Routledge
Taylor & Francis Group

Kōtuitui: New Zealand Journal of Social Sciences Online

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tnzk20

Covering tangata whenua in Aotearoa: a big data exploration of print media and Māori

Justin Bonest Phillips

To cite this article: Justin Bonest Phillips (2023) Covering tangata whenua in Aotearoa: a big data exploration of print media and Māori, Kōtuitui: New Zealand Journal of Social Sciences Online, 18:3, 195-212, DOI: 10.1080/1177083X.2022.2122517

To link to this article: https://doi.org/10.1080/1177083X.2022.2122517

Published online: 18 Sep 2022.

KOTUITUI: NEW ZEALAND JOURNAL OF SOCIAL SCIENCES ONLINE
2023, VOL. 18, NO. 3, 195–212
https://doi.org/10.1080/1177083X.2022.2122517

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Covering tangata whenua in Aotearoa: a big data exploration of print media and Māori

Justin Bonest Phillips

School of Social Sciences, University of Waikato, Hamilton, New Zealand

ABSTRACT

A large body of academic research documents harmful media coverage of indigenous populations across the globe. New Zealand is no exception. Aotearoa's indigenous people, Māori, share similar damaging experiences, leading one major NZ media company to publish an apology for their historically poor depiction of tangata whenua. This paper adds to that wealth of evidence using the automated methods of machine learning to examine coverage of Māori in NZ print media. Across roughly 800,000 sentences – spanning over two decades of coverage – this research investigates print media discourses involving Māori at a mass scale while demonstrating the applicability of such tools for further research. The results replicate a collection of existing findings at large-N scale, further documenting problematic discussions of violence, political representation, and culture, among several other concerns. The novel approach also hints at complex, obscure relationships embedded within problematic language in Aotearoa's print media, identifying notions of division - both implied and otherwise - along with notable instances of resistance.

ARTICLE HISTORY
Received 13 April 2022
Accepted 2 September 2022

KEYWORDS
Māori; print media; newspaper coverage; automated topic analysis; word embeddings; Aotearoa New Zealand

Introduction

'Nō mātou te hē'; 'We're sorry';

'A team of reporters investigated how *Stuff* and its newspapers have portrayed Māori.

From the first editions to now, our monocultural lens means we haven't always fairly represented tangata whenua.

We've been racist, contributing to stigma, marginalisation and stereotypes of Māori.

Ka whakapāha

mātou ki te Māori

We apologise to Māori.' (*Stuff*, November 2020)

Stuff's front-page apology represents a historic acknowledgement from one of New Zealand's largest media institutions. The *mea culpa*, published on *Stuff*'s website and its regional papers across Aotearoa (Stevens 2020), recognises the pain and hurt inflicted by a domestic media upon tangata whenua – New Zealand's indigenous people.

CONTACT Justin Bonest Phillips justin.phillips@waikato.ac.nz

Various historic examples of damaging coverage accompanied the apology. From colonial era newspaper headlines describing Māori as inferior to modern-day framing of Māori advocates as dangerous radicals, the examples reflect ongoing and persistent language that plague media depictions of NZ's indigenous people.

Stuff's harmful examples are certainly valuable from a public awareness perspective, but ultimately they only tell us what we already know. As nearly half a century worth of research has extensively demonstrated, media language discussing Māori is concerning in numerous complex, and evolving ways. To draw again from the previous example, problematic historical media coverage could likely be described as blatantly racist, whereas modern discourses offer far more subtle language – though potentially no less pernicious.

Indigenous cultures across the world face similar concerns both historically and otherwise, reflecting the wealth of scholarship devoted to media's portrayal of native peoples across time. That body of scholarship is useful on two immediate grounds: first, it features as a resistance against acts of intentional and unintentional institutional level oppression. Second, it (can) operate(s) as an instructive tool to revise and amend such detrimental reporting in the future – though that requires journalists (among others) willing to listen and adapt. One year on from *Stuff*'s apology, their editors similarly acknowledge this commitment to change, claiming '[w]e've still got work to do' (Jacobs 2021).

Qualitative methodologies, such as discourse and content analyses, feature as the backbone for this critical research into media coverage of indigenous peoples. These time intensive approaches offer invaluable insights into the deeper meanings of words, phrases, images, and beyond. Qualitative research, therefore, provides a level of sophistication that satisfies both these purposes of resistance and instruction just discussed. By demonstrating the damaging – and often implicit – meanings behind media content, scholarship in this field exposes the real world pain inflicted upon indigenous peoples and their cultures.

Such qualitative methods, by their nature, are however limited by the availability of resources. Ultimately the number of researchers dedicated to manually code material, and the time devoted to such exercises narrows the capacity of these studies to a moderate number of observations – often small cross-sectional or longitudinal assessments of media both in Aotearoa and elsewhere.

A useful complementary method of documenting harmful media coverage sits in the quantitative realm. Here automated machine learning methods can provide a much larger picture of the scope of this problematic phenomenon. Such work, for example, details how indigenous peoples are discussed in media across time and space at vast scales. Contributions in this area are useful from several perspectives. On the one hand, they provide space to replicate existing qualitative work at large N scales – contributing to a more robust evidence base. On the other, such methods can also help reveal the widespread nature of hurtful media discourses; test for changes over time; and investigate more complex (mathematical) relationships in language – among other benefits.

This study continues that important line of complementary quantitative work by examining how media discuss Māori in New Zealand. Utilising the machine learning methods of topic analysis and word embeddings, this paper examines over 800,000

sentences discussing Māori, from 130 NZ print media outlets, published between 1995 and 2021. The results not only replicate previous qualitative findings across a variety of instances but also hint at more complex relationships at a mass scale.

Literature review

A wealth of literature exposes racism embedded – and distributed – in our media. The evidence is vast, spanning radio (Spearman 2021), newspapers (Richardson 2004), and television (Entman 1992) broadcast in countries across the globe. As social media expanded over the past decade, so too did research identifying the prevalence of racism on these new media platforms (Matamoros-Fernández and Farkas 2021). This valuable scholarship on racism and media exposes specific instances of negative media discourses targeting – for example – immigrants (van Dijk 1989), the (racialized) poor (Gilens 1999), and indigenous cultures among many other groups.

