LSE Media and Communications

Media@LSE MSc Dissertation Series

Editors: Bart Cammaerts and Lisa Derand

Hate in the Mainstream: Proposing a 'Keyness-Driven' Framework to Surface Toxic Speech in the Public Domain

Pica Johansson

Published by Media@LSE, London School of Economics and Political Science ("LSE"), Houghton Street, London WC2A 2AE. The LSE is a School of the University of London. It is a Charity and is incorporated in England as a company limited by guarantee under the Companies Act (Reg number 70527).

In the interests of providing a free flow of debate, views expressed in this paper are not necessarily those of the compilers or the LSE.

ABSTRACT

'Toxic speech' is a subset of harmful language prevalent on social media, conceptually close to 'hate speech', but seen as a less aggressive form of abusive language compared to it's harsher neighbor. These discourses have seemingly been normalized over time, yet, to-date, there is no research which attempts to systematically study this phenomena longitudinally, ex post. Research shows that white supremacists have made opportunistic use of social media's unregulated state to 'slip hate into the mainstream', infiltrating the public domain with toxic discourse -- rather than explicitly hateful language -- as a tactic to broaden support for their political agenda. White supremacist discourse therefore makes for an excellent case study to pilot this papers contribution: a 'keyness-driven' framework which builds on automated textual analysis and natural language processing, used to surface a decade of toxic speech on Twitter, in turn modeled to uncover 10 years of latent discursive trends. Using a sizable corpus consisting of 3.3 million tweets and 59,000 posts from the white supremacist forum 'Stormfront', representing discourse on immigration from 2011 to 2020, this research exemplifies how to 1) systematically discover harmful narratives that share similarities with those from explicitly hateful groups and 2) track how they change over time. By using this framework it was revealed that toxicity on Twitter using language similar to that used by white supremasicsts increased by 28% in the past decade. In addition, the analysis indicates that 2016 - 2017 was a critical moment for this discourse after which a number of the dominating narratives dramatically shifted. These findings indicate that a 'keyness-driven' framework shows promise for identifying moments at which hateful discourses shift and evolve, thus offering an essential methodological contribution to the study of harmful language.

Hate In The Mainstream: Proposing a 'Keyness-Driven' Framework to Surface Toxic Speech in the Public Domain

Pica Johansson

INTRODUCTION

> "The white race is an endangered rare species of great ape. It can be argued that whites are being extra tolerant at this point in time and the posterity of the white race depends on not overflowing immigration with migrants from non-white countries."

Prejudice and intolerance manifest through language - the toxicity associated with these traits lingers in between words, permeating a statement rather than using singularly hateful terms. The preceding quote, which originated on Twitter, illustrates this point. It is not directly hateful, and represents an opinon, therefore it cannot be removed from the platform – but evidently this class of language is undesirable in the public domain. Imagining a cacophony of posts with similar sentiments quickly draws concern to the normalization of such speech. Toxic speech is a pernicious social problem with consequences for public life – it should also be treated as such, yet is barely addressed by academia.

The study of toxic speech is in its infancy, and is currently dwarfed by the substantial body of research which concerns 'hate speech', despite the former being more widespread. The limited research in the field is inherently a methodological challenge, as current algorithmic models do not represent this category of speech (Nario *et al*., 2020). It is seen to be particularly challenging to capture due to its covert, veiled or highly contextual nature, with only a small number of recent studies attempting to further the field, but none attempting to retroactively map toxic narratives and their evolution.

An increased focus on toxic language would entail a much needed shift away from reactive content moderation, to proactive, trend-driven analysis to better anticipate when online discourses go awry. As expressed by UN Secretary General Antonio Guterres, addressing harmful discourse does not entail limiting freedom of speech, rather, it means keeping hate speech from escalating into something more dangerous (United Nations, 2019). He continues, "Hate is moving into the mainstream – in liberal democracies and authoritarian systems alike. And with each broken norm, the pillars of our common humanity are weakened." (ibid).

Hate In The Mainstream: Proposing a 'Keyness-Driven' Framework to Surface Toxic Speech in the Public Domain

Pica Johansson

That hate has moved into the mainstream has in particular been said about the resurgence of the far-right, their increased popularity in the 2000's being attributed to highly strategic use of social media and palatable expressions of hate (Meddaugh & Kay, 2009), resulting in a shift in the so-called 'overton window' of what opinions are regarded as 'acceptable' in the public domain (Williams *et al.*, 2019). The public expression of toxic language gives it legitimacy and both reinforces and perpetuates social biases regarding the groups it most commonly targets (Munger, 2016; Sap *et al.*, 2020). A central concern is therefore that this language is echoed on social media, becoming subtly normalised within public discourse.

Using natural language processing (NLP) and automated textual analysis, this paper addresses this gap in literature and contributes with a novel approach by identifying harmful language on two fronts. First, this research provides a 'keyness-driven' framework to surface toxic language retroactively. This framework is then used to model latent topics in toxic language resembling that of white supremacists, and explore how this discourse has evolved over the past decade.

Subsequent sections of this paper are structured as follows. The literature review focuses on theoretical and methodological contributions which illustrate the challenges inherent to the study of harmful language. This is followed by a statement of the conceptual framework used for this research, after which the research objectives are presented. The methodological chapter presents the rationale for selecting 1) 'keyness' analysis as the method which underpins the 'keyness-driven' framework to surface toxic content and 2) 'structural topic models' for the analysis of toxic content over long periods of time. This section also outlines the data collection, preprocessing steps and model application for full reproducibility. Next, the findings are presented in relation to the two research questions, after which these are discussed ahead of concluding.

THEORETICAL CHAPTER

Literature Review

'Harmful language' serves as an umbrella term for the field of study which concerns the use of online language which can cause harm or distress to an individual or society (Munezero *et al.*, 2013). A comprehensive bibliometric study describes the field as highly interdisciplinary and consisting of many interrelated and overlapping disciplines, such as linguistics, law, and social sciences (Tontodimamma *et al.*, 2020). That the field draws on a multitude of disciplines has also impeded academics to reach an agreement on what hate speech, by far the most researched topic within it, in practice entails - many even approaching their studies with a "I-know-it-when-I-see-it" attitude (Poletto *et al.*, 20121: 488). What is agreed upon, however, is that labelling harmful language of all sorts is often a subjective task because it is contingent on the context of the utterance, as well as prevailing social norms (Saleem *et al.*, 2017). The inconsistent labelling of harmful language is a major problem for the area of research this paper draws: literature which aims to put forth quantitative methods using natural language processing to detect harmful language.

