Proceedings of LATERAISSE Workshop @LREC2022, pages 17–25
Marseille, 25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

A Cancel Culture Corpus through the lens of Natural Language Processing

Justus-Jonas Erker [1], **Catalina Goanta** [2], **Gerasimos Spanakis** [1]
[1] Maastricht University, [2] Utrecht University
{j.erker@student., jerry.spanakis@}maastrichtuniversity.nl
e.c.goanta@uu.nl

Abstract

Cancel Culture as an Internet phenomenon has been previously explored from a social and legal science perspective. This paper demonstrates how Natural Language Processing tasks can be derived from this previous work, underlying techniques on how cancel culture can be measured, identified and evaluated. As part of this paper, we introduce a first cancel culture data set with of over 2.3 million tweets and a framework to enlarge it further. We provide a detailed analysis of this data set and propose a set of features, based on various models including sentiment analysis and emotion detection that can help characterizing cancel culture.

Keywords: Social Media Analysis, Sentiment Analysis, Offensive Language Detection, Hate Speech Detection, Irony Detection, Emotion Detection, Cancel Culture

1. Introduction

Cancel culture is a phenomenon inherently linked to the Internet (Romano, 2019) that generally refers to the situation where an individual is 'professionally assassinated' (Carr, 2020). Unlike a movement, cancel culture on social media has 'neither leaders nor membership' (Mishan, 2020), but has rather emerged from earlier offline practices of public shaming such as call-out culture, boycotting (Mishan, 2020) or banishment (Kato, 2020) by attributing new meaning to terms (e.g. to cancel) cultivated in popular culture (Romano, 2019). Yet on the Internet, public shaming developed its own specific expressions, whether called 'the human flesh search engine' in the East (Shen, 2017) or 'cancel culture' in the West.

The resulting concept has a fuzzy meaning, challenging the limits of free speech on the one hand (Shen, 2017) and defamation on the other (Carr, 2020). Anyone can get cancelled, whether they are an online (e.g. Youtuber) or offline (e.g. politician) personality (Zurcher, 2021); whether they are a celebrity or an average citizen (Thomas, 2020). In addition, debates labeled as cancel culture have equally focused on non-human targets, such as children's books (Cantrell and Bickle, 2021).

The social justice aspects of cancel culture have raised acclaim in both popular and scientific literature. For instance, according to (Clark, 2020), 'cancel culture' reflects a critique of systemic inequality which has democratized public discourse. At the same time, given its virality, the concept transcended the queer communities of color it is said to have originated from (Clark, 2020), and has been used to often double as a mob intimidation technique (Romano, 2019). The gains and perils arising out of cancel culture very much depend on how this concept is framed. So far, academic scholarship investigated this phenomenon almost exclusively from a social science perspective, emphasizing power narratives connected to theoretical frameworks in critical studies (Bouvier and Machin, 2021) (Clark, 2020) (Veil and Waymer, 2021). As social media platforms are increasingly called upon to comply with state-mandated standards of content regulation, it is important to understand how cancel culture can be defined and measured.

This paper contributes to the debate by unpacking cancel culture and proposing a taxonomy of constitutive elements. Based on these elements, we propose a translation into a cohesive framework based on a collection of NLP tasks, which is currently missing from interdisciplinary as well as computational literature. We provide a detailed analysis of these measurements. Furthermore, we introduce a dataset of 22 cancel culture cases with over 2.3 million tweets, a data collection technique and a framework for enlarging this data set in the future. We discuss limitations of the proposed data gathering techniques as well as the limitations of measurements. With contributing this first data set, we hope to tackle these limitations in the future to get even more insights of cancel culture from an empirical perspective. To summarize, we investigate 2 research questions:

1. Can cancel culture incidents on Twitter be identified?

2. Can data gathering for cancel culture incidents be automated?

The paper is structured as follows: the first part of the paper describes the phenomena of cancel culture and maps it to a cohesive framework based on a collection of NLP tasks in §2. We describe our data collection technique in §3, followed by the description of our used

features in §4. Based on these features, we investigate cancel culture in §5 and build our mathematical framework for enlarging the provided data set in §6. Following, we will discuss and wrap up the results in §7 and §8.

2. Theoretical Framework

This section describes the current work and how the characteristics of cancel culture can be mapped to NLP tasks.