The effects of these media perpetuated racist discourses are similarly well documented. The immediate outcomes are perhaps the most obvious: targeted populations are dehumanised (Steuter and Wills 2010), subject to hate speech (Hawdon et al. 2017), and violence (Poynting and Perry 2007). Outcomes stretch well beyond immediate harms to marginalised groups, though. For nearly half a century, for example, we have known that (negative) opinions of minorities are often justified by purveyors with explicit references to media coverage (Hartmann and Husband 1974).

This distorted and damaging portrayal of minorities also tends to warp views within the broader electorate, with far-reaching consequences. To draw from one example, news media tends to racialize violent crime reporting in the United States, disproportionately characterising African American men as violent (Oliver 2003), persuading audiences to support more punitive remedies as a direct response to this coverage (Gilliam and Iyengar 2000). Such discourses additionally may prime citizens to judge candidates on their responses to crime (Valentino 1999). Likely in response to emerging electorate concerns – or more deep-seated ideological positions, political parties and candidates, therefore, capitalise with campaign promises (e.g. tough on crime stances) that often hold dire real world consequences for marginalised populations (e.g. three strikes laws; see Green et al. 2006).

Likewise, journalistic language has a covert tendency to more broadly divide society in the audience's minds, leading to hurtful societal conceptualisations of us and them (Van Dijk 2000). Such frames of division not only alienate some but also further strengthen dangerous in and outgroup mentalities (Mastro and Seate 2012; Roberts and Rizzo 2021). These are just some of the many potential reasons why assessments of media discourses on race are so deeply necessary.

Indigenous populations feature as an important subset of this work on media and race. Such investigations help detail how media's depictions of race may weaken indigenous people's sense of identity (Pietikäinen 2003), belonging (Love and Tilley 2013), and their right to customary claims (Wilkes et al. 2010). From a colonialism perspect, modern media, therefore, operates as a tool to subjugate and separate indigenous people from their land and their culture, in turn strengthening the state (and its people's) entitlement to place and resources. Research on media discourses of indigeneity

therefore understandably emerges in the aftermath of places and peoples affected by the reach of the imperial project (e.g. Australia, Canada, the US).

Māori are one of those affected people. Aotearoa New Zealand's tangata whenua (i.e. people of the land) centre as another target of harmful media coverage. Negative media frames – and broader discourses – of Māori mirror that of other indigenous peoples, with similar language involving crime (McCreanor et al. 2014); poverty (Beddoe 2014); disease (Rankine and McCreanor 2004); and work ethic (Nairn and McCreanor 1991). In some instances, these frames may also seek to explicitly question Māori indigeneity itself (Bayard 1998), speaking more broadly to the colonialism points just made.

Like explorations in other countries, researchers here have investigated how Māori are depicted in print (Wall 1997), radio (Adds et al. 2005), and television (Nairn et al. 2012) for decades. Collectively this work points to a number of problematic media discourses involving Māori. Barnes et al. (2012), for example, offer an extensive list of patterns in media coverage of Māori, recreated in Table 1. Many of those discourses are drawn from earlier work exposing Pakeha (white) constructed identities of Māori as represented in media (McCreanor 2005). This set of literature describing NZ-media discourses is particularly instructive, as it provides a set of frequent keywords deployed in this language. The relationship between these keywords features as a central research focus of this paper.

As Table 1 shows, media discourses of Māori in New Zealand similarly function within the us vs. them framework. Meaning, media language involving Māori treat Pakeha as a norm, with the 'other' extracting – undeserved – special rights, and privileges, for example. Where issues of politics or race emerge, Māori are depicted as hypersensitive or bent on stirring social and political disorder (e.g. via referencing Hobson's pledge upon the signing of Te Tiriti). Conceptualising these discourses (and others) within the context of Pakeha constructed identities helps narrow such language under a broader umbrella term, a useful exercise for the study to come.

Table 1. Patterns in media discourse – recreated from Barnes et al. (2012).

Pattern	Effect
Pakeha as norm	Constructs Pakeha as the ordinary normal citizen and culture of New Zealand
One people	New Zealanders are represented as a single culture in which all are to be treated the same
Rights	Individual Pakeha rights take precedence over collective Māori rights
Privilege	Māori are portrayed as having resources and access denied others
Ignorance and hypersensitivity	Pakeha offend Māori because of ignorance; Māori responses are unduly sensitive
Good Māori Bad Māori	Māori are seen as good or bad depending on the argument of the speaker; Pakeha are rarely described in this way
Stirrers	Those who challenge the social order are depicted as troublemakers who mislead others for their own ends
Māori crime, violence	Māori are seen as more likely to be criminal or violent than Pakeha
Māori culture	Māori culture is depicted as primited and inadequate for modern life, and inferior to Pakeha culture
Māori inheritance	Describes ancestry in fractions in a way that denies Māori concepts of whakapapa and self-identification
Māori resources	Critical of any return of significant resources to Māori as a denial of Pakeha rights to exploit such opportunities
Māori success	Small scale Māori projects that fit Pakeha business models and use Māori culture can be viewed as positive
Treaty of Waitangi	The Treaty is a historical document of little relevance to the contemporary setting – a barrier to development

While these discourses emerge in a variety of media (e.g. radio, television, and newspapers), print continues as the predominant medium for studying coverage of Māori in Aotearoa. This print focused research, along with complementary work on other mediums, provides the foundation from which broader thematic Māori discourses have been identified; Barnes et al.'s contribution above represents the comprehensive collection of such discourses. From a longitudinal perspective, research into print media's coverage of Māori also offers a well established legacy of evidence, spanning for at least three-quarters of a century (e.g. see Thompson 1953).

Though research on print media has revealed a number of important ways that Māori are discussed in NZ media, one primary finding sits at the broader categorical level. For example, scholarship emerging from the Whariki Research Group (WRG) – and thereafter – found that Māori are often covered within one of 16 topics: 'Arts, Business, Crime, Education, Fisheries, Foreshore and seabed, Health, Land, Māori/Pākehā relations, Military, Political representation, Religion, Sport, Social issues, Tikanga (items about Māori customs and procedures) and Treaty of Waitangi' (Rankine et al. 2014). Researchers have also found similar categorical results from separate content analyses of NZ media (Barnes et al. 2005), suggesting only minor fluctuations in the type of topics exist over time. These are notable findings because it suggests where Māori are explicitly discussed in media, such coverage is likely restricted into limited categories – a point made by the WRG academics.

