding the events, including adjacent discussions about people, policies, and issues relevant to the conferences. However, this method was deemed appropriate for the scope of our study. Stede and Patz [10] point out that few studies in this area use more sophisticated queries, and that a simple query such as ours is mainly problematic if it contains an inherent bias which affects the aspect of the data to be measured, such as a politically-charged search term. An extended search query would immediately cause the data to decrease in topicality; even using the term "COP" would introduce a large number of unrelated tweets.

Our data is also limited to English language Tweets, which in this case causes a large degree of obscurity due to the international nature of the conferences. Moreover, there is a clear case to be made that the patterns in the sentiments of tweets related to the conferences would be affected by the nationality of the tweet author, for instance in the case of unprecedented natural disasters around the time of the conference.

4.3 Validity of Our Model

The reliability of our analysis hinges on the accuracy and interpretation of our model predictions. Our model was trained on a large dataset of around 50,000 English tweets, with reliable human labels. The model achieved a reasonable accuracy — and it should be noted that there are limits to how accurate such a model can become anyway, due to human-judge disagreements.

As the training data comes from the same social media platform as the data for our study this endows the model with a degree of transferrability, although it could be argued that the sentiment in the discourse around climate change and politics is a task with a high degree of topicality. The natural language features constituting, say, a positive tweet on the platform in general may differ from what constitutes a positive tweet in the specific discourse around the COP conferences.

There is also a deeper ambiguity in the concept of sentiment itself which should be addressed. For instance, consider the following artificial tweet:

> 'I'm so happy that everyone is working so hard to solve the issue of climate change', said no-one ever. [1]

The tweet indicates a negative sentiment with regards to climate change. However, the use of the word 'happy' could potentially cause our model to classify this as a tweet with positive sentiment. Not only would this tweet reduce the accuracy of our network, but the tweet itself could be argued as inherently destructive to the model training process; it may suggest to the network that the word 'happy' ought to be synonymised with negative sentiment rather than positive (which it ought to in isolated context). Sarcasm of the form shown above, as well as other problematic language features such as irony, humour, satire, and exaggeration, make the task of defining sentiment itself complicated, before even building a model to predict it. We made up 10 tweets [2] to investigate the difficulties our model may have encountered in this regard during the training process, as illustrated in Figure 16. The model managed to classify only 2 of the 10 tweets correctly. Unsurprisingly, the model seems very confident in its classification of most of the sarcastic tweets, despite its evaluations being completely incorrect. Additionally, the tweets accumulating the lowest confidence value were signed the neutral sentiment label. This is most likely due to the class imbalance in our training set: in other words, when the model is not confident, it defaults to labelling the tweet as neutral.

	Content	Predicted Negative	Predicted Neutral	Predicted Positive	Predicted	Confidence	Actual
0	"I'm so happy that everyone is working so hard to solve the issue of climate change", said no-one ever. #sarcasm\n	0.094003	0.094084	0.811912	2	0.811912	0
1	It's been great to see so many politicians working so hard at achieving nothing with regards to climate change. Keep it up! #sarcasm\n	0.172700	0.026435	0.800865	2	0.800865	0
2	Coal and gas may harm our planet, but green power cannot hurt me. Go green! #climate\n	0.841626	0.145500	0.012874	0	0.841626	2
3	Most climate change activists will campaign strongly for change, yet will still use electronics and power-intensive devices. Sure, let's listen to them!\n	0.555325	0.402044	0.042631	0	0.555325	0
4	Why did the polar bear refuse to go the party? Because he heard that the ice caps were melting! Haha so funny... #sarcasm\n	0.035278	0.099813	0.864909	2	0.864909	0
5	I believe that we will never solve the issue of climate change. Actually, that is not what I believe.\n	0.697762	0.228066	0.074173	0	0.697762	2
6	Although the issue of climate change is a challenging one, it's not all doom and gloom. Progress is in the making!\n	0.587708	0.319563	0.092730	0	0.587708	2
7	The only thing sustainable in this world is my ability to sustain a fake smile, with the knowledge that nature is perishing. #climatechange #savetheearth\n	0.684905	0.153612	0.161483	0	0.684905	0
8	A lot is uncertain in this world. But one thing I am certain of is that we will be able to eradicate the horrific effects of climate change #wecandothis\n	0.286373	0.502397	0.211230	1	0.502397	2
9	2,4,6,8, who do we appreciate? Climate change activists!	0.290256	0.439584	0.270160	1	0.439584	2

Figure 16: A sample of made-up tweets designed to test the model's ability to handle ambiguous sentiments in language.

[1] Note that this made-up sentence does not necessarily reflect the opinions of any of the collaborators of this project.

[2] Again, please note that these tweets do NOT necessarily reflect our opinions!

4.4 Next steps

In order to improve the reliability of our analysis, we could increase the scope of our collected data. For instance, we could manually categorise more users to increase the number of tweets in the corresponding part of the analysis. In addition, we could incorporate a range of different languages to capture more global trends in the discourse around the COP conferences. However, this would require adapting our sentiment model.

Refining the sentiment model could involve finding more training data, or training data which is more relevant to the topic of our study. We could also use pre-trained sentiment models and investigate the reproducibility of our analysis with those differing models.

References

[1] Dahal Biraj, Sathish Kumar, and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, 9(24), June 2019.

[2] Dag Elgesem, Lubos Steskal, and Nicholas Diakopoulos. Structure and content of the discourse on climate change in the blogosphere: The big picture. *Environmental Communication*, 9:169–188, 04 2015. doi: 10.1080/17524032.2014.983536.

[3] M. Falkenberg, A. Galeazzi, M. Torricelli, N. Di Marco, F. Larosa, M. Sas, A. Mekacher, W. Pearce, F. Zollo, W. Quattrociocchi, and A. Baronchelli. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12):1114–1121, December 2022.

[4] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014. doi: 10.1609/icwsm.v8i1.14550. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/14550`.

[5] Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

[6] Neetu Pathak, Michael J. Henry, and Svitlana Volkova. Understanding social media's take on climate change through large-scale analysis of targeted opinions and emotions. *Conference: The AAAI 2017 Spring Symposium on Artificial Intelligence for Social Good (AISOC 2017), March 27-29, 2017, Stanford, California, 45-52*, 3 2017. URL `https://www.osti.gov/biblio/1361990`.

[7] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

[8] Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2088. URL https://aclanthology.org/S17-2088.

[9] Abhishek Samantray and Paolo Pin. Credibility of climate change denial in social media. *Palgrave Communications*, 5(1):1–8, December 2019. doi: 10.1057/s41599-019-0344-4. URL https://ideas.repec.org/a/pal/palcom/v5y2019i1d10.1057_s41599-019-0344-4.html.

[10] Manfred Stede and Ronny Patz. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlp4posimpact-1.2. URL https://aclanthology.org/2021.nlp4posimpact-1.2.

[11] UNFCCC. United nations official cop24 event page, 2018. URL https://unfccc.int/event/cop-24. Accessed: 9-Feb-2023.

[12] UNFCCC. United nations official cop25 event page, 2019. URL https://unfccc.int/event/cop-25. Accessed: 9-Feb-2023.

[13] UNFCCC. United nations official cop26 event page, 2021. URL https://unfccc.int/event/cop-26. Accessed: 9-Feb-2023.