2.1. Characteristics of Cancel Culture: A definitional overview

Given its varied usage, 'there is no single accepted definition of cancel culture' (Gerstmann, 2020). Mainstream media accounts have tried to pinpoint at the meaning of this phenomenon by framing it alongside moral lines, such as 'the public shaming of those deemed moral transgressors' (Mishan, 2020), or by focusing on the speakers: 'it is about unaccountable groups successfully applying pressure to punish someone for perceived wrong opinions.' (Gerstmann, 2020). Social science has led to more granular definitions. (Thomas, 2020) defines cancel culture as 'a way to call on others to reject a person or business', which can occur 'when the target breaks social norms - for example, making sexist comments - but it has also happened when people have expressed opinions on politics, business and even pop culture.' Focusing on a range of triggering causes for social justice, Ng portrays cancel culture as 'the withdrawal of any kind of support (viewership, social media follows, purchases of products endorsed by the person, etc.) for those who are assessed to have said or done something unacceptable or highly problematic, generally from a social justice perspective especially alert to sexism, heterosexism, homophobia, racism, bullying, and related issues.' Ng (2020). To 'cancel' a speaker has also been framed as 'an expression of agency, a choice to withdraw one's attention from someone or something whose values, (in)action, or speech are so offensive, one no longer wishes to grace them with their presence, time, and money.' (Clark, 2020; Bouvier and Machin, 2021). More succinctly, (Randall, 2021) believes the phenomenon to be a 'modern form of ostracism and harassment', while (Velasco, 2020) describes it as 'a sporadic collective social movement leveled against individuals who infringe on the loose norms of social acceptability'.

The range of definitions explored above generally converges on a few key components: the target committing a perceived social wrong, the cause relating to justice, and the call to withdrawing support. From this perspective, cancel culture represents a unilateral act in that it does not entail 'hearing and analyzing multiple and competing voices' in the context of conflicting moral values (Veil and Waymer, 2021).

2.2. Unpacking Cancel Culture

As indicated above, social science literature has so far focused on the social justice narratives behind cancel culture, to justify it as an expression of empowerment in the face of systemic inequality and unfairness. Given its significant legal and economical consequences, it is essential to contribute to existing discussions by identifying the constitutive characteristics of this complex socio-cultural phenomenon as it unfolds on social media. We therefore propose an original taxonomy which allows for a closer examination of the various aspects of cancel culture as outlined in popular depictions of its definition and scope.

Overall, we identified five main constitutive characteristics of cancel culture:

- **The target**: This is the object of cancel culture, and it covers a wide range of options. Not only individuals can be cancelled, but also businesses, and things such as children's books (Helmore, 2021) or other cultural products such as movies (Provost, 2020).

- **The ad hoc swarm**: This reflects the critical voices engaged in cancelling the target. Unlike coordinated raids organized on specific platforms (e.g. 4chan) and executed on others (Hine et al., 2017), cancel culture entails a more organic expression of moral righteousness (Chiou, 2020).

- **Perceived wrong**: This is the action (or inaction) perceived by the swarm as morally or legally unjust, and it also comes with a presumption of guilt attributed to the target.

- **Cause**: This reflects the nature of the perceived wrong as a type of injustice grave enough to the swarm to merit collective action.

- **Demands/actions**: This is the justice goal pursued by the swarm, a finality that is aimed as a punishment for the perceived wrong, and it can range from asking for someone to be fired, for a certain action to stop (e.g. not displaying a movie on Netflix). The demands pursued by the swarm are intended to bring attention to the perceived wrong, and in doing so, exercising pressure through comments, trending hashtags, etc.

2.3. Towards NLP Tasks

Based on the proposed taxonomy of the constitutive characteristics of cancel culture, we propose a series of NLP tasks that will be described in the following.

2.3.1. Text Classification

As explained in §2.1 a major characteristic of cancel culture is the targeting of an entity with a call to action for perceived wrong expressed in language. In the context of cancel culture, this language has been shown to

be tending towards negative emotions and in some situations even hate speech (Hooks, 2020, p. 21, 36). Depending on the domain and the corresponding audience (which will differ by average age, interests, etc.) of a potentially canceled entity, a wide range of different language forms is to be expected. Therefore, being able to classify certain types of speech and sentiments of the language, and detecting possible anomalies along a certain time period can be expected to support detecting cancel culture events.

2.3.2. Actor Analysis (Target Filtering)

Since cancel culture demands a target, actor analysis could be used to filter out tweets that do not concern a specific target. While Named Entity Recognition could be helpful for this, it is most convenient to just filter for tweets that mention/target the entity in question directly.

2.3.3. Action Analysis

Another big part defined in §2.1 is the presence of demand for action. As this demand can be very broad depending on the domain in which an entity and its community are operating in, a possible solution is extraction of verbs using a Part-of-Speech (POS) tagger, and using a statistical model to count the frequent use of negative verbs (such as fired, resign, etc.) that might indicate cancel culture.

3. Data Collection

As part of this paper, a set of cancel culture cases from Twitter has been collected. The dataset is available on a GitHub repository [1] with the necessary data statements (Bender and Friedman, 2018). Furthermore, we provide a detailed description of the data set in the Appendix C. To ensure the quality of the data set, we have derived, based on §2.1, the following collection procedure.