The forthcoming literature review encompasses three main scholarly fields which deal with these automated quantitative methods, first reviewing the relevant literature which concerns 'hate speech', next that on 'toxic speech' and finally reviewing scholars who have specifically studied the discourse of the far-right.

Hate Speech

Quantitative analysis of harmful rhetoric was first studied in the late 1990's (see Spertus, 1997), but the exponential increase of publications began twenty years later, once social media had taken root and the internet had established itself as the bedrock for public discourse (Fortuna & Nunes, 2018; Waqas *et al.*, 2019; Tontodimamma et a l., 2020). Since then, nearly all studies on harmful language concern hate speech, an area of research which is slowly reaching saturation.

Hate speech research takes countless directions, capturing a range of closely related linguistic concepts including '*verbal abuse*' (Nobata *et al*., 2016), *'insults'* (Mahmud *et al*, 2008; Sood *et al*., 2012), *'threats'* (Fiser *et al*., 2017; Hammer, 2017), *'profanity'* (Coats, 2021) , *'vulgarity'* (Xiang *et al*., 2012, Holgate *et al*, 2018), *'obscenity'* (Mubarak *et al*., 2017) , as well as language that can be seen as '*bullying*' (Xu *et al*., 2012) *'derogatory'* (Nobata *et al*., 2016) or *'inciting violence'* (Zampieri *et al*., 2019). Studies also concern hate speech directed towards specific groups, for instance on language that expresses misogyny (Fersini *et al*., 2018a, Guest *et al*.; 2020), homophobia (Davidson *et al*., 2017), sexism (Waseem & Hovy, 2016; Davidson *et al*., 2017), racism (Tulkens *et al*, 2016), islamophobia (Törnberg & Törnberg, 2016; Vidgen & Taha, 2019), and antisemitism (Bjola & Manor, 2020).

These various areas not only have different foci: they also stress different aspects of hate speech. Some emphasize the writer's intention, others the linguistic form, others still the potential effect on the victim (Poletto *et al*. 2020). This is to underscore that hate speech identification research has taken many shapes and has been studied from countless angles, to the degree that scholars have -- rather than publishing new models -- instead attempted to advance the field by consolidating existing models and resources (Noever, 2018; Vidgen & Derczynski, 2020; Risch *et al*., 2021) as well as developed benchmarking frameworks to compare models (Röttger *et al*., 2021).

Toxic Speech

As a reaction to the oversaturated field of hate speech, a small number of scholars have directed attention to the more subtle form of harmful language that this research situates itself within: toxic discourse. It is important to clarify that within the broader field of harmful language toxicity and hate speech have been used as synonymes (alongside other terms such as online violence and online hate) (Chandrasekharan *et al*., 2017). This paper adopts the views of scholars that have attempted to study this tier of harmful language, as well as Google's Jigsaw Initiative (Jigsaw, 2021), by regarding toxicity and hate speech as two distinct concepts

-- emphasising the importance of capturing speech that does not qualify as outright hate speech -- but is harmful nonetheless.

Toxic speech is defined as speech which is unsettling, disrespectful, or in other ways uncomfortable to the degree that someone would want to leave the conversation (Borkan *et al.*, 2019, Nario, 2020). It has been described as being subtle (Price *et al*. 2020; Gilda *et al.*, 2021), covert (Nario *et al.*, 2021), veiled (Han & Tsvetkov, 2020) and 'likely to offend' (Kolhatkar, 2019), as well as at times indicating implicit social bias (Sap *et al.*, 2020). *"If minorities in this country were truly oppressed you wouldn't have so many people desperately pretending to be one."* - is exemplified as a toxic statement because it is harmful without containing the overtly aggressive language typically targeted by hate speech models (e.g. *"you are an idiot"*) (Nario *et al*. 2020: 18).

To-date literature on toxic speech has been exploratory, and generally aims at publishing resources to improve the quality of online conversations. For instance, Kolhatkar and colleagues (2019) were the first to publish an annotated dataset aiming to improve the quality of online discussion, emphasising the shift in focus from moderating the single instances of hate speech, to understanding and improving the bulk of online conversations. But toxicity was only one of four categories they annotated, and the corpus remains limited to comments to online news articles. Although, compared to the studies by Price et al.(2020) and Nario *et al.* (2021), Kolhatkar *et al*. considered the widest range of 'toxic speech' as recognized by their scale: 'very toxic' - harsh, offensive or derogatory language directed towards a person, 'toxic' - for example ridicule, and 'mildly toxic' - language only considered toxic by some people[1], thereby providing helpful distinctions within the concept.

Price *et al*. (2020) build on Kolhatkar *et al.*'s (2019) study by using their dataset to study a smaller subset of comments which they deem 'unhealthy', in contrast to outrightly threatening, abusive or in other ways distinctly harmful language. This research contributes with a novel typology and outlines six models, each capturing a subcategory of 'unhealthy' comments: those that are hostile, antagonistic, provocative/trolling,

[1] 'Very toxic' would in this case likely be deemed hate speech according to the definition this paper follows

condescending/patronizing, dismissive or sarcastic, but conclude by acknowledging that there is no way to know how exhaustive these categories are, and what other categories of speech might still be missing.

Sap *et al*. (2019) contribute with an exceptionally comprehensive and novel study on social biases in language which moves the focus from what is *explicitly said* to the *implied meanings* that frame people's judgements about others. They provide both a new conceptualization to model *"the pragmatic frames in which people project social biases and stereotypes onto others"* as well as the publish the Social Bias Inference Corpus: 150,000 social media posts which are mapped across 34,000 implications of language about 1,000 demographic groups (Sap *et al*., 2019: 1). This contribution brings social implications of harmful language to the fore, by comparison to previous studies which exclusively focus on effectively modeling harmful language.

