Capturing Contentiousness: Constructing the Contentious Terms in Context Corpus

Ryan Brate*
KNAW Humanities Cluster
Amsterdam, The Netherlands
ryan.brate@dh.huc.knaw.nl

Andrei Nesterov*
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
nesterov@cwi.nl

Valentin Vogelmann
KNAW Humanities Cluster
Amsterdam, The Netherlands
valentin.vogelmann@dh.huc.knaw.nl

Jacco van Ossenbruggen
VU University Amsterdam
Amsterdam, The Netherlands
jacco.van.ossenbruggen@vu.nl

Laura Hollink
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
l.hollink@cwi.nl

Marieke van Erp
KNAW Humanities Cluster
Amsterdam, The Netherlands
marieke.van.erp@dh.huc.knaw.nl

ABSTRACT

Recent initiatives by cultural heritage institutions in addressing outdated and offensive language used in their collections demonstrate the need for further understanding into when terms are problematic or contentious. This paper presents an annotated dataset of 2,715 unique samples of terms in context, drawn from a historical newspaper archive, collating 21,800 annotations of contentiousness from expert and crowd workers.

We describe the contents of the corpus by analysing inter-rater agreement and differences between experts and crowd workers. In addition, we demonstrate the potential of the corpus for automated detection of contentiousness. We show that a simple classifier applied to the embedding representation of a target word provides a better than baseline performance in predicting contentiousness. We find that the term itself and the context play a role in whether a term is considered contentious.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**; • **Computing methodologies → Knowledge representation and reasoning**.

KEYWORDS

datasets; bias; crowdsourcing; knowledge capture

ACM Reference Format:
Ryan Brate, Andrei Nesterov, Valentin Vogelmann, Jacco van Ossenbruggen, Laura Hollink, and Marieke van Erp. 2021. Capturing Contentiousness: Constructing the Contentious Terms in Context Corpus. In *Proceedings of the 11th Knowledge Capture Conference (K-CAP '21), December 2–3, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3460210.3493553

*Both authors contributed equally to this research.

K-CAP '21, December 2–3, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8457-5/21/12…$15.00
https://doi.org/10.1145/3460210.3493553

NOTE: Some examples in this paper may be shocking or offensive. They are provided as illustration or explanation of the work and do not reflect the opinion of the authors or their organisations.

1 INTRODUCTION AND MOTIVATION

Cultural heritage institutions harbour vast collections that have often been compiled over long periods of time. Collection and documentation practices therefore reflect the cultural and societal norms of the various time periods during which they were compiled. As a result, they may contain terms that are inappropriate in modern society. An example of a contentious term that we find in historical documents is 'half-blood' to denote people of mixed descent. Nowadays, this term is considered offensive when discussing people, although it is still acceptable when discussing for example animals or plants.

Many institutions recognise the problem of outdated language in their collections. For example, the Amsterdam Museum published a statement in 2019 that they would not use the term "Golden Age" anymore in their exhibitions to refer to the Dutch 17th century.[1] The National Archives of the Netherlands states that they "explore the possibility of explaining language that was acceptable and common in the past and providing it with contemporary alternatives", meanwhile "keeping the original descriptions, because they give an idea of the time in which they were made or included in the collection".[2] The size of heritage collections makes investigating and manually replacing or contextualising problematic language impossible. To illustrate this: the digital collection of the National Library of the Netherlands consists of more than 120 million pages.

This work is a first step towards aiding heritage professionals to investigate and chart contentious language used in their collection on a large scale. In this paper, we present an annotated corpus of contentious terms from historical Dutch newspaper archives, including the textual contexts in which they appear, with an analyses of the signals that indicate contentious language use.

The contribution of this paper is twofold. The first contribution is the corpus of contentious terms in context (ConConCor) consisting of 2,715 unique text snippets, most of which (2,395) annotated by at least 7 annotators as to whether they consider a target word

[1] https://www.amsterdammuseum.nl/nieuws/gouden_eeuw (in Dutch) Last visited: 10/09/2021
[2] https://www.nationaalarchief.nl/taalgebruik-in-onze-archieven (in Dutch) Last visited: 10/09/2021

in the snippet to be contentious or not. By providing the potentially contentious target words in the textual context in which they originally appeared, we aim to facilitate analyses of how (much) context influences contentiousness. We use the term 'contentious' to refer to all (potentially) inappropriate or otherwise sensitive words. For example, words suggestive of some (implicit or explicit) bias. Because contentiousness is likely a subjective and hard to define notion, it is all the more important to have a relatively high number of annotators per snippet. This enables, for example, the selection of a sub-corpus of snippets on which many annotators agreed to train a machine learning model; or an analysis of snippets on which annotators did not agree. To obtain a sufficiently large number of annotations per sample, we asked both expert and crowd annotators. The code and resources used to generate ConConCor, general documentation and the corpus' accompanying datasheet (based on [7]) are available at https://github.com/cultural-ai/ConConCor.

The second contribution of this paper is a multilayered analysis of the agreement between annotators, and the distributional semantic properties of contentious words and contexts. Here, we provide insights into the following questions: 1) In which cases and to what extent do expert and crowd annotators agree on the contentiousness of a term in a given context? These analyses provide insights into the feasibility of creating reliable annotations for a concept as subjective and complex as contentiousness; 2) Can we automatically detect contentious words outside of their contexts? To answer this, we predict the contentiousness of terms based on word embeddings to establish a baseline; and 3) Do the contexts of contentious terms have any properties that distinguish them from the contexts of non-contentious terms? We answer this question by clustering the context terms based on their embeddings, and discuss the results. Our analyses show that contentiousness is a complex concept that warrants the attention of humanities scholars and computer scientists alike, if we are to develop intelligent systems that can deal with problematic language use.