3.1. Cancel Culture and Google Trends

As previously described, some components of cancel culture like the ad hoc swarm can not be seen as an attribute with some threshold that leads to a binary classification of cancel culture, but rather as a spectrum. If an ad hoc swarm becomes larger and larger, the attention from media towards the cancel culture candidate increases correspondingly. As soon as the cancel culture case has a sufficient attention (i.e. the ad hoc swarm is of sufficient size), newspapers are going to pick it up as cancel culture. As soon as this happens, an amplification loop begins where more and more people start searching the web regarding this cancel culture case, which on the other hand pushes the news even further in the most popular queries ranking. If the case becomes big enough, the given target will correlate with the key word "cancel culture" on Google Trends for the given time period. Previous work has shown that Google Trends can be used as a reliable source to measure the interest in conservation topics and the role of online news within the internet (Nghiem et al., 2016), especially also for exploring cancel culture (Etheve, 2020). We are going to pick up on these insights by crawling through short time periods on Google Trends and investigating search terms (cancel culture candidates) that correlate to "cancel culture".

3.2. Collection Procedure

The cancel culture cases are collected as follows:

1. find candidates on Google Trends (e.g. using the Google Trends API)
2. check in newspapers if cancel culture case
3. identify first occurrence of cancel culture case
4. gather cancel culture case from Twitter (before and after)

Once a cancel culture candidate "entity" is found on Google Trends, news articles from that time referencing that entity are investigated. For this, we use Google's advanced search, that allow us to query news articles in the corresponding time window of the gathered tweets. If this entity is canceled according to §2.1, the candidate is added to the corpus. This step is essential, as an entity can be associated with cancel culture just by speaking out on the subject.
Following up, the found articles are explored to determine the first date of the cancel culture case. To be able to analyze the data, tweets mentioning the target entity are scraped before and after the first occurrence of cancel culture.

3.3. Shortcomings of Data Collection

The proposed data collection technique requires a lot of manual work, which is very time-consuming. Furthermore, the cancel culture cases investigated are on top of the cancel culture spectrum, i.e. the ones which got the most global attention.

4. Feature Generation

We use the following features: various output probabilities from pre-trained language models that capture the affective state (which we will describe in 4.1) and action features in the form of verb frequency and tweet frequency (which we will further explore in 4.2). Verb frequency aims to estimate cancel action and tweet frequency aims to estimate the ad hoc swarm size. Respectively, a time interval T is introduced that corresponds to some cancel culture case D. This cancel culture case has $D_t \subset D$ that contains all tweets of one day t. The generated features f for a time interval t are all part of the feature vector F_t. For this feature vector, many models are combined to support the modelling process that we will explore further. Measuring the size of ad hoc swarm is done by counting

[1] https://github.com/Justus-Jonas/Cancel-Culture-Corpus

the number of tweets per t and then normalizing them over the observed time span T. This normalized tweet frequency for the corresponding time span t is added to the feature vector F_t.

4.1. Text Classification

The text classification approach is based onsec an existing RoBERTa models from TweetEval tweeteval, which are used for five tweet classification tasks. The following models with the corresponding features are included:

- Sentiment analysis with *positive, neutral and negative*
- Offensive language detection with *offensive*
- Hate speech detection with *hate*
- Irony detection with *irony*
- Emotion detection with *anger, joy, sadness and optimism*

Before the data are fed into the model, non-textual noise is removed (like links, images, etc.). For every tweet of a day t in the time interval T, all Softmax Scores outputs per tweet F_s of the models are generated, which are then averaged for each day. This gives a set of sentiment and classification features (F_{s_t} for each day, as equation 1 shows.

$$\forall f \in F_s, t \in T : \overline{f_t} = \frac{\sum_{d \in D_t} f_t(d)}{|D_t|} \qquad (1)$$

Finally, the aggregated outputs of each day are used as features for the mathematical framework. Due to the size of tweets processed, the number of investigated tweets is limited per day to 10000 for computational reasons.

4.2. Action Analysis

In order to be able to measure the frequency of cancel culture as described in §2.3.3, all cases are scraped with 10 days prior to the initial cancel culture event. The data are split into two, prior cancel culture and during cancel culture. Both data sets are preprocessed, in which we remove non-textual noise and apply lemmatization. To identify verbs, Part-of-Speech tagging is applied. Following, the frequency of every verb is calculated over all cancel culture cases (with its prior cancel culture data). Now, the frequencies of the two vectors V_C (verb frequency cancel culture) and V_B (verb frequency before cancel culture) are subtracted from each other $V = V_C - V_B$. The most frequent verbs are added to the cancel culture verb dictionary.

Applying Action Analysis

With the generated cancel culture verb dictionary, terms are counted for every time step t and are aggregated together to form a continuous value. Similar to the tweet frequency (`frequency_normalized`), this continuous value is $0 - max$ normalized (`verb_freq_normalized`) and both are given as an additional feature as F_A in the modelling process. These features are concatenated with F_s to one feature vector F.

5. Data Analysis

As part of the data analysis, investigated the generated feature vector F of 22 manually identified cancel culture cases. Overall, all cases follow a similar pattern with little variation. In the following we will describe general characteristics of cancel culture and special cases we observed. Nonetheless, we provide a more detailed analysis in the appendix C in table 1 where we investigate the Pearson correlation between the tweet frequency (the size of the ad hoc swarm) and other features.