This narrowing of Māori focused coverage may further reinforce the us and them stereotype discussed above. According to this logic, the language of 'us' transcends categories (e.g. spans the newspaper), but Māori are othered by NZ media into predefined generic topics. Given its foundational position in the literature, identifying and assessing these generic categories of Māori coverage features as one primary focus in this paper.

Moving beyond where Māori are covered (and categorised) in media, the other focus of this paper explores how Māori are discussed in media. In an effort to address this secondary question, many of the major studies in this space have narrowed their explorations of Māori focused discourse to generic frames emerging within specific topics, including gender and crime (Deckert 2020), health (Rahiri et al. 2018), sport (McCreanor et al. 2010), business (McCreanor 2011), and politics (Sullivan 2008), to name just a few. This rich body of literature provides invaluable insight into the framing of Māori in Aotearoa New Zealand's media – in print and otherwise. To date, the bulk of such explorations have relied primarily on small-N samples to provide necessary time and space for deeper, sophisticated qualitative readings of media texts. Though larger studies certainly exist, this field of scholarship in Aotearoa has yet to see contributions at mass scales using big data analysis tools. Such a contribution would help further advance this body of work – just as it has others – in new and valuable ways.

For example, natural language processing (NLP) – the automated processing of texts via machine learning – has numerous applications ranging from the arguably mundane (e.g. email spam filters) to the life saving (e.g. revealing human trafficking: Dubrawski et al. 2015, tracking online extremism: Johansson et al. 2017). At its core, NLP tools permit researchers to explore texts at a mass scale, relying on the machine to categorise, organise, and identify language patterns from any number of sources (e.g. books, newspapers, social media posts). From a media perspective, NLP toolkits have been used to

identify common news frames (Ylä-Anttila et al. 2018), and topics (Liu et al. 2020), among others.

The advance is twofold here, as the researcher somewhat removes themselves from the analysis (a potential wide-ranging methodological critique of content analyses), while simultaneously expanding the number of observations to potentially address sampling bias. The benefits reach well beyond methodological contributions though. NLPs can help researchers identify patterns in discourse that might have otherwise been obscured. For example, from exposing differences in how men and women communicate (Park et al. 2016) to comparing the evolution, language, and ideological positions of political parties (Hjorth et al. 2015), such toolkits have a wide array of applications.

This is partly why automated tools are now being deployed in research areas involving media and indigeneity. Wallace's (2021) exploration of Canadian print media's framing of its indigenous people represents the most recent and directly relevant example. That study utilised word similarity algorithms (e.g. dendrograms) to thematically categorise print media language across three decades of Canadian newspaper coverage, exposing the prevalence of stereotypical portrayals of indigenous peoples as impoverished and criminal – among several other problematic frames. While print media commands the focus of the current article, similar automated methods have be extended to larger and more complex datasets like social media. Moeke-Pickering et al. (2018) accomplish just this, by using an analysis package from Sysomos, a social media analytics company in Toronto, to identify the types of discourses emerging during the #MMIW (missing and murdered Indigenous women) movement on Twitter. That social media research, spanning over 100,000 tweets, showed important instances of contestation and resistance against the Canadian Government's Truth and Reconciliation Commission.

These examples highlight how automated methods offer real value in the indigenous research context, though there are some noteworthy limitations. For example, conceptually, frames differ from topics and the relationships between individual words. The beginning of Entman's (1993) well known attempt to clarify the paradigm of framing across disciplines is instructive here: 'To frame is to select some aspects of a perceived reality and make them more salient in a communicating text … '. While automated tools can tell us what is being discussed (e.g. topics), and they can identify important relationships between words (e.g. embeddings), they can never substitute for manually exposing how texts define problems, interpret causes, morally evaluate, nor propose treatments (i.e. Entman's four framing characteristics). The contribution of such tools is therefore not in substitution of qualitative methods, but in complementary analysis: that is, natural language tools can uniquely reveal patterns that are consistent with the use of – for example – frames within communicating texts.

Acknowledging the complementary nature of automated approaches, the potential applications of such machine learning tools to media coverage of Māori requires brief discussion here. From a broader perspective, an investigation of this nature would add to the (aforementioned) growing body of scholarship of media and indigeneity, using natural language analysis. Such a study would potentially give an expansive picture of how Māori are covered in media; something that can be used both to test existing theory and expand knowledge in new areas. NLP informed research would therefore not only considerably expand the existing evidence set involving Māori-focused

discourse but may additionally serve as a platform for more detailed, narrowed investigations going forward – much like the sophisticated qualitative work previously outlined. This is the purpose of the present study.

Methodology

This study involves the automated analysis of approximately 500,000 articles, published across 130 print outlets in New Zealand. Data was collected via the Newztext search function, housed by The Knowledge Basket – Te kete o te wānanga. Newztext represents New Zealand's largest digital repository of print news, containing NZ-based newspapers owned by both Stuff (previously Fairfax) and New Zealand Media and Entertainment. The publication time period spans the entire dataset: from January 1995 to May 2021.

Following methodological approaches from previous studies (e.g. see Sullivan 2008), the search parameter simply looked for any article in the Newztext dataset that contained the keyword 'Māori'. The resulting half-million articles were then downloaded, processed, and input into a separate database.

For this study, over 800,000 sentences were extracted from the Newztext articles that include the keyword Māori. Those sentences were then parsed using standard natural language analysis procedures (e.g. removing stopwords and stemming where necessary).

The (parsed) 800,000 sentences were then analysed using 2 machine learning methods: (1) topic analysis; (2) word embeddings. Automated topic analysis (1) separates content into categories based on the relationship and frequency of words. Word embeddings (2) place words into vector form, which allows the researcher to more comprehensively explore meanings and relationships between words. Both tools operate as replication tests of existing work and more detailed investigatory tools.

The analysis works as follows. First (1), using MALLET (MAchine Learning for LanguagE Toolkit), sentences were classified into individual topics, via the Latent Dirichlet Allocation algorithm. Two topic models were created: a WRG model, named after the Whariki Research Group that proposed 16 Māori focused topics in NZ print media; and an ideal model constructed via standard diagnostic analysis. Put more simply, the machine was instructed to classify sentences into 16 categories for the WRG model; diagnostics were then run to construct an additional model with an optimal number of categories (i.e. the 'ideal' model). Finding this ideal model involved several statistical tests, including comparing the optimal conference (UMASS) and log likelihood harmonic means, while also qualitatively considering topic consistency (i.e. subjectively interpreting potential models). This diagnostic process ultimately found six topics to be optimal. Both models (i.e. WRG: 16 topics; and 'ideal': 6 topics) result from 200,000 iterations through Mallet's LDA algorithm.