2 RELATED WORK

While the topic of contentiousness touches many different fields, we focus our related work on contentiousness in the cultural heritage domain and on automatic detection of sensitive language.

2.1 Contentiousness and cultural heritage

As exemplified by the Terminology working group at Rijksmuseum Amsterdam, the Dutch National Archives offensive language reporting page, the Amsterdam Museum's decision to stop using the term *Golden Age* and the Dutch National Museum of World Cultures' *Words Matter* publication, decolonisation, and its associated contentious language use, is high on the agendas of cultural heritage institutions. This goes beyond internally inspecting their collections, but also involves engaging with the wider research community and stakeholders at symposia such as Inward Outward,[3] and Decolonizing Museums.[4]

Imbalances in heritage representation and presentation have been flagged as problematic by researchers from various fields such as bioethics [4], museology [1, 15] and social sciences [5, 14]. The debate revolves around how the visibility of historically marginalised people can be increased and how a multifaceted perspective on colonial pasts can be presented better to make museums more inclusive. However, the computer science or big data perspective that this issue needs, remains underexplored in these initiatives.

2.2 Detecting contentiousness

The computational linguistics community has been working on detecting explicit examples of harmful language such as hate speech and offensive language on Twitter [2] and through various initiatives such as the Workshop on Online Abuse.[5] Kaplun et al. [8] created a crowdsourced corpus of news articles annotated with whether terms are controversial or not. Controversy is however different from contentiousness, as the controversial lexicon used is very much a contemporary lexicon and does not necessarily contain sensitive terms that have for example a colonial connotation, but terms referring to topics that people may rather not discuss or have strong opinions on such as *finances* or *abuse*.

A survey on bias in 146 computational linguistics research papers found that there is little consensus on what bias means exactly and how it relates to research outside the compuational linguistics community [3]. Among the investigated papers, 17 deal with hate speech detection, 15 with sentiment analysis, and 8 with machine translation, which we consider closest to ConConCor. The majority of the papers are concerned with embeddings, coreference resolution and language modelling or dialogue generation.

Closest in aims to our work is [13]. Whilst not computational, Schulzke investigates the use of contentious language and how its meaning changes in various contexts through quantitatively analysing the language of one episode of *South Park*.

3 CREATING CONCONCOR

In this section, we describe the contents and creation process of the Contentious terms in Context Corpus (ConConCor). We first collected a list of potentially contentious terms as well as alternatives followed by obtaining relevant text snippets from the larger Dutch historical newspaper corpus. Finally, we discuss the three-stage approach to obtaining manual annotations regarding the contentiousness of terms in context.

3.1 Seed list of (non-)contentious terms

Our initial seed list of potentially contentious and non-contentious terms was derived from *Words Matter: an unfinished guide to word choices in the cultural sector* compiled by the Dutch National Museum of World Cultures [11]. It is available in both Dutch and English and contains descriptions of sensitive terms, their historical usage and implications, and appropriate and inappropriate usage contexts. Additionally, alternative words that could be used in relevant contexts are provided.

We selected only Dutch unigram terms from Words Matter. Compound terms, for example, "kleine mensen" ('small people') were not included in the study but we plan to include these in future experiments. In total, 91 terms were selected in this study, 76 of which were described as contentious (in some contexts) and 13 were listed

[3] https://www.beeldengeluid.nl/en/visit/events/inward-outward-symposium Last visited: 10/9/2021

[4] https://decolonizingmuseums.pl/ Last visited: 10/9/2021

[5] https://www.workshopononlineabuse.com/ Last visited: 10/09/2021

as alternative words (to contentious terms), and a further 2 were either contentious or alternative words depending on the context.

3.2 Text snippet collection

The textual contexts in which (non-)contentious terms are used, were selected from The Europeana Dutch Newspaper collection, which is a subset of the Dutch National Library's historical newspapers archive[6] spanning the period 1890-01-01 to 1941-12-31. It contains scans as well as plain text obtained by Optical Character Recognition (OCR).Due to variations in the printing process as well as the collection process, OCR errors may occur. We retrieved from the archive documents that contain one or more seed terms and that are categorised as 'article', thus excluding other types of content such as advertisements and family notices. We selected a stratified sample of potential text snippets from the retrieved documents over seed list term, decade, and newspaper issue distribution metadata. Rather than sampling text snippets uniformly within each stratum, we gave snippets sampling weights proportional to their probabilities, as estimated from the initial set of retrieved documents via trigram frequencies, for two reasons:

(1) Albeit small in comparison, the set of documents selected for annotation should be representative of the larger archive in terms of language use and semantic content. We use the probability of a document in the archive as a proxy for representativeness;
(2) There are many OCR errors in the archive, often leading to unintelligible examples for annotation. Documents with large amounts of errors are characterised by lower probabilities, as standard spelling tends to be more common than deviations introduced by OCR.

This sampling process resulted in 2,715 unique newspaper samples that contain one or more (non-)contentious terms from the seed list. Due to copyright issues, only limited snippets (140 characters maximum) can be shared directly as part of the downloadable corpus, but the repository contains the necessary code to regenerate the dataset when access to the source data is obtained.