5.1. Tweet Frequency

As shown in the first sub-figure (left) of figure 1 Jimmy Fallon got canceled on day 7. While the negative sentiment increases rapidly, the most obvious increase in the total number of tweets, which is linked to the size of the ad hoc swarm. This behavior generalizes to most observed cancel culture cases, but might differ in its extremes. An observed cancel culture case from `@Pepsi` multiplied tweets by a factor of around 53 within two days while in the case of Jimmy Fallon only a factor of about 15 is observed. In cases where the attention before was very low (`@Shanemgillis`) where we observed jumps from 27 tweets a day to over 120000 tweets a day. While we did not run in any problems of extreme fluctuations of Sentiment values due to the small number of tweets before the cancel event, this still might be something to keep in mind when working on cancel culture cases where an entity usually does not retrieve as much public attention.

Nonetheless, we also found some special cases (see appendix `@gabecake`) where we observe a significant raise in frequency prior to the actual cancel culture event. After investigating Gabi DeMartino related news articles on the 11/30/2020, one day before the corresponding Twitter account `@gabecake` got canceled, we found that she launched a new product and teased a new song that gave her a lot of positive public attention on that day [2]. One day after that she posted a video on a platform which got her banned for ethical and legal concerns [3] and caused the phenomena of cancel culture. We investigated two similar cases where either the dynamic of the public opinion changed due to new events happening in a debate `@UnburntWitch` [4] or where a large supportive movement emerged simul-

[2] https://www.justjaredjr.com/2020/11/30/gabi-demartino-launches-new-fragrance-beautiful-mess-

[3] https://www.buzzfeednews.com/article/tanyachen/onlyfans-suspended-youtuber-gabi-demartinos-ac

[4] https://bentcorner.com/zoe-quinn-alec-holowka-suicide/

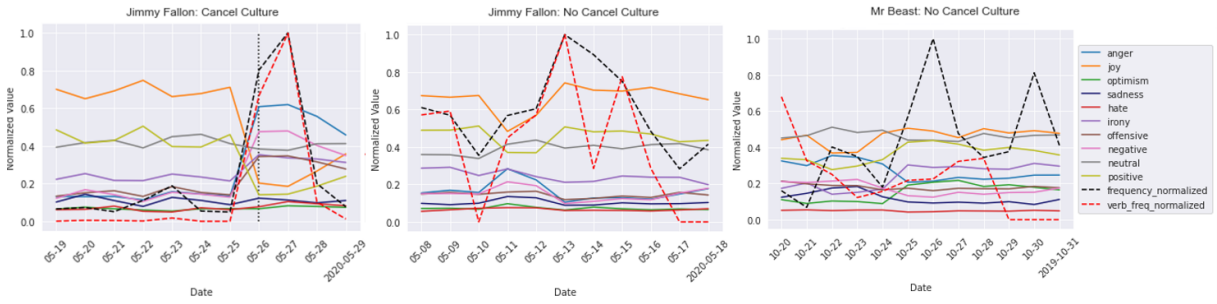

Figure 1: A case of cancel culture (05-19-2020 - 05-29-2020) vs a non cancel culture (05-08-2020 - 05-18-2020) vs. an anti cancel culture event Mr. Beast (10-20-2019 - 10-31-2019)

taneously `@Lin_Manuel` [5] [6]. The pearson correlation coefficients between negative Sentiment values and frequency reflect on that (see Appendix C).

5.2. Cancel Culture associated Verbs Frequency and Text Classification

Apart from the increase of negative and decrease of positive sentiment, the normalized frequency of cancel culture associated verbs strongly correlates with a cancel culture event, while it randomly fluctuates in cases of non cancel culture as it can be seen in subfigure (middle) of figure 1. This dataset is also of Jimmy Fallon, ten days before the figure on the left begins (prior to the cancel culture event). The third graph shows an "anti" cancel culture case when Mr Beast, a famous YouTuber, started Team Trees, an initiative to plant a lot of trees, which got him a great amount of positive attention. Demonstrating that the standalone feature of tweet frequency is not sufficient for identification of cancel culture. One such a feature that helps to distinguish cancel culture is anger. As can be seen in the first graph, anger increases on the day that Jimmy Fallon is canceled. Similar can be observed for negativity, offensive language, irony and the negated for joy and positivity.

The first graph shows an interesting characteristic of cancel culture. After Jimmy got canceled for about 2 days, the frequency of tweets about him dropped drastically, indicating that people lost interest in actively tweeting about him. However, the amount of anger and negativity lingers after the amount of tweets drops. Further exploration shows that this phenomenon generalizes to most other cancel culture cases. The average duration of the spike in frequency, which is calculated by the difference between the day with the highest increase in frequency and the day with the largest decrease in frequency, is only 1.95 until it approaches its baseline again. The same is calculated for the spike in negative sensitivity, and there the average duration is 2.95. From the gathered data, it can therefore be concluded that the negativity in general stays longer than the increase in attention.