With a slightly different focus, but similar in spirit to the three aforementioned papers, Han and Tsvetkov (2020) empirically prove that the industry leading toxicity classifier by Google overwhelmingly misclassified subtle toxicity. In reaction to Han and Tsvetkov (2020), a group of researchers from Google attempt 'covert toxicity' modeling to capture 'microaggressions' , 'obfuscation', 'suggestive emojis', 'sarcasm/humor' and 'masked harm' (Nario *et al*. 2021). This veiled toxicity contains use of language that may not be immediately obvious, and can use dark humor, code-words or emojis to convey a hateful message. The exploratory study shows early promise for using machine learning classifiers to capture toxicity, although also acknowledge the challenge of using crowdsourced annotation for the niche-topic. Fundamentally, research suggests that the language 'missed' by classifiers is typically not understood to be harmful to all readers but requires context and most of the time targets a specific group (Poletto *et al*., 2020, Nario *et al*. 2021), as illustrated by Table 1. In isolation, none of the words are indicative of toxic speech, but are so when placed in context.

Target	Example of toxic discourse
Immigrants	"Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home." (Burnap & Willians, 2016: 10)
Women	"Literally the only thing that matters for choosing a wife - the woman's chastity." (Han & Tsvetkov, 2020: 7736)
Muslims	"Wearing a Burkha doesn't feel very #UK" (Vidgen & Yasseri, 2020: 69)
Jews	"What's the difference between boy scouts and Jews? Boy scouts come back from camp." (Nario et al., 2020: 14)

Table 1: Examples of contextually targeted toxic speech

Toxic Discourse of the Far-Right

Language of the far-right has been described as characterised by having attributes of targeted, subtely harmful, discourse, as a result of their expressions of deeply ingrained bigotry being key to their often white supremisict, anti-immigrant or anti-semitic ideology (Klein, 2012). Holt *et al.*'s (2020) comprehensive content analysis of eight far-right extremist forums lends empirical support for this, findings indicating that the most prevalent ideological sentiments expressed in users' posts involved anti-minority comments, often conveying negative sentiment overtly, or by using slurs. Similarly, Warner & Hirschberg (2010) find demeaning language referring to minorities and targeted groups often make use of the veiled language of stereotypes, as each stereotype is context dependent and has a "language of its own, with one-word epithets, phrases, concepts, metaphors and juxtapositions", conveying strong bias at best and malice at worst (Warner & Hirschberg, 2010: 19).

Other research has emphasized how the far-right's *"toxic yet effective messages"* which convey cultural intolerance and racial superiority are becoming normalized through social networks, *"such that the lines that once separated racism from political extremism are harder to distinguish."*

(Klein, 2012: 428). Klein (2012) has coined this practice 'information laundering', conceptualising how racial hate speech has become legitimised as the far-right's social networks have expanded (visualized in Appendix B). Thus, hate speech what was once prevalent in fringe movements of society, has steadily become a toxic discourse considered part of the mainstream. Similarly, Bonilla-Silva (2002) argues that racism has taken new form in the post-civil rights era, characterized by a *"slipperiness and apparent nonracialism"*(ibid: 41), and Meddaugh & Kay (2009) suggest that online white supremacist rhetoric appears to have become *"less virulent and more palatable to the naive reader"*, by tatical use of using implicitly hateful messaging (ibid: 251).

Researchers have also proffered that the normalization of harmful discourse by the far-right has led to a shift in the 'Overton Window' (Marantaz, 2019). A shift in the 'Overton Window' signals that an issue or position now falls within the realm of acceptable discourse, typically due to it being extensively used within public domains and/or adopted by politicians (Marantaz 2019; Williams *et al*., 2019). The far-right are argued to have been engaging in these practices for decades, Williams *et al*. (2019) going as far as saying that their role has been pivotal in shifting xenophobic online discussions to where they are today, starting with the worlds first hate site 'Stormfront'. The theory attributes this evolution to the steady expansion of social platforms which have allowed content to travel far beyond fringe sites and circles of the devout, thus unwittingly enabling *"purveyors of bigotry to infiltrate into mainstream spaces of public discourse"* (Klein, 2012: 427).

This literature review has identified the difficulties in the detection of harmful language and outlined concerns specifically pertaining to the far-right's use of toxic language. Despite researchers' concerns with toxic language entering the public domain, few empirical studies track the evolution of this language, an important step in identifying root causes and solutions. In response to this gap in the literature, this paper proposes a framework which uses longitudinal data from the white supremacist forum 'Stormfront'[2] to retroactively surface toxic posts from a mainstream domain (Twitter) and explores methods to analyze latent topics and trends from the past decade. As such, this study aligns itself most closely with the positionings of scholars who urge that research on harmful language should be more nuanced, and that resources must be redirected from real-time moderation to proactive intervention (Schmidt & Wiegand, 2017; Jurgens *et al.*, 2019; Sap *et al.*, 2020).

Conceptual Framework

The conceptual framework which the present research draws on consists of two important parts. The first makes use of Poletto *et al.*'s (2020) typology of harmful language which sets various concepts of harmful language in relation to one another. This typology provides the conceptual background for how the research perceives the larger field of harmful language. Next, the rationale for undertaking the research is described as echoing Schmidt and Wiegand's (2017) call for a paradigm shift to 'anticipatory governance' of social media, an conceptual backing which situates this paper normatively.

Typifying Hateful Language

As the vast majority of studies on harmful language concern hate speech, it follows that other areas of harmful language have been overlooked. Poletto *et al.*'s (2020) contribution to the field is distinct to research on harmful language as it attempts to typify the ambiguity related to the phenomenon (Figure 1). This excellent typology brings much needed clarity to how the concepts relate to one another, but also makes clear that harmful language has nuances. Hence, whilst hate speech is an instance of harmful language, there also exists harmful language that does not qualify as hate speech. Similarly, the typology takes into account that the harmful language directed towards specific groups, such as racism, is not always explicitly hateful, but

[2] Prior studies of white-supremacist discourse exist. To date, Stormfront has been used to examine overt hate (Figea *et al.*, 2016, Gibert *et al.*, 2018) and to explore expressions white supremist ideology (Wong *et al.*, 2015; Perry & Scrivens, 2016), in particular the manifestation of xenophobia, such as its antsemitic (Dentice, 2018; Dentice, 2019) and racist beliefs (Meddaugh & Kay 2009**;** Klein, 2012; Figea *et al.*, 2016). Little, therefore, is known about the properties of toxic language in this speech, if it has changed, and even less on how language similar to it prevails on social media platforms. Brief methodological details on this research can be found in Appendix A.

often toxic. Most importantly for the present research, the positioning of these concepts makes evident the fact that when research focuses exclusively on hate speech, one disregards other instances of harmful language, not least toxic speech -- being more prevalent than hate speech -- yet severely understudied.