3.3 Annotating contentiousness

We provided volunteers with the potentially contentious terms as well as the text snippets in which the terms occur, and asked them to label whether or not they deem the terms contentious given their context. If they could not decide, they could select "I don't know" or "Illegible OCR" to indicate that OCR errors prevented them from taking a decision. The annotation process included three stages: pilot annotation, expert annotation, and crowd annotation. All stages required the participation of Dutch speakers.

The pilot stage was intended for testing the layout of the annotation form, the clarity of the instructions, and the optimal number of sentences in the snippet of text that was provided as context around the potentially contentious term. Six anonymous members of the Cultural AI Lab each annotated the same 40 samples, where either a context of three or five sentences was given, and gave feedback in four open questions following the annotation task. Based on this, the instructions and layout were improved, and the number of context sentences was set to five. This data is not included in the corpus as it was a part of the task development.

	Experts	Crowd workers	Total
Nr. of batches	3	57	60
Nr. of unique samples	150	2,570	2,715
Nr. of annotations	1,000	20,800	21,800
Nr. of annotators	N/A	416	>416

Table 1: General statistics of ConConCor

In the second stage, humanities scholars from the KNAW Humanities Cluster[7] volunteered to label terms in context as contentious or not. Participation was anonymous and no demographic information was collected. Although there was no monetary reward, we gave the participants a chance to win one of 10 Cultural AI mugs by entering their email address. This stage included 3 unique batches of 50 samples. 2 of 3 batches were annotated by 7 participants and 1 by 6 participants. The number of unique annotators is unknown as one expert could annotate more than one unique batch. In total, there are 1,000 annotations divided over 150 unique samples.

The goal of the expert annotation stage was threefold: 1) to obtain judgements on the contentiousness of terms from domain experts; 2) to identify samples on which all expert annotators agree to be used as "control samples" in the crowdsourcing phase; and 3) to enlarge our list of potentially contentious terms by asking experts to point out other contentious terms in the context snippets.

Based on the domain experts' feedback, minor adjustments were made to the annotation instructions. 17 additional potentially contentious terms were included, while terms that denote historical toponyms (for example, "Bombay") were excluded from the next phase of the study as they proved difficult for experts and caused disagreement. Historical geographical information resources such as the World Historical Gazetteer[8] could be used to detect outdated toponyms, whereas such resources are not available for other types of contentious terms. Five control samples came out of the expert annotation phase: snippets that contain the contentious terms "kaffer" ("kaffir") "neger" ("negro"), and "dwerg" ("dwarf"), and the non-contentious terms "gemengd" ("mixed"), and "achtergrond" ("background").

The third annotation stage was a crowdsourcing task distributed via the Prolific platform.[9] In this stage, crowd workers received compensation for their efforts, advertised at Prolific's standard "good" rate of £7.50 per hour for the expected task duration, and ultimately being paid £12.16 on average in our task. The crowdsourcing was not anonymous as Prolific assigns unique IDs to their users and collects demographic data. The participants' IDs were anonymised in the public dataset. The demographic data is stored separately and was not considered in the analysis of the results at this stage. 416 people took part in the study. 57 unique batches of 50 samples were distributed among them. Each batch consisted of 5 control samples, 5 samples that included terms suggested by the expert annotators of the previous phase, and 40 samples based on our original seed list of contentious and non-contentious terms. Together, the crowd

[6] https://delpher.nl (in Dutch) Last visited: 10/9/2021

[7] https://huc.knaw.nl Last visited: 10/9/2021
[8] https://whgazetteer.org Last visited: 10/9/2021
[9] https://prolific.co Last visited: 10/9/2021

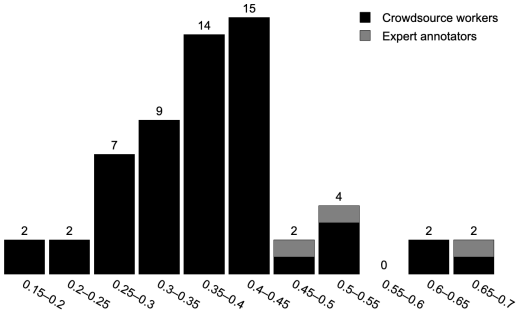

Figure 1: Nr. of batches grouped by Krippendorff's α ranges.

workers annotated 2,570 unique samples, with each sample annotated by between 1 (batch 50 was annotated only by 1 annotator due to an error in the distribution process) and 416 people (5 control samples were annotated by all annotators), amounting to 20,800 annotations. Table 1 presents the overall corpus statistics in terms of number of samples, unique samples and number of annotators.

4 CONCONCOR ANALYSIS

In this section, we present our corpus analyses, starting from the annotations, followed by the contentious terms, and the contexts in which they occur.

4.1 Annotations Analysis

We use Krippendorff's α as a measure of overall inter-rater agreement. Krippendorff's α is applicable to situations with multiple annotators per sample and is robust against missing data [9]. The α values range from -1 to 1, where 1 indicates perfect agreement, 0 means agreement as expected if annotations had been done randomly, and negative values indicate inverse agreement. In addition, we report the percentage of annotators who agreed with each other - i.e. "percentage agreement" - to convey agreement per sample. Samples with only one annotator (batch 50) are excluded.