This first-stage analysis tests for surface level (replicatory) comparisons between findings from prior literature (i.e. the WRG model) and the evidence from this dataset, while providing the basis to examine consistencies and discrepancies of language between (and within) categories.

The second (2) test uses word embeddings to more deeply investigate the meanings (and relationships) between individual words used in print media. Using fastText, an open source word to vector programme written by Facebook, words from a large corpus can be translated into vectors which in turn can be used to mathematically

detect relationships between words (e.g. synonyms, word associations). This is particularly useful, as the tool can expose problematic word relationships in print media that qualitative researchers have previously identified (e.g. 'bad' Māori and 'bludgers'). Barnes et al.'s comprehensive list of anti-Māori frames is instructive here (Table 1), given they list quite a few keywords found across decades of qualitative research.

These two automated tools, topic analysis and word embeddings offer complementary avenues to explore how Māori are discussed at mass scales in New Zealand print media, both as a replicatory and investigatory tool.

Before proceeding, though, two remaining points require elaboration. First, from an epistemological perspective, the present project assumes that language can be assessed objectively via mathematical tools; this is not a universally accepted position. Postpositivist academics, for example, would obviously contest this notion. Indigenous researchers would also further express concern over this ontological foundation (Wilson 2001), but the process of stripping language of context and subjectivity in relation to cultural meaning would also feature as a primary objection. For example, one relevant Māori customary perspective defines data, in this case language, as taonga – something with cultural significance (Hudson et al. 2017). The automated tools discussed above transform data in a mathematical sense, but according to customary positions they also – critically – disconnect words from culture, and to some (e.g. see Taiuru 2018) might even extinguish important associations between mauri (life force) and the data connected to both the living and dead. This is the first point requiring acknowledgement and reflection, given the current paper's approach.

A final word on indigenous data sovereignty is similarly required in this section. As decades of research meticulously demonstrates, Māori are indeed framed in Aotearoa in a number of concerning ways. As a Pakeha researcher, it is necessary that I acknowledge how I engage with and seek to transmit knowledge in indigenous academic areas and recognise those that hold true sovereignty over this important communicative space. This is essentially at the heart of the indigenous data sovereignty concept. Professor Smith (2021) makes a related argument involving research that potentially '"steals" knowledge from others and then uses it to benefit those who "stole" it'. Words such as these demand inclusion here before proceeding.

Results

Figures 1 and 2 provide comparison clouds for the two topic models created in this study. Words in these clouds represent differences between one topic from another (i.e. a frequent word in one topic but absent in others). The size of words in these figures indicates the magnitude of difference between the topics. Meaning, large words – closer to the centre in the figures – appear far more often under the selected topic than elsewhere.

As discussed in the literature review, prior research suggests that Māori are largely discussed within 16 topics in NZ media. The WRG model parameters split the Newztext dataset into 16 categories, representing a surface level, automated replication of this prior qualitative work. Figure 1 does show many of the same topics identified by qualitative researchers over a decade ago, this time using machine learning.

Here, for example, the WRG model identifies coverage of art, crime, education, health, political representation, sport – all topics that coincide with findings from previous

Figure 1. WRG model comparison cloud.

research. The model also appears to combine some previously conceptually separated topics (e.g. land and fisheries into natural resources); splits some into multiple topics (e.g. Māori-Pakeha relations appears across numerous topics); and offers new topics altogether: local (government) and comparison. A breakdown of the distribution of topics can be found in Table 2.

The 'comparison' topic requires brief elaboration before proceeding to the ideal model. In this category, Māori are positioned in direct comparison to other population groups. For example, the language in this topic routinely compares Māori outcomes (e.g. health) against other groups, often Pacific peoples.

Prior research has noted the distribution of these topics has slightly changed over time, and the results in Table 2 reflect this. For example, the 'health' topic fluctuates between 13% and 7% across the 2004 and 2007 datasets. The WRG model in the current paper finds a middle ground, as it assigns roughly 10% of all sentences containing the Māori keyword under the health topic. Given the dataset for this study spans decades,