[5] https://mickeyblog.com/2020/07/05/review-hamilton-is-the-best-movie-of-2020/

[6] https://edition.cnn.com/2020/07/07/entertainment/lin-manuel-miranda-hamilton-slavery/index.html

6. Mathematical Framework for identifying Cancel Culture

As described in §3.2, investigating news articles is necessary to identify whether cancel culture is present in a particular case. This section presents a technique that automatically identifies cancel culture and therefore allows a complete automation of the proposed gathering process introduced in §3. We define a mathematical framework based on the data analysis of §5. The complete model is split up into two phases, as shown in Figure 2. First, a model is used to determine whether cancel culture is present in the given dataset or not. If cancel culture is present, a different statistical model detects on which day the target got canceled exactly.

6.1. Cancel Culture Identification

In order to detect whether cancel culture is present in the dataset, the features that are generated by the Text Classification and the action analysis are aggregated per day. Additionally, the normalized frequency of tweets is added as an additional feature. Now that the features are aggregated, the model adds all scores of negative emotions
$F_n = \{\text{anger, sadness, hate, irony, offensive, negative}\}$ together, aggregated by day, which gives a 'negativity score'. Then, the day with the highest negativity score is compared to the day with the lowest negativity score before t occurred to calculate the slope. The difference between the feature values F_D, calculated in equation 4 of those two days specified in Equations 2 & 3 is then used as a feature vector for a cancel culture dataset D to detect cancel culture.

$$F_{D_{max}} = \max_{t \in T}\Big(\sum_{f \in F_{n_t}} f\Big) \tag{2}$$

$$F_{D_{min}} = \min_{t' \in \{0,..,t_{max}-1\}}\Big(\sum_{f \in F_{n_{t'}}} f\Big) \tag{3}$$

$$F_D = F_{D_{\max}} - F_{D_{min}} \tag{4}$$

Once the final features F_D are generated, a Support Vector Machine (SVM) classification model is used to test our hypothesis whether Cancel Culture on Twitter can be identified. The SVM uses an RBF kernel with

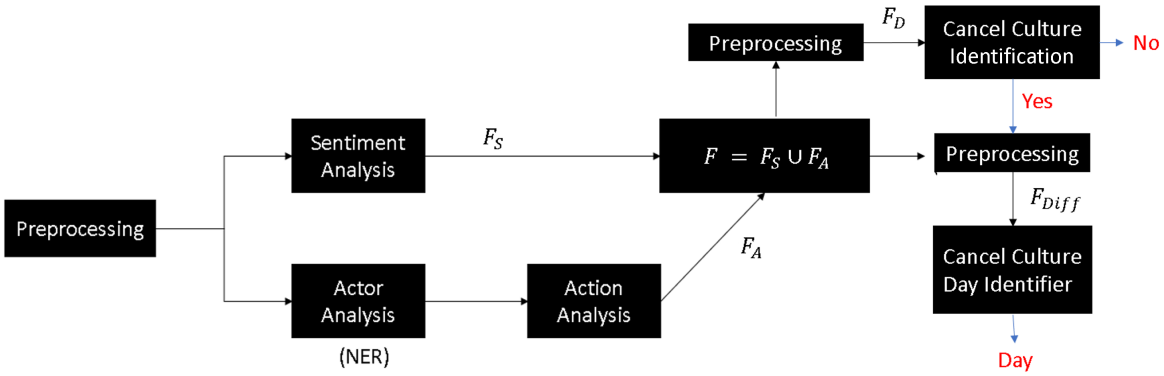

Figure 2: From Feature Generation to the General Model containing the Cancel Culture Identification and Cancel Culture Day Identifier.

a regularization parameter of 2 to be able to create a decision boundary, where all data points on one side of the decision boundary indicate an absence of cancel culture, and all data points on the other side of the decision boundary indicate a presence of cancel culture.

6.2. Cancel Culture Day Identification

In order to detect the date on which the target was canceled, a date identifier algorithm is used after a cancel culture case is predicted. In particular, the difference between every feature for each day is computed so the value of increase or decrease respectively can be distinguished. Furthermore, the day that cancel culture case has occurred is selected according to the maximum negative increase of activity. In other words, the day that has the highest negative change on the calculated difference of the features is declared to be the day that cancel culture occurred. Finally, a delta time value is calculated, which is the difference between the sum of the largest changes and the value of the first day, in order to distinguish the deviation. Below, the mathematical formula 5 shows how the difference is calculated, while equation 6 shows the process on finding the biggest slope of negative increase F_{Diff} using the negative features F_n for a time step t with $F_{Diff_{n_t}}$.

$$F_{Diff} = F_t - F_{t-1} \tag{5}$$

$$t_{start} = \max_{t \in T}\Big(\sum_{f \in F_{Diff_{n_t}}} f\Big) \tag{6}$$

t_{start} is then selected as the first day of cancel culture.