Inter-annotator agreement. We calculated Krippendorff's α for every batch of annotated samples. When all four options ("Contentious", "Non-contentious", "I don't know", and "Illegible OCR") are used for the calculation, the median α is 0.31. This is low, which would be expected of a task with a high degree of subjectivity and/or complexity. If we take out those annotations where people selected "I don't know" or "Illegible OCR," agreement increases to an α of 0.39. Figure 1 shows the distribution of α values of the batches (based on only the 'contentious' and 'non-contentious' options). Most of the batches annotated by crowd workers have an agreement between 0.35 and 0.45. The agreement within the three batches done by expert annotators is higher: 0.46, 0.50, and 0.65.

Inter-rater agreement among crowd workers can potentially be improved by filtering out annotations of under-performing annotators. The corpus as well as the (unpublished) associated demographic data contains various indicators that can potentially be used for this purpose: completion time of the task, reported Dutch language proficiency, responses to the control questions and the agreement of an annotator with others. We test the effect of two such indicators, both available as part of the corpus.

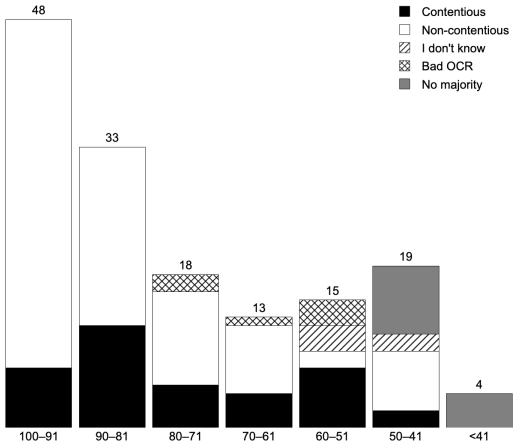

Figure 2: Nr. of samples grouped by percentage agreement, incl. majority votes of expert annotators.

Firstly, we score how many control questions each annotator got 'right', i.e. in line with the answer that experts unanimously gave. When we remove all data by annotators that got three or more control questions 'wrong' (30 annotators), the overall median α increases from 0.39 to 0.41. While this method does improve overall inter-annotator agreement, it has the downside that it is non-trivial to select meaningful control questions. Out of the five control samples, which we selected because experts unanimously agreed on them, three got high agreement among crowd-workers (93-95% of crowd annotators agreed). However, two got low agreement (40% and 56% agreed), stressing again the difference between experts and laypersons and the subjectivity of the task. Secondly, we calculate pairwise agreement between annotators using Krippendorff's α. When we remove all data by annotators who have an average pairwise $\alpha < 0.2$ (83 annotators), the overall median α of the corpus increases from 0.39 to 0.50. This suggests that careful selection of annotations by crowd workers can lead to results that are on par with results of experts. In the remainder of this paper we will base our analysis on the annotations of all annotators, i.e. without annotator selection.

Agreement per sample. We calculate percentage agreement per sample to investigate how often and in which cases annotators were able to reach a high level of agreement. Figures 2 and 3 display the number of samples grouped by percentage agreement, as well as their majority vote label, among experts and crowd workers, respectively. The control samples are excluded from the analysis.

We observe that expert annotators reach over 80% agreement in more than half of the samples (in 48+33 samples out of 150, Figure 2). Crowd workers reach over 80% agreement in more than 40% of the samples (in 553+494 samples out of 2,520 samples, Figure 3), and over 70% agreement in over 60% of samples. This confirms that agreement among expert annotators is higher than among crowd workers. In general, a reasonable level of agreement has been reached for over half of the annotated samples, in both the expert and crowd-sourcing groups. This is a positive result showing that even in a subjective and complex task such as ours, with low overall

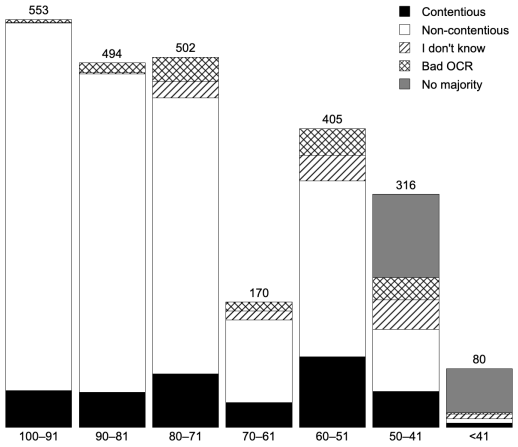

Figure 3: Nr. of samples grouped by percentage agreement, incl. majority votes of crowd annotators.

inter-rater agreement, it is possible to obtain a reliably annotated dataset of reasonable size using experts or crowd workers. We note that disagreement between annotators in subjective tasks is more than noise and can be valuable information [6].