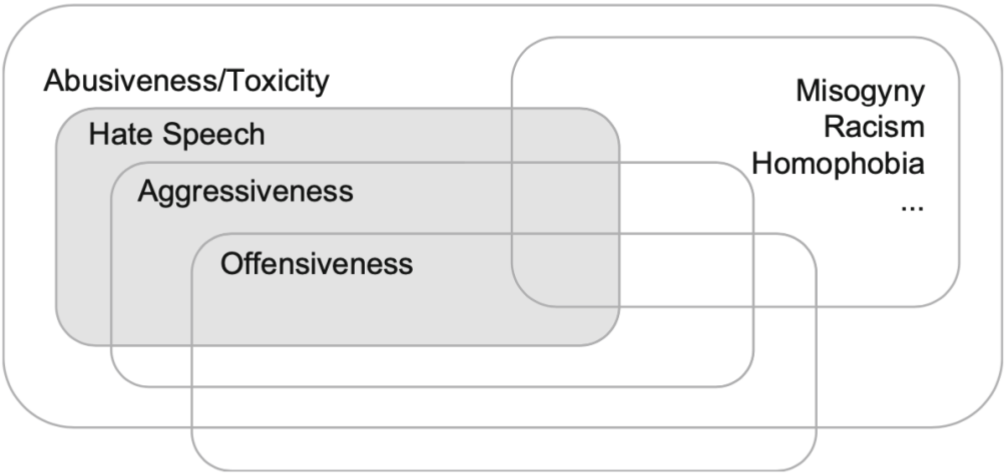


Figure 1: Relationship between hate speech, toxic speech targeted towards groups, and related concepts (Poletto *et al*., 2020: 482)

Anticipatory Governance of Social Media

Recalling the earlier statements by Antonio Guterres, Guterres proclaimed that hate had moved into the mainstream, and that taking action against these discourses should be seen as keeping hate at bay -- not limiting free speech (United Nations, 2019). Within academic scholarship, this can be equated to what Schmidt and Wiegand (2017) describe as a move towards 'anticipatory governance' of social media. Redirecting the focus of research would mean a shift from detecting individual, isolated hateful comments, to instead gaining insight into the overall proportion of negative commentary over a certain time in an attempt to identify and counter larger systems of harmful language (ibid).

This marks a significant divergence from the current approach to regulating social media. Under the current model, governments pressure tech companies to improve algorithmic content moderation of their platforms, entailing that the focus is exclusively on explicitly hateful language and close to real-time monitoring of harmful language (Lucas, 2014). Case-by-case moderation and keeping an eye on the most recent harmful narrative weeds out a proportion of the most grave and harmful language prevalent online, but little is known about the pockets of language that harbour these toxic discourse before they escalate. A paradigm of 'anticipatory governance' thus calls for new avenues of research which keeps abreast of the development of harmful language over time, and tracks their development holistically.

Research Objectives

The below section outlines the research objectives and contributions of the present research.

Research Objectives & Contributions

The primary motivation for this research is to explore the potential for a novel method to retroactively map toxic narratives, given the absence of studies concerning the evolution of toxic language on mainstream platforms. Using natural language processing and automated textual analysis, this paper addresses the aforementioned gap in literature and provides an original contribution to academia on harmful language on two fronts.

The research firstly contributes a unique framework, which systematically surfaces toxic posts which contains linguistic similarities with white supremacists. It also contributes to early stage findings which highlight latent topics within the identified instances of toxic language, exploring how such discourse has changed over the course of a decade.

Research Questions

The paper examines the following research questions:

I. To what extent can a 'keyness-driven' framework be used to retroactively surface toxic content from Twitter?

II. What latent topics exist on Twitter which mirror the toxic language used by white supremacists? How have these changed over time?

METHODOLOGICAL CHAPTER

The fact that language is a rich data-source underpins the suite of methods which systematically research big data in textual format, as well as the well-established methods of content analysis (Krippendorff, 1980) and discourse analysis (Fairclough, 1989; Van Dijk, 1993). What separates 'text-as-data' methods from the, aforementioned, traditional modes of researching language is how text is used. Traditional methods use text for their meaning, thereby understanding texts that we read, by contrast text-as-data methods which mine texts and abstract information from them (Benoit, 2020). These automated methods thereby benefit the present research which aims to systematically analyse informational patterns across large volumes of text, rather than approaching the documents one-by-one.

The methodological chapter is divided into four sections: first addressing the methodological rationale, next the data collection, the preprocessing steps undertaken and finally model application.

Methodological Rationale

The following sections elaborate on the the methadological rationale underpinning the present research by first outlining known challenges in researching harmful language using quantitative methods. Next, the basics of the proposed 'Keyness-Driven' framework are outlined, after which the two methods the research uses, keyness analysis and structural topic models, are explained.

Challenges in Researching Harmful Language using Quantitative Methods

Research to date on both hate speech and toxic speech tends to use machine classification, a method which requires researchers to identify texts that serve as examples of harmful language, which in turn are used to train a model that aims to predict whether new, unseen, text can be classified as such.