6.3. Evaluation Results

To test our hypothesis of the provided framework, the 20 cancel culture cases are mixed with 23 negative events, of which 20 are prior cancel culture data of the corresponding cases (the week before the cancel culture event) and 3 ”anti” cancel culture cases like Mr. Beast (see §5) that demonstrate an adhoc swarm. The model is able, apart from one data point, to separate all cases from each other. For the Cancel Culture Day Identifier, the standard deviation is calculated for the amount of days the statistical model is off on its prediction. On the current dataset, the day identification model has an average deviation 0.59 days. We also have used this model to enlarge the data set. Specifically, we gathered 4 more cases (2 positive and 2 negative) of which the model was able to identify cancel culture correctly. The two positive cases (`Bob Baffert` and `pepe le pew` have been added to the data set while the negative samples (`Katt Williams` and `Rowan Atkinson`) have been added to the appendix B. Based on news article investigations, we could confirm that the identified starting days t_{start} were in both positive cases correct.

6.3.1. Permutation Feature Importance

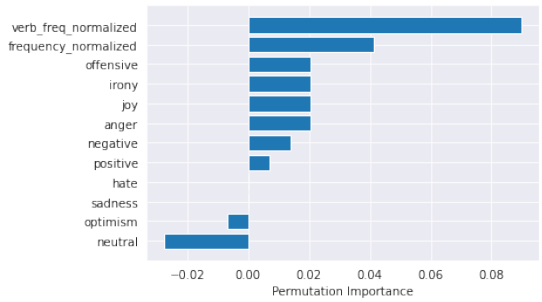

Figure 3: Permutation Feature Importance of identifying cancel culture

To get an idea of the importance of each feature, the SVM is given the same data set as a classification task, where the permutation feature importance is measured. As figure 3 demonstrates, the normalized verb frequency is most important, which aligns with the hypothesis stated in §5. Furthermore, frequency plays

a significant role in detecting a cancel culture event. Offensiveness, joy, anger, irony, followed by negativity are the most important sentimental and emotional features, which also aligns with the hypothesis in §5. While the hate language detection model has already shown in the data analysis 1 that it seems not to correlate with cancel culture, similar to the positive score, we interpret this insignificance due to the redundancy of other correlating features. The assumptions in §5 of the irrelevance of sadness and optimism amplified.

7. Discussion

Concluding from §6.3.1, the sufficient elements to identify cancel culture appear to be a combination of sentiment analysis, emotion detection and irony detection (of tweeteval) together with the frequency of tweets (i.e. a measurement for the size of the ad hoc swarm) and the presence of verbs that correlate with cancel culture (action analysis).

7.1. Limitations

While both models seem to be able to create decision boundaries that make the feature vectors of cancel culture and non cancel culture events separable, it is important to note that as described in §3.3 the gathered data is biased by the attention from the media like news organizations. Entities of public interest are therefore more likely to be picked up by our gathering technique, which is why our findings are only representative to cases of public figures.

While we had some cases like Goya (see §C) where the baseline prior was only a few tweets a day, this might get even worse when looking at more privately preserved cases. This could easily lead to fluctuations in the frequencies as well as text classification values, leading to a misclassification due to the way that the values per day are aggregated, a day with a very small amount of tweets can easily become an outlier. For example, if on a certain day only two people tweet about them and both of these tweets are negative, this can create a spike in negativity that might trick the model into believing that the target gets canceled on that day, especially if on the other days the number of tweets was even lower.

Moreover, we addressed the problem of dynamic changes within public opinion that might change very quickly or create movements that happen simultaneously, making the identification more difficult (see §5 for `@gabecake` etc.). Considering the Cancel Culture Day Identifier, one limitation is that the fact that the only consideration is the increase of features per day. However, if a person starts getting canceled at the end of the day, this increase might not be apparent immediately, and it will only become visible on the next day. In order to circumvent this, the data could be split per hour instead of per day. A downside to this is that more data would be needed, this is especially a problem if only a few number of tweets per day are given as baseline.

7.2. Research Questions

Given the previous analysis, we can conclude that cancel culture can be identified with the limitation to public figures, answering our first research question. Similarly, as far as RQ2 is concerned, we have demonstrated that using Google Trends as well as the provided mathematical framework can be used for automatically expanding the data set.

8. Conclusion

This paper introduces cancel culture to the computational literature and demonstrates that it's a phenomenon that can be empirically observed and studied using a combination of NLP techniques, including sentiment analysis and emotion detection. We find that cancel culture is rather short-lived, with an attention peak of 1.95 days and a peak in negative expression of 2.95 days (§5). Furthermore, we introduce a first public data set with over 2.3 million tweets of 22 cancel culture cases and a mathematical framework for automatically enlarging it in the future. Gathering such large number of tweets is very time-consuming, not only because of API constraints but also considering that every single tweet has to be processed by 5 different Transformer models. We hope that with a joint call to the interdisciplinary social medial analysis community, we can scale up this data set together to get more insight in this new emerging social phenomenon.

References

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bouvier, G. and Machin, D. (2021). What gets lost in twitter 'cancel culture' hashtags? calling out racists reveals some limitations of social justice campaigns. *Discourse & Society*, 32(3):307–327.

Cantrell, K. and Bickle, S. (2021). Cat in a spat: scrapping Dr Seuss books is not cancel culture.