Agreement per target word. Figure 4 shows all 91 target words in the corpus with the proportion of samples labelled as 'contentious', 'non-contentious', 'I don't know' or 'Illegible OCR' by the majority of annotators. Out of 91 target words, 6 were labelled by the majority of annotators as contentious in all the samples they appear in. For example, the target word 'bosneger' ('forest negro') was labelled as contentious by the majority of annotators in all eleven samples. Eleven target words were labelled by the majority as non-contentious no matter in which sample they appeared. Examples are 'allochtoon'('immigrant') and 'bediende' ('servant'). According to the majority, most target words (52) are contentious or non-contentious depending on the samples they are used in. To illustrate this, we look at the target words 'barbaren' ('barbarians'). It is contentious in 7 samples and non-contentious in 15 samples. All 7 annotators deem it contentious in sample *#619*, where it is used to describe people as uncivilized:
#619 "Als nu een rijke Chinees, dien het 'noodlot' onder de ***barbaren*** *(niet-Chineezen) gebracht heeft, naar een gemalin van zijn eigen stam verlangt <...>" / "If a rich Chinaman, who brought 'fate' to the* ***barbarians*** *(non-Chinamen) desires a consort of his own tribe <..>"*[10]

In sample *#271*, the word is used in a more metaphorical sense, referring to the concept of an uncivilised enemy, and it is marked as non-contentious by all 8 annotators:
#271 "Wij zullen ons wreken niet met de wapenen der ***barbaren****, maar met het geestes zwaard onzer propaganda. Iedere man en Iedere vrouw, wien het ernst is met het streven der sociaal-democratie, moet tot die wraak bijdragen." / "We shall take revenge, not with the* ***barbarians'*** *weapons, but with the spiritual sword of our propaganda. Every man and every woman, who is serious about striving for the social-democratic cause, must contribute to this revenge"*[11]

This shows that some terms seem to carry an inherent contentiousness or non-contentiousness, but for the majority of our target words the context is a deciding factor in the annotators' judgments. We see this also when we investigate the annotations per annotator: 220 annotators each individually annotated at least 1 target word both as contentious and non-contentious across different samples.

Some target words appear relatively often in samples for which annotators chose 'I don't know' (e.g. 'baboe' / 'nanny') or on which annotators did not reach a majority vote (e.g. 'mulat' / 'mulatto'). These cases are interesting as here contentiousness seems to be more dependent on the personal opinion of an annotator and less on a context or the word itself. Finally, some target words appear in samples that annotators flagged as 'Illegible OCR' (e.g. 'birma' / 'Burma').

4.2 Target and Context Analysis

With evidence for high subjectivity and contextuality of contentiousness in terms of agreement across annotations, we turn to computational linguistics. Specifically, we look at distributional semantics and the analysis of the influence of terms and their contexts' semantics on contentiousness. In doing so, we aim to show that contentiousness in ConConCor can serve as basis for meaningful statistical analyses and potentially for enriching AI with the cultural concept of contentiousness. To best serve both of these aims, we opt for simple and well-established methods to increase the chances of obtaining interpretable results.

Preprocessing. From this background corpus, we construct word embeddings by training a Word2Vec skip gram model. We use the Gensim API [12] and its default training parameters – embedding dimensionality of 100, based on a window of 5 words, ignoring words with fewer than 20 occurrences and training for 10 epochs – which yielded good enough results for our purposes of establishing baseline analyses and did not prompt us to experiment with other model architectures or parameter settings.

All control samples and annotations labeled 'I don't know' and 'Illegible OCR' are removed from the ConConCor corpus for statistical analysis. Samples are then labeled either 'contentious' or 'non-contentious' based on majority vote over either the annotators of a given sample or the samples of a given target term, depending on the experiment. In the former case, this results in 493 contentious samples (18.3%), 2,068 non-contentious samples (79.6%) and 127 samples with no majority (4.7%), for a total of 2,688 samples. In the latter case, this results in 2,870 contentious samples (29%), 6,474 non-contentious samples (67%), 281 no-majority samples (2%), for a total of 9,344 samples.

Target Analysis. Our first question in assessing the semantic nature of contentiousness is to what extent the target terms themselves predict their contentiousness. To this end, we fit a logistic regression to assess the predictive power of the target terms' distributional semantic features with respect to the annotators' judgments. As

[10] https://www.delpher.nl/nl/kranten/view?coll=ddd&identifier=ddd:010339930:mpeg21:a0007 (in Dutch) Last visited: 21/09/2021

[11] https://www.delpher.nl/nl/kranten/view?coll=ddd&identifier=ddd:010760348:mpeg21:a0127 (in Dutch) Last visited: 21/09/2021