Price *et al*. (2020) outline three reasons which make toxic speech particularly challenging to accurately classify, even more so than hate-speech. First, toxic messages are more muted and less likely to use explicit or inflammatory language (ibid), whilst a number of studies show that the main attributes that drive successful classifiers are in fact lists of hateful words (Noever, 2018) or sequences of characters (ngrams) (Waseem & Hovy, 2016, Nobata *et al.*, 2016). Second, toxic language is more context dependent than explicit hate speech, captured by scholars describing it as often being "veiled" (Jurgens *et al.* 2019; Han & Tsvetkov, 2020; Nario *et al.* 2021). This is troublesome for machine classification as it relies on manual annotation, known to systematically miss veiled language, stereotypes and identity attacks. For instance, a test conducted with a simple statement of explicit toxicity, "Peter is an idiot" had a miss rate of 43% by annotators (Nario *et al.* 2021: 15). Finally, as classifiers work with binary classes (i.e. either belonging to a category or not), the ambiguity related to toxic speech means that models run an even greater risk of classification error (identifying 'false positives' and 'false negatives') than hate-speech models.

In attempts to account for this uncertainty, the few scholars who have successfully classified toxic language have used bidirectional encoder representation from transformers (BERT) word embeddings (Price *et al.* 2020; Nario *et al.*, 2021). By using it's 110-340 million parameters, BERT embeddings can somewhat accurately account for the context in which words are used, and therefore improve classification significantly (Milutinović & Kotlar, 2021). Despite their promise of being able to provide models at 70-80% performance for researching toxicity, BERT embeddings require computing capacity far beyond that at hand for this research.

A 'Keyness-driven' Framework

Instead of approaching the study of toxic speech as a classification task, this paper proposes a new method, understood as a procedural framework containing a number of steps, seen to have the added benefits of 1) systematically accounting for changes in language over time 2) controlling for context by narrowing the scope of the texts and 3) being semi-automated and therefore ridding the research of subjective human annotation.

The statistical method 'keyness analysis' underpins the proposed framework. Given a corpus known to contain toxic language, as well as an equivalent corpus on the same topic from a mainstream domain, a keyness analysis will present a list of statistically significant terms which can be used to subset similar texts from the mainstream domain. Hence, this approach inductively creates a custom filter of features that have relatively high prevalence in a toxic text, compared to a benchmark of mainstream public discourse.

From the onset, the framework needs the texts to meet three basic criteria: 1) the posts must represent similar time-periods, 2) be of similar length and 3) relate to a specific topic or theme. The present research uses the time period 2011 to 2020, processes texts from Stormfront and Twitter to be of equivalent length, and all have the common theme of 'immigration'. This approach is rationalized by the body of research which indicates that lexical features (i.e. words) are strong indicators of harmful language (Nobata *et al*., 2016; Waseem & Hovy, 2016; Noever, 2018) and that quantitative language models produce best results when the scope of the texts is narrow (Benoit, 2020; Grimmer & Stewart, 2013).

The approach is referred to as an exploratory 'framework' due to the many steps (as will be outlined), qualitative inputs, and aforementioned criteria needed to surface toxic language in this way, by no means claiming to be a well-tested and fully automated algorithmic model. Once all steps are undertaken, the 'keyness-driven' framework is used to extract a subset of tweets, and by being an outcome of this process are said to have *high lexical resemblance* to toxic posts from Stormfront.

Primary Method: Keyness Analysis

Keyness analysis is a statistical measure which identifies significant differences in the use of features, most often terms, between two corpora (Gabrielatos, 2018). The idea of capturing keyness initiated with Scott (1997), who introduced the word 'key word' to describe "a word which occurs with unusual frequency in a given text [...] by comparison with a reference corpus of some kind" (ibid: 236). By creating lists of words that were key to a topic, Scott (1997: 233) established that words, when contextually grouped together in 'culturally significant ways' would 'provide a representation of socially important concepts'. To mitigate against the risk that terms have various meanings in various contexts, automated textual analysis benefits from studying texts with narrow themes (Benoit, 2020. Therefore, this research focuses specifically on toxic speech on the topic of immigration, in turn hoping to enhance the aforementioned representations of socially important concepts on this particular topic. Thus, what this research aims to capture is the cluster of terms[3] truly 'key' to the Stormfront posts on immigration ['target text'], compared to the same topic on Twitter ['reference text']. This type of frequency analysis can be computed either in a 'focused' or 'exploratory' manner (Gabrielatos, 2018), the prior used when researchers set out to research a specific set of predefined terms. This research is seen to use a hybrid of both, given that the terms all relate to immigration, but are not set a-priori - but are rather extracted inductively for exploratory analysis, described as a 'way into texts' or to generate terms to inspire further study (Gabrielatos, 2018: 227).

Secondary Method: Topic Modeling

To address the second research question this paper uses topic modeling, a set of unsupervised machine learning models which allow for latent topic discovery in unknown corpora (Blei *et al.*, 2003). It is an inductive[4] model which clusters collections of words into topics which the researcher then labels by hand. This approach is particularly advantageous when studying the subtle attributes of toxic language, as these unique models capture linguistic patterns that are cumulatively frequent and only observable across thousands or millions of words (Stubbs, 1994), making inferences as if "from a birds eye view" (Grimmer & Stewart, 2013). The statistical method underpinning the method which allows the model to capture latent linguistic patterns is based on identifying the connections between recurring co-occurrences of words that run through texts and across documents (Blei, 2012).

Stemming from the family of unsupervised machine learning, the models second advantage is that they allow researchers to approach new datasets without having in depth knowledge of a specific discourse (Mishra, 2017). This mitigates against researchers' subjectivity, biases

[3] Note that Wilson (2013) suggests 'key item' as a more inclusive term, useful for this analysis given that features are stemmed and include bi-grams, thus technically no longer being 'terms' in their true sense. For simplicity, the present research refers to 'items' as key terms.

[4] Referring to the inference of "universal statements" from "singular statements", such as hypotheses or theories (Popper, 1959: 426).

and other preconceptions, as has previously proven challenging for research on toxic text, contrasting deductive that has been used so-far to research toxic speech classification, relying on annotated training data with a set number of predefined labels and clear frameworks of what defines each category from the onset.