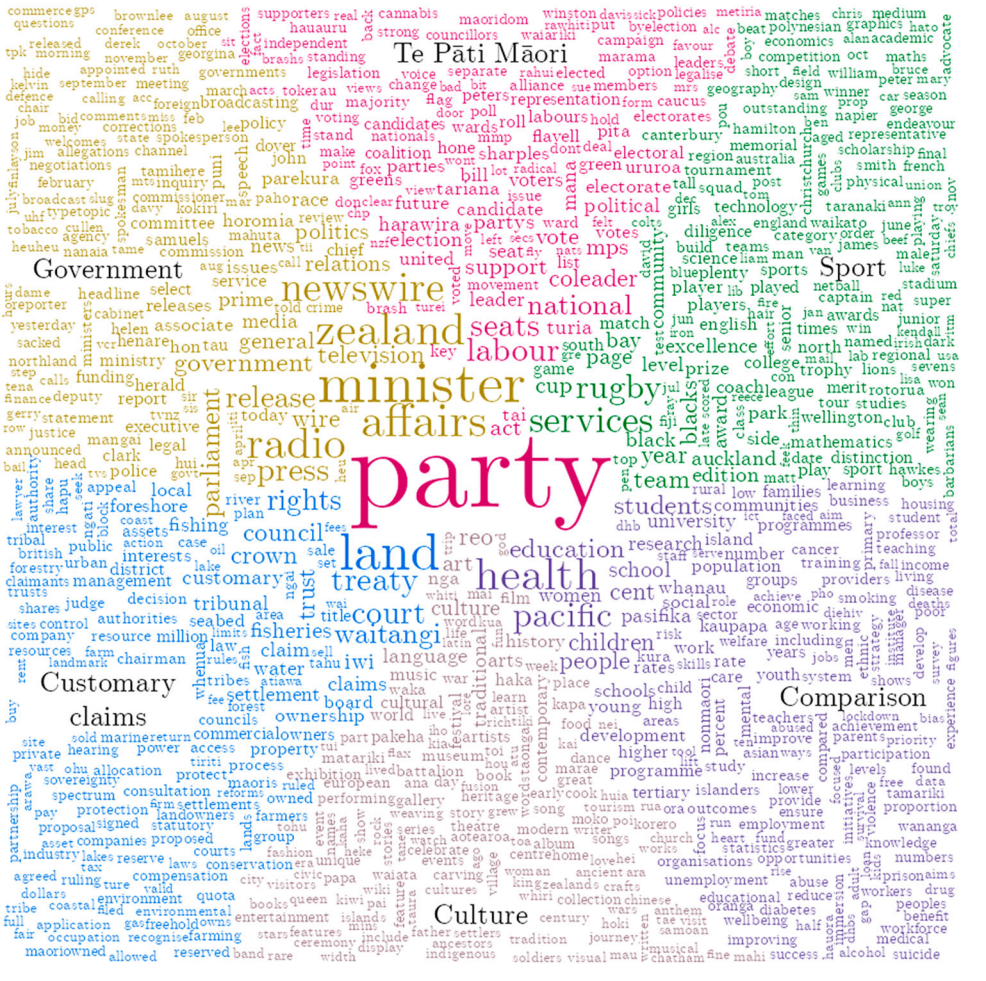

Figure 2. Ideal LDA model comparison cloud.

including the years studied in aforementioned research, 10% seems to be a reasonable allocation for the health topic.

The proportion of other topics, like arts, crime, sport, and political representation (considered Te Pāti Māori in WRG) appear to roughly overlap with distributions in findings of previous research. However, there were also some notable differences – and absences.

For example, the WRG model appears to simplistically divide broad discussions of Māori resources and customary rights. Previous research extrapolated such topics into a conceptually distinct set of individual categories: such as land, fisheries, foreshore and seabed, and the Treaty. This is another valuable finding because it suggests media coverage of these customary and resource-based topics deploys overlapping language.

Overlapping language is a core feature of the ideal LDA model. Here the algorithm – based on diagnostics – indicates the optimal topic model only contains six topics: comparison, culture, customary claims, government, Te Pāti Māori, and sport. Meaning, the other topics identified in the WRG model are assimilated into one of six categories. The

Table 2. Topic distributions.

	Topic models: WRG 2004 (1)	Rankine et al. 2014 (2)	WRG LDA (3)	Ideal LDA (4)
Culture				23%
Arts	4%	17%	4%	–
Religion	2%	0.6%	–	–
Tikanga	–	4%	–	–
Comparison				24%
Business	7%	10%	–	–
Education	9%	10%	5%	–
Financial probity	2%	–	–	–
Health	13%	7%	10%	–
Customary claims			8%	23%
Fisheries	4%	1%	–	–
Foreshore and seabed	7%	0.2%	–	–
Land	10%	6%	–	–
Te Tiriti	14%	6%	–	–
Politics			6%	11%
Political representation	18%	6%	–	–
Military	–	2%	–	–
Te Pāti Māori			8%	16%
Māori–Pakeha relations	9%	9%	–	–
Sport	1%	6%	5%	10%
Crime	–	12%	10%	–
Observations	353	858	830,212	830,212

Note: For the purpose of comparison, the 'politics' topic is interpreted as 'government' in the Ideal LDA model.

model, therefore, identifies overlapping language between a variety of seemingly independent categories.

This overlapping language is additionally notable because it identifies problematic similarities discussed by prior research. Take natural resources, for example. The language from this WRG-identified category overwhelmingly (∼77%) sits within the customary claims in the ideal topic model. This is not a completely unexpected finding since *prima facie* discourse involving natural resources (e.g. be it customary claims or existing uncontested titles) is likely to overlap.

However, the words that most accurately differentiate the customary claims topic from others (e.g. 'land', 'treaty', 'court', 'waitangi'), suggest any overlap that exists with Māori operated natural resources mirrors an overarching narrative of Māori led resource dispute elsewhere. Seen in this light, anti-Māori discourses that Barnes et al. identify, like Māori rights, privilege, resources, and the Treaty are likely all embedded within this 'customary' topic – hinting that similarities in language between words themselves is a worthwhile exploration.

More deeply exploring similarity in language is precisely the focus of the next NLP tool used in this study. Here the fastText algorithm is used to mathematically identify a word's nearest neighbours (i.e. via a cosine similarity range of 0–1), which can be used to directly test the presence for keyword associations that prior research has identified. In this phase of the test, I survey the twenty nearest neighbours for problematic relationships previously identified.

For example, according to prior research, violence often features when Māori are discussed in the context of crime. The words violent (cosine similarity: 0.52) and violence (0.51) are indeed some of the nearest neighbours to the word crime in this

dataset. Like detailed elsewhere in the world (e.g. the crime suspects script, see Gilliam and Iyengar 2000), and in NZ, physical characteristics of offenders – including skin colour – operate as a primary neighbour to the keyword 'offender': light-skinned (0.63), dark-coloured (0.58), olive-skinned (0.58), dark-skinned (0.58). Assailant (0.57) also represents another nearest neighbour of the 'offender' keyword, further reflecting the relationship between skin colour, violence, and crime in sentences discussing Māori in NZ media. The following exemplifies some of these embedded relationships:

> Maori violent crime a 'fact of life': A respected academic is blaming Maori for high rates of violent crime that have helped give Hamilton the country's second-highest number of 'three strike' offenders. (Marlborough Express, 9 May 2013)

> Claims by a New Zealand researcher that Maori carry a 'warrior' gene that makes them more prone to violent and aggressive behaviour has again put the spotlight on the over-representation of Maori in violence statistics. (The Press, 10 August 2006)

> When most of the violent offences and other crimes are committed by young Maori, why should the rest of us have any sympathy for racist stirrers. (The Daily News, 18 April 2000)

These types of mathematical relationships can be applied well beyond crime coverage, however. For example under the rights and privilege discourses identified by Barnes et al., the keywords of democracy and equality are often used in instances where Māori are discussed as infringing on the rights of society (i.e. Pakeha). Some of the concerning word embedding neighbours for 'democracy' include: 'antidemocratic' (0.65), 'undemocratic' (0.60), 'lgnz' (0.46), 'disenfranchise' (0.42), 'tyranny' (0.41), 'seats' (0.41). To those familiar with NZ political discourse, these nearest neighbour words routinely embed in discourses criticising Māori political representation – and election – at both our Parliamentary level (which includes the Māori 'seats' as allocated by the Māori electoral roll), and the local level (where local councils are increasingly adopting Māori wards). The following selection highlights these criticisms of Māori representation at both the local and national level:

> Maori seats a mockery of democracy. (The Independent, 1 August 2001)

> The imposition of a non-consulted, undemocratic and racially divisive Maori board would be a costly and totally wrong move. (Nelson Mail, 4 April 2012)

> Of course, they are misguided; clinging to anti-democratic and sexist customs will eventually sound the death knell for Maori culture. (The Dominion, 2 February 1997)

> It is clearly undemocratic that voters on the Maori roll can have their cake and eat it by voting for Labour with their MMP party vote and electing Maori Party MPs in reserved electorates in which no one else can vote. (New Zealand Herald, 20 September 2005)

Anti-Māori language involving Māori's perceived ('hyper') sensitivity regarding ('harmless') Pakeha ignorance also appears prevalent in word embeddings. Here the concerning nearest neighbours for 'sensitive' and 'ignorance' include: 'offensive' (0.39), 'unsuitable' (0.38), 'promulgated' (0.38), 'intolerance' (0.79), 'arrogance' (0.54), 'misrepresentation' (0.50), 'illinformed' (0.50), 'prejudices' (0.48). Interestingly, these words are used in the dataset both as anti-Māori discourses, and in instances of resistance –

another important finding worth noting. Examples help fully demonstrate the wide spectrum of this contested space:

> When will pakeha stop denying their faults by blaming the problem on Maori paranoia and victim mentality, and begin to dialogue in reasonable debate with a preparedness to admit wrongdoing and to change their superior attitudes and offensive behaviour. (The Daily News, 24 May 2001)

> A tongue-in-cheek sign that has thrilled tourists for years at one of Rotorua's top tourist attractions has been removed after being labelled a grotesque racial stereotype. (The Daily Post, 3 August 2018)

> I applaud Maori culture – however, I have had a gutsful of reading/hearing about it as if, before every way we turn and act, we must consider their culture in case we offend. (The Press, 19 June 2009)

> He admitted 'in retrospect' that it could be insensitive to Maori, and has apologised if anyone was offended, but claims when he showed it to his Maori friends 'they laughed'. (Southland Times, 23 November 2016)

> Efforts to intrude Maorification into university protocol, ceremonial, etc (which by implication suggests former existence) is a falsification and often rude and offensive to those of other 'cultures'. (Waikato Times, 17 July 1996)

That final claim, one of expressly questioning the history and legitimacy of Māori culture, introduces the final word most worth discussing here: culture. Culture's nearest neighbours (generally) reflect seemingly positive values: 'uniqueness' (0.56), 'language' (0.46), 'richness' (0.51), 'traditions' (0.56), 'heritage' (0.52), 'identity' (0.52), 'juxtaposition' (0.50). This collection of apparently complimentary words highlights two additional, potentially problematic discourses, though. First, as has previously been argued, there is a danger in conveying Māori culture as 'being frozen at the point of first contact' (Barnes et al. 2012). The nearest neighbours similarly position culture as largely pointing back in time, instead of forwards. Though, notable examples of resistance exist here as well:

> Tourists will enter through a time tunnel which will tell the story of Maori culture. (The Press, 7 September 2002)

> Maori culture is not by nature frozen, as the vibrant explorations by today's artists, singers and writers are demonstrating. (The New Zealand Herald, 1 February 2000)

The keyword 'juxtaposition's relationship with culture further highlights the language of comparison previously identified by the topic models as well. According to this conceptualisation, Māori sit merely as a facet of comparison, while potentially further embedding us vs. them discourse – in this case directly related to the culture of other minorities, be it Pacific or otherwise. In fact, the keyword also provides space to critique Māori in stereotypical ways against a complimentary (implied) vision of 'us', though here too examples of resistance exist:

> I think it is such an almost surreal juxtaposition of the very English, very royal and precise, with the Maori – kind of tribal and very emotional. (New Zealand Herald, 28 September 2016)

> The focus on closing the gaps tends to juxtapose groups in opposition to each other, but really what New Zealand has failed to grasp is that what's good for Maori is good for the country and that it's actually in the national interest that all those gaps that continue are remedied. (New Zealand Herald, 16 March 2014)

Discussion and conclusion

In this paper, machine learning tools were applied to over 800,000 sentences containing the keyword Māori. These sentences represent a substantial amount of Aotearoa New Zealand's print media coverage of Māori, spanning over 130 publications across 2 decades. Building on the foundation of sophisticated qualitative research involving NZ media's depiction of Māori, this paper finds areas both of replication, and problematic divergence with prior research.

From the replication perspective, both automated topic analysis and word embeddings in the current paper reinforce academic observations made over half a century. NZ-based print media does indeed appear to cover Māori within a narrow set of topics, leaving the 'other' to either implicitly or explicitly occupy broader space across the newspaper. The language used to discuss Māori reflects similar prior findings of concern; the well documented (global) relationship between media reporting of crime, race, and violence perhaps represents the most obvious well explored example. Language portraying Māori as hyper sensitive in response to ('harmless') Pakeha ignorance represents another, among several other areas of overlap. There were also clear – problematic – relationships between how Māori culture and political representation were discussed in print during this time. Machine learning, in this case, therefore provides further evidence of demonstrably harmful language deployed in NZ-based print media across decades.

The tests in this study also expose important differences with prior research. For example, the ideal topic model – automatically – compressed all discussion of Māori land, sea, and other resources into a single customary claims category. What this suggests is the machine fails to distinguish – in a meaningful fashion – a difference between Māori ownership of natural resources and broader customary claims. If a natural language processor sees fit to combine these topics, then it is little surprise that some audiences fail to see differences either. Such a merger might similarly overlap between several anti-Māori discourses ranging from rights, privilege, to inheritance, resources, and Te Tiriti.

The ideal model further identifies Māori as a feature of comparison with other cultures and ethnicities. Viewed through this lens, Māori outcomes in health, education, and otherwise merely serve as a metric to how well (or poorly) they perform in comparison to others. This operates in two concerning ways. For one, it often explicitly positions Māori in contrast to other minorities (e.g. Pacific, Asian). Implied within these direct comparisons with minorities sits a separate category: us. Put another way, the comparison category appears not only to reinforce us vs. them framing but simultaneously juxtaposes minorities within the 'them' context.