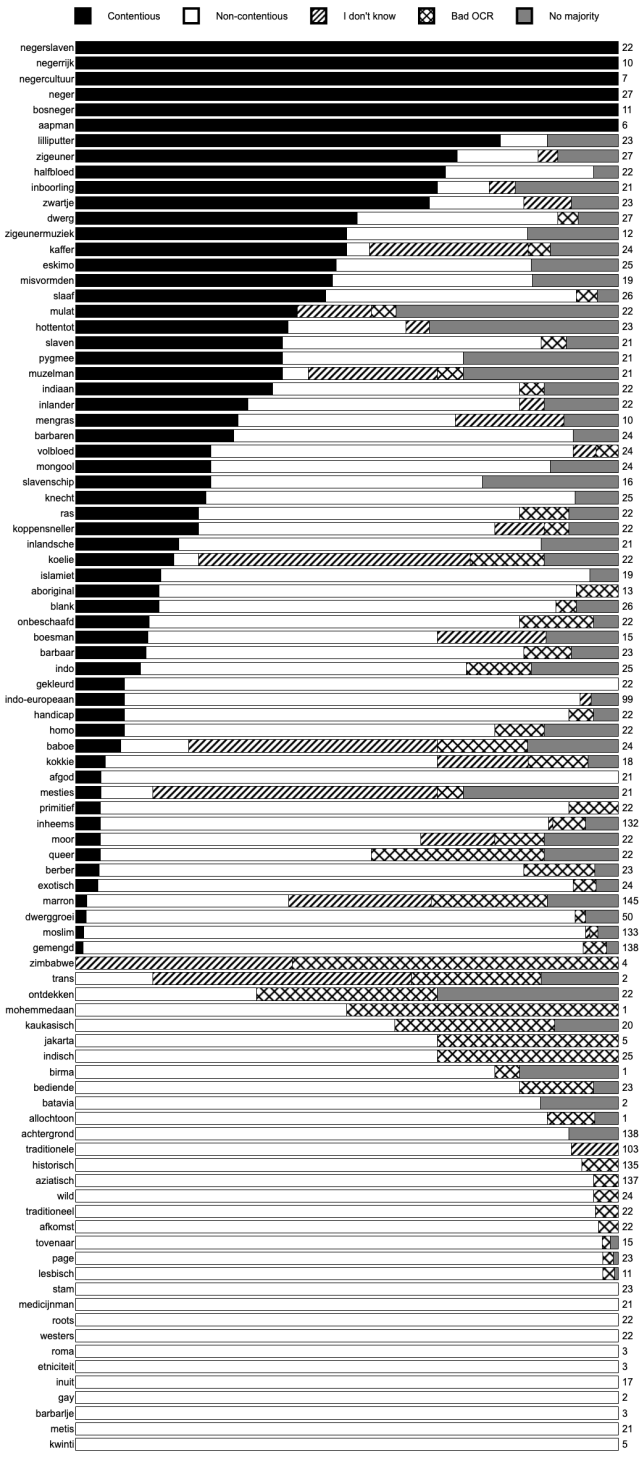

Figure 4: Proportion of majority votes of 91 target words. The number of samples per target word is on the right.

predictors, we use the Word2Vec embeddings described above and the majority contentiousness vote over contexts as the dependent variable.

We explicitly account for the influence of the annotators in the data by adding each sample's annotator as a categorical variable to the predictors of the regression model and, correspondingly, taking the majority vote over contexts *per annotator*. In this way, we do justice to the subjectivity of the task, as noted in Section 4.1, and additionally ensure that patterns identified by the regression model are solely due to relationships between target term embeddings and majority vote labels, rather than patterns in annotators' judgements. In probability theoretic terms, the regression model is thus equivalent to estimating a conditional distribution $P_a(contentious \mid embedding(target))$ *per annotator* a.

Similarly, we then construct a baseline model by counting the proportion with which an annotator annotated any target term in any context as contentious. This model is equivalent to the distribution $P_a(contentious)$ for each annotator a. Given this distribution, the baseline's prediction for a given annotator and any given target is then equivalent to majority voting, i.e. $P_a(contentious) \geq 0.5$.

The logistic regression is performed as a 30-fold cross validation on a 75-25% train-test split, given the low number of unique target terms and complexity of the prediction task. This implies that the regression's performance depends strongly on the specific train-test split, as reflected in the relatively large confidence intervals of the performance measurements provided in Table 2.

We summarise the results of the target analysis in Table 2: most saliently, neither model is able to identify a strong statistical relationship. This is unsurprising given the complex task and relatively little data (recall that ConConCor comprises 'only' 91 target terms). However, our main and important finding is that the regression model significantly outperforms the baseline on almost all measures; except for accuracy. Here the baseline is significantly better which can be attributed to the relatively high label imbalance in the data.

Furthermore, we find that a large number of the regression model's coefficients are significant (around 70%) which indicates that many of the dimensions of the embedding space contribute to whether a term is contentious. We can therefore tentatively state that contentiousness is a complex concept and determined by many semantic properties of a given term.

Given these results, we conclude that there is significant predictive power between the structure of word embedding spaces onto non-contextual contentiousness labels. To the extent that Word2Vec embeddings capture the underlying semantic properties of words, contentiousness is inherent in the semantics of words. This conclusion was not expected: on one hand, general-purpose word embeddings such as Word2Vec need to capture a multitude of semantic concepts, many of which have far greater importance than the concept of contentiousness. On the other hand, we think of contentiousness as a strongly contextual concept (in the sense that many words might be labelled as contentious in one context but not in another) which would lead to mainly meaningless labels when taking the majority vote over contexts. We hypothesise that both aspects, namely specialised word embeddings and contextual factors, could help models achieve better performance than our simple logistic regression model does.

To begin investigations into the contextuality of contentiousness, we perform an additional series of experiments using hierarchical clustering on contexts.

	Accuracy	Balanced Accuracy	Precision	Recall	no. Signficant Coeffs.	AUC
Majority Baseline	[0.70 0.73]	[0.54 0.55]	[0.58 0.60]	[0.12 0.13]	[11.41 21.97]	[0.54 0.55]
Logistic regression	[0.69 0.72]	[0.76 0.78]	[0.75 0.78]	[0.76 0.78]	[66.13 70.73]	[0.80 0.83]

Table 2: Performance measurements for statistical algorithms and baselines. All ranges given are 95% confidence intervals obtained from 30-fold cross-validation. *AUC* refers to the area under the curve of the receiver operating characteristic. The model outperforms the baseline on nearly all performance measures, indicating a clear, albeit not strong, signal for contentiousness in the semantic space.