This research applies an STM, which builds off the most common 'latent dirichlet allocation' model (LDA) (Blei *et al.*, 2003), and extends three other models: the correlated topic model, the Dirichlet-Multinomial Regression topic model, and the Sparse Additive Generative topic model (Roberts *et al.*, 2013). Compared to LDA, STM's have the added advantage of being able to incorporate document-level metadata as model covariates (Roberts *et al.*, 2013). This combats the most restrictive of LDA assumptions -- that topics can only be modeled based on the document content -- thereby disregarding if they are from the same author, date, or obtained through a specific sampling strategy. STM's thus account for that word-use is dynamic and that different words can be described to discuss the same topic over time (Lebryk, 2018), as this research aims to explores topic evolution over time, the STM is therefore undoubtedly the best suited method.

Data Collection

This research uses data collected from two sources: Twitter and the white supremacist forum Stormfront. The following section outlines the rationale for selecting these two corpora, details how the data was gathered and ends with comments on the ethics of text-mining.

Twitter - a Mainstream, Highly Networked, Microblogging Platform

Twitter's popularity in research stems from it being a rich data source which affords the ability to explore high volumes textual data on an abundance of topics, in a readily accessible format thanks to their Application Programming Interface (API) (McCormick *et al.*, 2015). Social media data has the benefit of being unprompted by researchers and therefore represents utterances of language and broadcasts of opinions that are unfiltered yet public (ibid). The research also benefits from Twitter having a large user base and consistent activity during the period of interest (2011 - 2020). Finally, its greatest value for this research lies in it being the most established and mainstream microblogging platform, distinct in its users sharing political opinions, news, and, in contrast to most other social platforms, facilitating interactions with a network much larger than one's personal or immediate one. This research was granted permission to Twitters full historical archive, accessed through Twitter's API v2 endpoints[5] .

Stormfront - the World's Largest Online Forum for White Supremacists

Stormfront is a rich data source, providing ample data spanning many years -- since its creation in 1995 has attracted over 13 million posts and 300,000 members contributing across its message boards (Wong *et al*., 2015). Given that this research concerns white supremacy, Stormfront was a suitable platform, which has also been used for other studies to "track extremist attitudes about topics on race, immigration, and politics" (Dentice, 2019: 147). When studying harmful language, scholars have tended to deal with the imbalance in classes (harmful vs. non-harmful) by collecting data from pages that are expected to contain higher proportions of harmful language (Tulkens *et al.* 2016; Saleem *et al.*, 2017; Schmidt & Wiegand, 2017). Further studies have shown promise for accessing expression of 'ideologically sensitive' sentiments using forums as the data is 'naturally-occurring' (produced with few social constraints) and often public (Holtz *et al*., 2012, 2020). The data from Stormfront was provided by courtesy of far-right scholar Anton Törnberg[6], and consists of a full record of all posts published from the forums inception in 1995 to 2020[7].

Ethics & Reflexivity

This research makes use of text mining from both Twitter and Stormfront. The ethics of using this data was considered, as studies using such texts can 1) question the morality of cultural labor and 2) concern the researcher's transparency with the users whose texts are mined (Kennedy, 2016). The justification for using this data aligns with Holtz *et al*.'s (2012) reasoning

[5] https://developer.twitter.com/en/products/twitter-api/academic-research

[6] Gothenburg University - https://www.gu.se/en/about/find-staff/antontornberg

[7] For peer-reviewed work that has used this dataset, see Törnberg & Törnberg (2018) and Törnberg (in press).

for using forum data from political interests groups specifically: this communication can be justified as 'public behavior' because the platform is used to convey the groups agenda. Similarly, researchers generally regard Twitter as a public social media platform, thereby being acceptable to source data from. The pilot does not collect any identifiable names or pseudonyms.

Sampling

This section first outlines how a taxonomy capturing 'immigration' was created. Next, sampling from Twitter and Stormfront is explained.

Creating a 'Immigration' Taxonomy

To sample data from both platforms text that related to immigration, a taxonomy of keywords was created[8]. A comprehensive taxonomy must take many factors into account due to the idiosyncratic language use on social media (Tulkens *et al.*, 2016). Within this research immigration is broadly defined as the "process of coming to a country in order to live in it permanently" (Cambridge Dictionary, 2021), therefore seen as an umbrella term encompassing both migrants and refugees (Amnesty, 2021).

First, the taxonomy made use of so-called "wildcard" matching (Beniot, 2020), capturing any of the main keywords "refugee", "immigrant" or "migrant" by their stem[9], to account for various endings to the terms. The 19 most common misspellings of these three terms were also added. Finally, the taxonomy was further expanded by accessing the open data initiative Hatebase's API[10], containing a crowdsourced multilingual set of hate words. 18 out of their 1500 terms in English specifically related to immigration, for instance: "border hoppers",

[8] Note that the initial criteria applied to both datasets was 1) only keeping posts in English 2) limiting the sample to posts published 2011 - 2020.

[9] This is done to allow for the characters at the end of the term to vary, for instance "immigr*" will capture "immigrant", "immigrants", and "immigration". All matches were also checked manually. See Appendix D for a sample of the extracted terms using wild-card matching.

[10] https://hatebase.org/search_results

"reffo", "anchor babies" and "wetback". The final taxonomy consisting of 38 terms can be found in Appendix C.