Carr, N. K. (2020). How Can We End #cancelculture - Tort Liability or Thumper's Rule? *The Catholic University Journal of Law and Technology*, 28:15.

Chiou, R. (2020). We need deeper understanding about the neurocognitive mechanisms of moral righteousness in an era of online vigilantism and cancel culture. *AJOB Neuroscience*, 11(4):297–299. PMID: 33196355.

Clark, M. D. (2020). Drag them: A brief etymology of so-called "cancel culture". *Communication and the Public*, 5(3-4):88–92.

Etheve, S. (2020). Exploring cancel culture.

Gerstmann, E. (2020). Cancel Culture Is Only Getting Worse.

Helmore, E. (2021). 'It's a moral decision': Dr Seuss books are being 'recalled' not cancelled, expert says, March.

Hine, G. E., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G., and Blackburn, J. (2017). Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. *arXiv:1610.03452 [physics]*, October. arXiv: 1610.03452.

Hooks, A. M. (2020). Cancel culture: Posthuman hauntologies in digital rhetoric and the latent values of virtual community networks. page 107.

Kato, B. (2020). What is cancel culture? Everything to know about the toxic online trend, July.

Mishan, L. (2020). The long and tortured history of cancel culture, Dec.

Ng, E. (2020). No grand pronouncements here...: Reflections on cancel culture and digital media participation. *Television & New Media*, 21(6):621–627.

Nghiem, L. T. P., Papworth, S. K., Lim, F. K. S., and Carrasco, L. R. (2016). Analysis of the capacity of google trends to measure interest in conservation topics and the role of online news. *PLoS ONE*, 11.

Provost, C. (2020). 'Cuties' culture war is dramatic and global – but no surprise.

Randall, M. E. (2021). Cancel culture and the threat to progress in radiation oncology. *Practical Radiation Oncology*.

Romano, A. (2019). Why we can't stop fighting about cancel culture, Dec.

Shen, W. (2017). Online Privacy and Online Speech: The Problem of the Human Flesh Search Engine. *University of Pennsylvania Asian Law Review*, 12:44.

Thomas, Z. (2020). What is the cost of 'cancel culture'? *BBC News*, October.

Veil, S. R. and Waymer, D. (2021). Crisis narrative and the paradox of erasure: Making room for dialectic tension in a cancel culture. *Public Relations Review*, 47(3):102046.

Velasco, J. C. (2020). You are cancelled: Virtual collective consciousness and the emergence of cancel culture as ideological purging. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 12(5).

Zurcher, A. (2021). Cancel culture: Have any two words become more weaponised? *BBC News*, February.

A. A special case of Cancel Culture

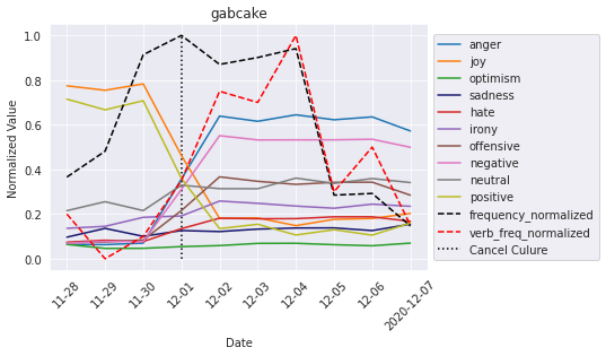

Figure 4: gabcake cancel culture case showing a significant increase in frequency right before cancel culture event.

B. Data Gathering: Negative Samples

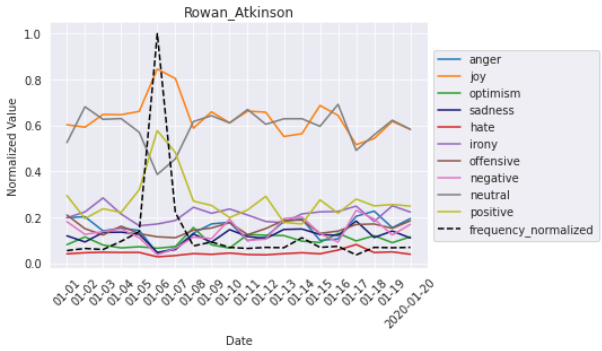

Figure 5: Rowan Atkinson talked about cancel culture which led to a Google trends correlation

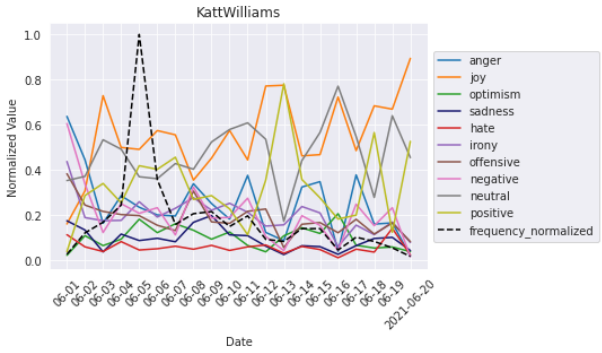

Figure 6: Katt Williams talked about cancel culture which led to a Google trends correlation

C. Analysis Cancel Culture Cases

Table 1: Every Cancel Culture Case (in order) by name, number of total tweets, min number of tweets per day, max number of tweets per day, first day of gathered data, last day of gathered day, the identified first day of cancel culture and the Pearson correlation coefficients between the frequency (the size of the ad hoc swarm) and the other features.