t	Terms associated with contentious samples
10	neger, kleurling, uitvoerig, kleinen, slaaf, blank, veilig, landbouwer, stelen, zioh, snellen, ruw, totaal, indianen, gat, haar, ophouden, waarvoor, suriname, vader
20	neger, blank, suriname, vader, blik, gezicht, verdwijnen, indiaan, zes, rekenen, dik, dame, zoodra, geschikt, hulp, hopen, gevoel, zwart, bevinden, gelaat

Table 3: Top 20 terms significantly associated with majority vote contentious samples, by greatest association. Each of these context terms occurs together with a minimum, *t*, unique target terms. We observe tokens that may be offensive and with clear racial and ethnic connotations.

t	Terms associated with non-contentious samples
10	nlet, ai, probleem, bespreken, verdedigen, hemel, nader, aangenaam, co., vergadering, bijv., baron, ontnemen, zilver, aankondigen, zullen, eventueel, voorzitter, ned., uiterst
20	politiek, onderzoek, rusland, snel, band, ziel, tevens, openbaar, leger, amsterdam, zon, verband, streven, algemeen, kiezen, aandacht, vast, god, sluiten, europa

Table 4: Top 20 context terms significantly associated with majority vote non-contentious samples, by greatest association. In contrast to Table 3, the context terms contain no obviously offensive or racially-charged language.

Context Analysis. That the contentiousness of a target term, as represented by vectors in an embeddings space, can be predicted above majority baseline performance demonstrates that contentiousness is somewhat captured by distributional semantics. Hence, the target term alone, removed of context has inherent contentiousness, as we would expect. However, we also see examples in ConConCor of the same annotator annotating the same target word differently across contexts. By removing the target term and annotator as factors, we investigate the association between contentiousness and context.

We represent concepts in terms of groups of tokens related by an underlying concept. We used an agglomerative hierarchical clustering algorithm to suggest such clusters of related tokens. We assemble a hierarchical clustering matrix where, to keep the experiment computationally manageable, we reduce the embeddings to 2 dimensions via UMAP [10] prior to clustering based on euclidean distance. Upon inspection, this approach yielded sensible clusters.

If a cluster (of context terms, related by a concept) is not strongly associated with contentious or non-contentious samples, we would expect it to be approximately distributed between contentious and non-contentious samples in their respective proportions. that is, a cluster with no significant associations would be expected to occur approximately 18.3% of the time with majority vote contentious samples and 79.6% of the time with majority vote non-contentious samples. Hence, significantly contentious clusters were identified as those clusters whose proportion of occurrence with contentious samples was significantly greater than 18.3%, according to a 95% significance 1-tailed binomial test. Similarly, significantly non-contentious terms were identified as those terms whose occurrence rate in non-contentious samples was significantly greater than 79.6% according to a 95% significance 1-tailed binomial test.

Individual context terms can be considered as clusters of a single token. We identify 461 significantly contentious context terms, and 98 significantly non-contentious context terms. Table 3 and Table 4 highlight significant context terms that are also co-occurrent with a minimum number of unique target terms (t). A higher minimum broadly represent greater applicability to the full range of target terms and contexts in the corpus.

Examining more complex clusters of many tokens, we find that clusters at a medium depth, i.e. 12-15 levels deep from the leaf nodes, frequently yielded clusters with broadly interpretable underlying concepts. Crucially, at this depth, the corresponding clusters generally yield a reasonably sized set of context terms present in the corpus to which statistical tests can be applied. Suggested clusters with significant association with contentious or non-contentious samples, and with inferable underlying concepts, were then manually pruned of tokens that were not deemed strong enough indicators of the underlying concept. We observed a number of significant clusters which can be found in our GitHub repository.

5 USE CASES AND OUTLOOK

Whilst this project is focused firstly on investigating contentiousness in the context of cultural heritage collections, we see many other domains where this topic is relevant: search & recommender systems, social media analysis, and journalism, as it is in no organisation's interest to offend their audience. Our broader aim is to develop more culturally aware AI systems, which are implicitly or explicitly aware of the subtle and subjective complexity of human culture. As our work shows, contentiousness can be understood as a complex and deeply cultural and social concept, it provides an interesting challenge in the pursuit of culturally aware and sensitive AI, both for developing new methods and testing existing ones. We foresee the following immediate avenues of research:

- Can we, and if so, to what extent, measure how contextual contentiousness is? To what extent and in what way does the context of a term influence how contentious it is perceived?
- If we can identify contentious contexts using ConConCor, can these contexts be generalised and used to detect additional contentious terms and/or contexts?

- Contentiousness is about perspective, as seen in the complex patterns of agreement above. Based on these patterns, and identity variables about the participants, can we recognise and analyse different perspectives?
- To what extent can ConConCor be used to investigate bias in other types of data such as other cultural institutions' collections or contemporary newspapers?
- Can contentiousness be modelled in structured resources such as Knowledge Graphs to reason over it?

6 CONCLUSION

In this paper, we presented ConConCor: the Contentious terms in Context Corpus. A set of 2,715 unique samples from historical Dutch newspapers annotated with information on whether a particular target term in the context is contentious or not according to at least 7 annotators for most of our samples. The corpus contains two sets of annotations: 150 unique samples annotated by domain experts, and 2,570 unique samples annotated by crowd workers for a total of 21,800 annotations. We analysed the corpus along the dimensions of inter-annotator agreement, term statistics and context statistics.