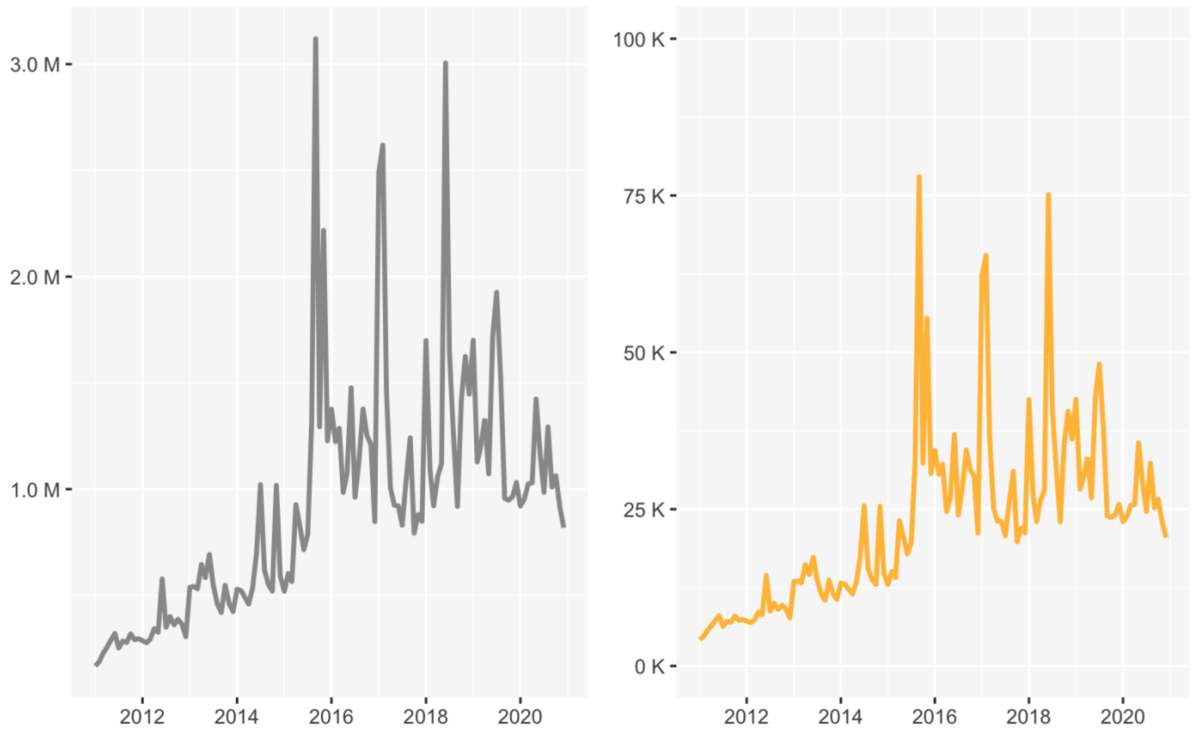
Twitter Sampling

Surfacing harmful language on social media is equatable with "looking for a needle in a haystack" (Marantz, 2020). Despite many studies using Tweets, there is surprisingly little consensus on how to appropriately draw samples from the platform (Lewis *et al.*, 2013). Generally, scholars use a stratified 'constructed week sampling' or variations of simple random sampling (Harlow & Johnson, 2011; Artwick, 2014, Vidgen & Yasseri; 2020), empirical tests comparing the two indicating that latter proves a more representative sample Kim *et al.* (2018). This study follows Takahashi *et al.*'s (2015) random sampling strategy for time-series data by collecting Tweets from a window of time -- set at random -- for each day for the full period of interest.

The size of the sample was constrained both by computing capacity, and Twitter's rate limit for researchers (10M Tweets/month). Sampling proceed by exhausting all Tweets[11] from a two-hour window each day set at random for the 10 year period. Each Tweet had to mention at least one of the aforementioned keywords on immigraton from the custom taxonomy and retweets were not collected, as the research aims to explore as many variations of language as possible. Next, these samples were aggregated by month, and from each of the 120 groups (12 months x 10 years) a random sample was drawn to balance the sample in a way that reflected a 2.5% proportion of the total volume of tweets posted in that month[12] (Plot 1) (Hino & Fahey, 2019) .

[11] To stay clear of hitting the rate-limit, a cap was set at 3000 tweets per two hour time slot, only hit on few occasions during data collection.

[12] The sample initially reflected 5% (N = 6,516,911) of Tweets but was reduced in size due to limits in computing capacity



Plot 1: Distribution of total Tweets (left) compared to the collected sample (right)

Preprocessing

The proceeding section outlines the two preprocessing steps undertaken: toxicity scoring and text-cleaning.

Automated Toxicity Scoring for Stormfront posts

When studying keyness, the 'target' corpus represents the text one wishes to compare to use as a baseline, or 'reference', corpus. The terms this research is concerned with are 'toxic' terms, therefore the target corpus needs to largely represent toxic language. To identify toxic posts from Stormfront to use as the 'target' text, an model developed by Google was used (as seen in Nario *et al.*,2020). The model is accessible through the Perspective API[13] and was created by Jigsaw and the Google Counter-Abuse Technology Team to help improve online conversations (Jigsaw, 2021). The model uses advanced machine learning to assign individual comments a