Name	number tweets	min tweets	max tweets	first date	last date	cancel date	Pearson correlation coefficients to frequency of Tweets									
							anger	joy	sadness	optimism	irony	hateful	offensive	positive	neutral	negative	verb freq
Lin_Manuel	173423	3025	57053	7/1/20	7/12/20	7/4/20	-0.37	0.33	0.08	-0.43	0.14	-0.32	-0.23	0.63	-0.83	-0.31	0.94
UnburnWitch	5214	10	2192	8/24/19	9/6/19	8/31/19	-0.5	-0.23	0.77	0.64	-0.19	-0.55	-0.37	0.43	-0.46	-0.18	0.86
alisoneroman	21756	59	14844	5/6/20	5/15/20	5/8/20	0.39	-0.36	-0.1	-0.02	0.26	0.55	0.42	-0.41	-0.13	0.44	1.0
arniebummer	7374	41	1797	1/8/21	1/17/21	1/10/21	0.47	-0.41	-0.51	0.27	0.5	0.15	0.69	-0.67	0.57	0.65	0.86
bobbafferi	2595	22	869	5/7/21	5/17/21	5/9/21	0.58	-0.49	0.14	-0.37	0.62	0.53	0.61	-0.47	-0.58	0.62	0.9
CarsonKing2	8777	34	4745	9/16/19	9/29/19	9/25/19	0.6	-0.67	0.43	-0.11	0.18	-0.14	0.36	-0.62	0.38	0.55	1.0
dojacat	185425	3978	49731	5/21/20	5/30/20	5/22/20	0.41	-0.58	0.85	-0.7	0.59	-0.38	0.67	-0.62	-0.43	0.62	0.98
gabcake	11657	276	1881	11/28/20	12/7/20	12/1/20	-0.02	0.04	-0.28	-0.2	0.12	-0.04	0.04	0.04	-0.06	-0.03	0.5
gmacanno	264909	849	121959	2/8/21	2/17/21	2/10/21	0.5	-0.53	-0.28	-0.29	0.61	-0.11	0.44	-0.46	-0.16	0.55	1.0
goya	228223	1461	108553	7/6/20	7/13/20	7/9/20	0.6	-0.6	-0.13	0.57	0.63	0.48	0.56	0.63	-0.65	0.57	1.0
jamescharles	132301	2503	32749	5/7/19	5/16/19	5/10/19	0.28	-0.28	0.53	-0.22	0.59	0.25	0.36	-0.39	0.37	0.39	0.98
jimmyfallon	46476	884	18108	5/19/20	5/28/20	5/26/20	0.86	-0.86	0.23	0.18	0.84	0.6	0.87	-0.83	-0.71	0.88	0.96
jk_rowling	89803	1147	28968	9/9/20	9/18/20	9/13/20	0.62	-0.61	0.53	-0.47	0.72	-0.06	0.85	-0.8	-0.36	0.69	0.99
LanaDReyOnline	349450	10374	152633	5/16/20	5/25/20	5/21/20	0.74	-0.74	0.68	-0.52	0.62	0.63	0.75	-0.68	-0.65	0.69	1.0
MorganWallen	59051	846	19968	2/1/21	2/9/21	2/3/21	0.68	-0.69	0.67	-0.32	0.46	0.1	0.73	-0.67	-0.87	0.77	0.99
pepe le pew	42960	36	10775	3/1/21	3/14/21	3/6/21	0.56	-0.61	0.4	0.11	0.4	-0.69	0.45	-0.72	-0.1	0.41	0.82
pepsi	298336	2679	156846	3/30/17	4/8/17	4/4/17	0.75	-0.7	0.33	-0.12	0.7	0.16	0.7	-0.69	-0.7	0.73	1.0
projared	71702	27	40000	5/6/19	5/15/19	5/9/19	0.45	-0.42	0.04	-0.43	0.48	0.52	0.58	-0.39	-0.6	0.55	0.97
sebastian stan	4609	146	1897	7/9/20	7/18/20	7/14/20	0.83	-0.82	0.7	-0.72	0.63	0.28	0.8	-0.77	-0.5	0.84	0.99
seuss	167600	383	43424	2/15/21	3/6/21	2/27/21	0.79	-0.79	0.38	-0.71	0.76	0.48	0.78	-0.74	-0.38	0.76	0.98
Shanemgillis	50889	6	16193	9/10/19	9/20/19	9/12/19	0.53	-0.55	0.49	0.15	0.31	0.24	0.52	-0.5	-0.49	0.59	0.95
starbucks	165210	7454	49505	6/6/20	6/15/20	6/10/20	0.91	-0.92	-0.12	0.13	0.8	0.59	0.89	-0.92	-0.92	0.94	0.96