Regarding research question 1, we see that while the overall inter-annotator agreement appears low, a major part of the samples in the corpus is annotated with a high percentage agreement between annotators. This stresses the value of having a large number of annotators per sample. Agreement among crowd workers is lower than among experts, but can be improved by filtering out underperforming annotators. An analysis of the annotations shows that in many cases, contentiousness is not a property of a word itself, but depends on the context it appears in: most words are in some contexts deemed contentious and in others non-contentious. There are, however, also words that are (non-)contentious regardless of context, or that seem to be a matter of personal opinion.

Research questions 2 and 3 concern the automatic detection of contentiousness: Our analyses show that contentiousness is a complex concept, making automatic detection difficult outside of context. We do, however, find clear signals in terms' semantics that relate to contentiousness. In answer to research question 3, we find that hierarchical clustering can identify certain concepts that seem indicative of contentiousness or non-contentiousness, which could be further developed in a tool to explore the contentiousness dimensions of galleries, libraries, archives, and museum collections.

With this research line, we hope to make cultural heritage collections more sensitive to different societal tendencies, leading to more diverse and inclusive collection descriptions.

AUTHOR CONTRIBUTIONS

RB: Stratified sampling of snippets from the Europeana newspaper Archive, building the annotation UI, coordinating the Prolific task, and compilation of the dataset. Main author of the ConConCor datasheet. Co-author of section 4.2. AN: Inter-annotator agreement analysis and data visualisation (Section 4.1), corpus and code documentation, contributed to the Introduction, Related work, and Creating ConConCor sections. VV: Implementation of text snippet sampling and of computational analyses, co-author of section 4.2. JvO: Introduction, overall manuscript editing. LH: Conceptual design, introduction, analysis of inter-annotator agreement, conclusions. MvE: Conceptual design, project coordination, related work, outlook, conclusions & overall manuscript editing.

ACKNOWLEDGMENTS

This work was funded by the EuropeanaTech Challenge for Europeana Artificial Intelligence and Machine Learning datasets, 'Culturally Aware AI' funded by NWO, and SABIO funded by the Dutch Digital Heritage Network. The authors would like to thank the Cultural AI Lab and KNAW HuC colleagues for their comments and annotations and the anonymous Prolific annotators. Special thanks to Mirjam Cuper (National Library of the Netherlands) for guiding KB and Europeana procedures, Lynda Hardman (CWI) for the suggestions on the article editing, and the anonymous reviewers for their constructive feedback.

REFERENCES

[1] Csilla E. Ariese. 2020. Amplifying Voices: Engaging and Disengaging with Colonial Pasts in Amsterdam. *Heritage & Society* 13, 1-2 (March 2020), 117–142. https://doi.org/10.1080/2159032X.2021.1901335

[2] Akanksha Bisht, Annapurna Singh, H. S. Bhadauria, Jitendra Virmani, and Kriti. 2020. Detection of hate speech and offensive language in twitter data using LSTM model. *Advances in Intelligent Systems and Computing* 1124 (March 2020), 243–264. https://doi.org/10.1007/978-981-15-2740-1_17

[3] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[4] Rosagemma Ciliberti, E. Fulcheri, Paolo Petralia, and Anna Siri. 2021. Sharing ethics of displaying human remains in museums. *Medicina Historica* 4, 3 (Feb. 2021), 1–8. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101736438&partnerID=40&md5=c0d03a2b83637982387afaf4437062ae

[5] Carol Ann Dixon. 2012. Decolonising the museum: Cité Nationale de l'Histoire de l'Immigration. *Race & Class* 53, 4 (March 2012), 78–86. https://doi.org/10.1177/0306396811433115

[6] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. (2018). arXiv:1808.06080

[7] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2020. Datasheets for Datasets. (2020). arXiv:1803.09010

[8] Kateryna Kaplun, Christopher Leberknight, and Anna Feldman. 2018. Controversy and Sentiment: An Exploratory Study. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence* (Patras, Greece) *(SETN '18)*. Association for Computing Machinery, 1–7. https://doi.org/10.1145/3200947.3201016

[9] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

[10] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3, 29 (Sept. 2018), 861. https://doi.org/10.21105/joss.00861

[11] Wayne Modest and Robin Lelijveld. 2018. *Words Matter: an unfinished guide to word choices in the cultural sector*. Technical Report. The National Museum for World Cultures (Tropenmuseum, Afrikamuseum, Museum Volkenkunde, Wereldmuseum). https://www.tropenmuseum.nl/en/about-tropenmuseum/words-matter-publication

[12] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (Valletta, Malta). University of Malta, 46–50. https://is.muni.cz/publication/884893/en

[13] Marcus Schulzke. 2012. Contentious Language: South Park and the Transformation of Meaning. *Journal of Popular Film and Television* 40, 1 (March 2012), 22–31. https://doi.org/10.1080/01956051.2011.624136

[14] Brenda Trofanenko and Avner Segall. 2014. *Beyond Pedagogy: Reconsidering the Public Purpose of Museums. Introduction*. SensePublishers, Rotterdam. https://doi.org/10.1007/978-94-6209-632-5_1

[15] Johanna Turunen and Mari Viita-Aho. 2020. Changing interpretations of the Gallen-Kallela Museum's Africa collection. *Historiallinen Aikakauskirja* 118, 4 (2020), 466–480. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098751112&partnerID=40&md5=796942175e7a5a6bcf711bef454e13ba