These areas of conformity and divergence potentially represent new avenues for research pressing forward. Again, take the problematic relationship between crime, race, and violence. Methods like those used in this paper can explore whether this type of news reporting abated (or evolved) over time. Similarly, longitudinal work is

also particularly useful in areas where change is to be expected: one obvious contribution might test *Stuff*'s commitment to improve their coverage in the future (Stevens 2020). Machine learning methods would also prove useful to test similar institutional promises elsewhere.

Future researchers should acknowledge the limitations of solely relying on natural language processing tools to accomplish these goals, though. Where possible, academics should look to bolster their automated analyses with context and qualitative evidence. The discussion of culture above provides one such justification: without more detailed readings of the texts one might miss important examples of resistance to problematic discourses, as were present in this data. Complementing automated methods with a wealth of qualitative historical research offers one such method, but pairing in-depth qualitative readings with these tools may provide many more valuable insights.

Identifying categories of Māori focused discourse across other (new) media is likely another area for future contribution. For example, social media spaces operate both as forums for observing institutional and individual behaviour. A tweet from Radio New Zealand involving Māori – for example – can continue to tell us how media cover tangata whenua, but responses to that tweet can expose quite a bit more about the Māori focused discourses of individuals both in Aotearoa and elsewhere.

That relationship between institutions and individuals speaks to the larger point of a paper such as this. Hurtful coverage of Māori can spill into both spheres: it can alter the opinion and behaviour of audiences, while serving as a foundation for resisting much needed, institutional level changes. Documenting how media discuss Māori, therefore, represents one tool that can help shift that relationship in a more positive direction. In line with decades worth of prior research, the hope is that this paper similarly helps echo an important call for journalists to reform the way media covers tangata whenua in Aotearoa.

Acknowledgements

I would like to thank Dr Waikaremoana Waitoki and Dr Damian Scarf for their encouragement and support for this paper. My thanks also go to the reviewers and the editorial team for their thoughtful comments and suggestions throughout the revision process.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This research was supported by a Ministry of Business, Innovation and Employment Endeavour Research Programme, 'Working to End Racial Oppression' (UOWX2002).

ORCID

Justin Bonest Phillips http://orcid.org/0000-0002-2984-4472

References

Adds P, Bennett M, Hall M, Kernot B, Russell M, Walker T. 2005. The portrayal of Maori and te Ao Maori in broadcasting: the foreshore and seabed issue. Wellington: New Zealand Broadcasting Standards Authority.

Barnes AM, Borell B, Taiapa K, Rankine J, Nairn R, McCreanor T. 2012. Anti-Maori themes in New Zealand journalism – toward alternative practice. Pacific Journalism Review. 18(1):195–216.

Barnes AM, Gregory M, McCreanor T, Nairn R, Pega F, Rankine J. 2005. Media & Te Tiriti o Waitangi 2004. Auckland: Kupu Taea, Media and Te Tiriti Project.

Bayard D. 1998. White turnips and mythical Moriori: combatting folk-linguistic and folk-anthropological myths in the popular press. New Zealand English Journal. 12:6–20.

Beddoe L. 2014. Feral families, troubled families: the spectre of the underclass in New Zealand. New Zealand Sociology. 29(3):51–68.

Deckert A. 2020. Indigeneity matters: portrayal of women offenders in New Zealand newspapers. Crime, Media, Culture. 16(3):337–357. doi:10.1177/1741659019873771.

Dubrawski A, Miller K, Barnes M, Boecking B, Kennedy E. 2015. Leveraging publicly available data to discern patterns of human-trafficking activity. Journal of Human Trafficking. 1(1):65–85. doi:10.1080/23322705.2015.1015342.

Entman RM. 1992. Blacks in the news: television, modern racism and cultural change. Journalism Quarterly. 69(2):341–361.

Entman RM. 1993. Framing: toward clarification of a fractured paradigm. Journal of Communication. 43(4):51–58.

Gilens M. 1999. Why Americans hate welfare: race, media, and the politics of antipoverty policy. Chicago (IL): University of Chicago Press.

Gilliam FD, Iyengar S. 2000. Prime suspects: the influence of local television news on the viewing public. American Journal of Political Science. 44(3):560–573. doi:10.2307/2669264.

Green EGT, Staerklé C, Sears DO. 2006. Symbolic racism and whites' attitudes towards punitive and preventive crime policies. Law and Human Behavior. 30(4):435–454. doi:10.1007/s10979-006-9020-5.

Hartmann P, Husband C. 1974. Racism and the mass media.

Hawdon J, Oksanen A, Räsänen P. 2017. Exposure to online hate in four nations: a cross-national consideration. Deviant Behavior. 38(3):254–266. doi:10.1080/01639625.2016.1196985.

Hjorth F, Klemmensen R, Hobolt S, Hansen ME, Kurrild-Klitgaard P. 2015. Computers, coders, and voters: comparing automated methods for estimating party positions. Research & Politics. 2(2):205316801558047. doi:10.1177/2053168015580476.

Hudson M, Anderson T, Dewes TK, Temara P, Whaanga H, Roa T. 2017. "He Matapihi ki te Mana Raraunga" – conceptualising big data through a Māori lens.

Jacobs M. 2021. Our truth, Tā Mātou Pono: one year on, what's changed and what still needs to. Stuff. https://www.stuff.co.nz/pou-tiaki/300465776/our-truth-t-mtou-pono-one-year-on-whats-changed-and-what-still-needs-to.

Johansson F, Kaati L, Sahlgren M. 2017. Detecting linguistic markers of violent extremism in online environments. In: Johansson F, Kaati L, Sahlgren M, editors. Artificial intelligence: concepts, methodologies, tools, and applications. IGI Global; p. 2847–2863.

Liu Q, Zheng Z, Zheng J, Chen Q, Liu G, Chen S, Chu B, Zhu H, Akinwunmi B, Huang J, et al. 2020. Health communication through news media during the early stage of the COVID-19 outbreak in China: digital topic modeling approach. Journal of Medical Internet Research. 22(4): e19118. doi:10.2196/19118.

Love T, Tilley E. 2013. Temporal discourse and the news media representation of Indigenous-non-Indigenous relations: a case study from Aotearoa New Zealand. Media International Australia. 149(1):174–188. doi:10.1177/1329878X1314900118.

Mastro D, Seate AA. 2012. Group membership in race-related media processes and effects. In: Giles H, editor. The handbook of intergroup communication. Routledge; p. 379–392.

Matamoros-Fernández A, Farkas J. 2021. Racism, hate speech, and social media: a systematic review and critique. Television & New Media. 22(2):205–224. doi:10.1177/1527476420982230.
McCreanor T. 2005. 'Sticks and stones may break my bones … ': talking Pakeha identities. In: Liu JH, editor. New Zealand identities: departures and destinations. Wellington: Victoria University Press; p. 95–128.
McCreanor T. 2011. Maori business news in the mass media. Sites. 8(2):32–56. doi:10.11157/sites-vol8iss2id173.
McCreanor T, Rankine J, Barnes AM, Borell B, Nairn R, Gregory M, Kaiwai H. 2010. Māori sport and Māori in sport: mass media representations and Pākehā discourse. AlterNative: An International Journal of Indigenous Peoples. 6(3):235–247. doi:10.1177/117718011000600304.
McCreanor T, Rankine J, Moewaka Barnes A, Borell B, Nairn R, McManus A-L. 2014. The association of crime stories and Maori in Aotearoa New Zealand print media. Sites. 11(1):121–144. doi:10.11157/sites-vol1iss2id240.
Moeke-Pickering T, Cote-Meek S, Pegoraro A. 2018. Understanding the ways missing and murdered Indigenous women are framed and handled by social media users. Media International Australia. 169(1):54–64. doi:10.1177/1329878X18803730.
Nairn R, Barnes AM, Borell B, Rankine J, Gregory A, McCreanor T. 2012. Māori news is bad news: that's certainly so on television. MAI Journal. 1(1):38–49.
Nairn RG, McCreanor T. 1991. Race talk and common sense: patterns in Pakeha discourse on Maori/Pakeha relations in New Zealand. Journal of Language and Social Psychology. 10 (4):245–262.
Oliver MB. 2003. African American men as "criminal and dangerous": implications of media portrayals of crime on the "criminalization" of African American men. Journal of African American Studies. 7(2):3–18. doi:10.1007/s12111-003-1006-5.
Park G, Yaden DB, Schwartz HA, Kern ML, Eichstaedt JC, Kosinski M, Stillwell D, Ungar LH, Seligman MEP. 2016. Women are warmer but no less assertive than men: gender and language on Facebook. PLoS One. 11(5):e0155885. doi:10.1371/journal.pone.0155885.
Pietikäinen S. 2003. Indigenous identity in print: representations of the Sami in news discourse. Discourse & Society. 14(5):581–609. doi:10.1177/09579265030145003.
Poynting S, Perry B. 2007. Climates of hate: media and state inspired victimisation of Muslims in Canada and Australia since 9/11. Current Issues in Criminal Justice. 19(2):151–171. doi:10.1080/10345329.2007.12036423.
Rahiri J-L, Gillon A, Furukawa S, MacCormick AD, Hill AG, Harwood MLN. 2018. Media portrayal of Māori and bariatric surgery in Aotearoa/New Zealand. New Zealand Medical Journal. 131(1479):72–80.
Rankine J, Barnes AM, McCreanor T, Nairn R, McManus A-L, Abel S, Borell B, Gregory A. 2014. Content and source analysis of newspaper items about Māori issues: silencing the 'natives' in Aotearoa? Pacific Journalism Review. 20(1):213–233. doi:10.24135/pjr.v20i1.195.
Rankine J, McCreanor T. 2004. Colonial coverage: media reporting of a bicultural health research partnership. Journalism. 5(1):5–29. doi:10.1177/1464884904040397.
Richardson JE. 2004. (Mis)representing Islam: the racism and rhetoric of British broadsheet newspapers. Amsterdam: John Benjamins.
Roberts SO, Rizzo MT. 2021. The psychology of American racism. American Psychologist. 76 (3):475–487. doi:10.1037/amp0000642.
Smith LT. 2021. Decolonizing methodologies: Research and indigenous peoples. London, United Kingdom: Bloomsbury Publishing.
Spearman L. 2021. Low hanging fruit: how sports talk radio hosts discuss racism. Communication & Sport. 9(6):934–953. doi:10.1177/2167479520916800.
Steuter E, Wills D. 2010. 'The vermin have struck again': dehumanizing the enemy in post 9/11 media representations. Media, War & Conflict. 3(2):152–167. doi:10.1177/1750635210360082.
Stevens M. 2020. Stuff's apology to Māori – our truth, Tā Mātou Pono. Stuff. https://www.stuff.co.nz/pou-tiaki/our-truth/300165985/stuffs-apology-to-mori–our-truth-t-mtou-pono.
Sullivan A. 2008. The Māori party and the media: representations in the mainstream print leading to the 2005 election. Pacific Journalism Review. 14(1):131–149.

Taiuru K. 2018. Why data is a Taonga: a customary Māori perspective.
Thompson RHT. 1953. Māori affairs and the New Zealand press. The Polynesian Society. 64 (4):366–383.
Valentino NA. 1999. Crime news and the priming of racial attitudes during evaluations of the president. Public Opinion Quarterly. 63(3):293–320. doi:10.1086/297722.
van Dijk TA. 1989. Mediating racism: the role of the media in the reproduction of racism. In: Wodak R, editor. Critical theory. Vol. 7. Amsterdam: John Benjamins; p. 199–226. doi:10.1075/ct.7.15dij.
Van Dijk TA. 2000. New(s) racism: a discourse analytical approach. In: Cottle S, editor. Ethnic minorities and the media: changing cultural boundaries. Buckingham: Open University Press; p. 33–49.
Wall M. 1997. Stereotypical constructions of the Maori 'race' in the media. New Zealand Geographer. 53(2):40–45.
Wallace R. 2021. "Warriors don't sleep til noon": colonial rhetoric and the framing of Indigenous recipients of welfare in Canadian print news, 1990–2015. Politics, Groups, and Identities. 9 (2):300–318. doi:10.1080/21565503.2019.1629311.
Wilkes R, Corrigall-Brown C, Ricard D. 2010. Nationalism and media coverage of Indigenous people's collective action in Canada. American Indian Culture and Research Journal. 34 (4):41–59. doi:10.17953/aicr.34.4.05q6170n1m987792.
Wilson S. 2001. What is an indigenous research methodology? Canadian Journal of Native Education. 25(2):175–179.
Ylä-Anttila T, Eranti V, Kukkonen A. 2018. Topic modeling as a method for frame analysis: data mining the climate change debate in India and the USA. SocArXiv. doi:10.31235/osf.io/dgc38.