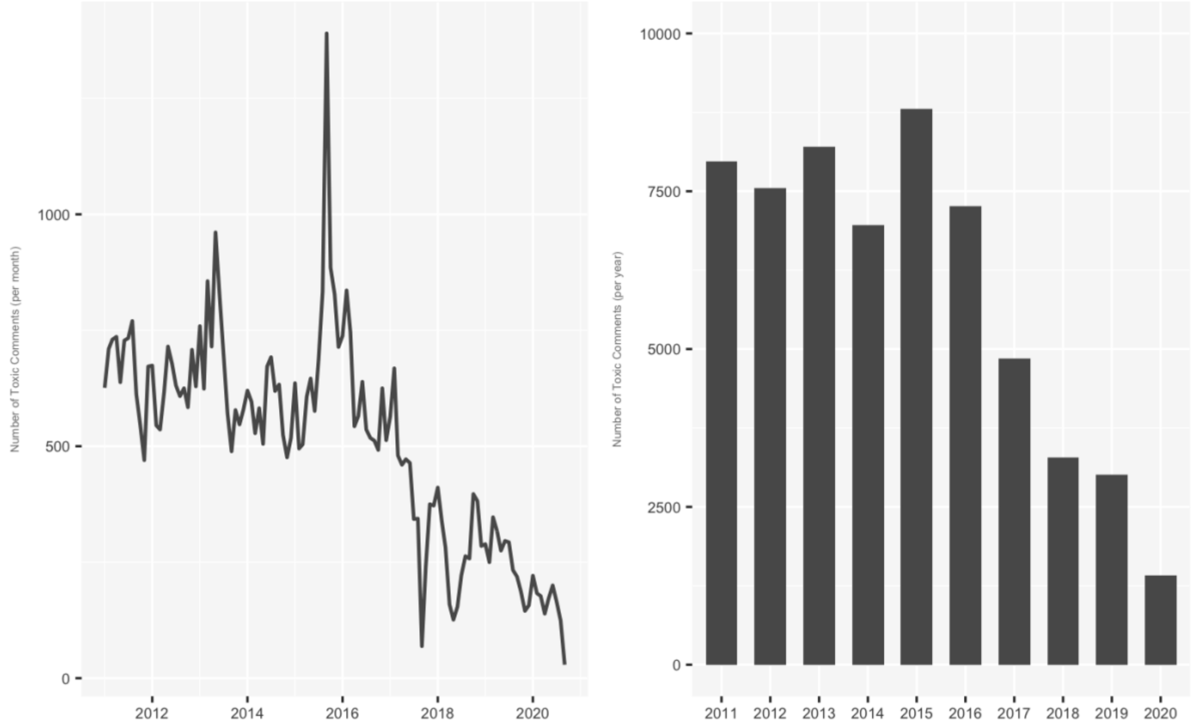
[13] https://www.perspectiveapi.com/

score between 0 to 1, 1 being 'most likely' to be toxic[14]. Generally, 'off-the-shelf' classification tools are ill-advised (see Vidgen *et al*. (2020) relevant for criticism), but for toxicity scoring specifically, the model was deemed satisfactory.

Before passing the posts to Perspective, each comment was divided into its sentences as a noise-reduction measure, to hone in soley on the part of the comment that had a match with the 'immigration' taxonomy. Sentence level analysis also made the posts on Stormfront more suitable for comparison to Tweets in terms of length. A total of 177,000 sentences (hereinafter 'posts') were scored by Perspective[15], subsequently filtered by a 'toxicity score' of 0.4 ≥, as per Nario *et al*.'s (2020) work on toxic speech. After this step, the dataset contained 59,000 sentences, distributed per year as follows (Plot 2). Examples of highest and lowest scoring texts examined before proceeding with the data in Appendix E.

[14] Perspective also includes other models, such as 'severe toxicity', 'insult', 'sexually explicit', 'profanity', 'treat' and 'identity attack' (Jigsaw, 2021), though for the purpose for this research only 'toxicity' is measured.

[15] This was a rather unconventional approach as the tool is primarily used for real-time content moderation and therefore usually does not process large quantities of comments at once. This project was granted special access to an increased query per second limit, making it possible to score such a high number of posts in a relatively short amount of time (7 days).



Plot 2: Distribution of Stormfront corpus, per month (left), per year (right), N = 59301

Text-Cleaning

The following preprocessing steps were undertaken to model the data and make it machine readable prior to modeling (Benoit, 2020):

1. Texts were removed that were shorter than 7 characters in length (these included single words only - mainly variations of "immigration" without further context).
2. Duplicate texts were removed (likely spam), decreasing the Twitter and Stormfront corpora each by 2-3%.
3. Corpora were cleaned from numbers, punctuation and other symbols.
4. Terms were stemmed (reducing words to their canonical form).
5. Terms were lowercased (equating 'USA' and 'usa').
6. Bi-gram were constructed (taking into account word-pairs or multiword expressions, e.g. 'New York').
7. Stopwords were removed based on a pre-defined list (such as "the", "and", "of"), as well an extended list of 412 non-discriminative stopwords including prepositions, conjunctions, definite articles, and pronouns, each extracted from the TidyText package (Appendix F) (Silge & Robinson, 2016).

Next, the corpus was converted into a 'document feature matrix', a statistical representation of the corpus sufficient to infer substantive properties of text (Hopkins & King, 2010). Table 2 presents summary statistics for the 20 corpora.

Corpus	# of documents	# of features	# types	# tokens	Avg. tokens per document
Twitter_2011	75,035	494,526	1,251,368	1,279,218	17
Stormfront_2011	15,140	112,035	278,419	285,268	19
Twitter_2012	104,711	646,624	1,766,913	1,808,890	17
Stormfront_2012	15,140	112,035	278,419	285,268	19
Twitter_2013	155,917	923,738	2,585,097	2,629,397	17
Stormfront_2013	15,300	113,943	282,101	289,247	19
Twitter_2014	181,799	1,066,265	3,071,446	3,126,580	17
Stormfront_2014	14,820	111,302	272,787	279,824	19
Twitter_2015	341,785	1,783,323	5,684,050	5,765,305	17
Stormfront_2015	15,357	116,240	284,947	292,162	17
Twitter_2016	347,741	1,895,599	5,906,910	5,984,135	17
Stormfront_2016	15,497	118,891	290,307	297,525	19
Twitter_2017	322,110	2,031,319	5,963,771	6,038,444	19
Stormfront 2017	11,708	94,159	218,758	224,218	19
Twitter_2018	414,391	3,318,382	11,097,977	11,373,828	27

Stormfront_2018	8,129	67,416	148,550	152,518	19
Twitter_2019	380,834	3,287,803	10,989,315	11,295,390	30
Stormfront_2020	6,172	54,479	116,562	119,678	19
Twitter_2020	271,474	2,677,257	8,133,010	8,367,557	31
Stormfront_2020	4,349	39,993	8,1420	83,398	19

Table 2: Summary statistics for each corpus

Model application

Next it is explained how 'key' terms are extracted, and how these are used to subset tweets. After these two steps the structural topic model is applied.

Extracting 'key' terms

This research uses Pearson's chi-squared test statistic to compute keyness, as the metric has shown promise to extract features to improve machine classification (Bahassine *et al*. 2020). The chi-squared test statistic measures the difference between observed (O) versus an expected frequency (E) if the independent variable (origin of corpus) had no effect on the distribution of the term, thereby providing an "indicator of a keyword's importance as a content descriptor" for the target text (i.e. Stormfront) (Biber *et al*., 2007: 138). The difference's significance is indicated by the associated p-value. Pearson's chi-squared is annotated as follows:

$$\sum_{i=1}^{n}(O_i - E_i)^2/E_i = X^2$$

The objective of this analysis is "not to maximise, or minimize, the number of key terms, but to derive as true a picture as possible of the differences [...] of item frequencies between two corpora" (Gabrielatos, 2018: 233). As a means to this end, thoughtful choices were made regarding how to best analyse the keyness of a corpus spanning 10 years, deciding to perform one analysis per year. This was motivated by, first, that the analysis deals with very large samples. P-values are known to be sensitive to sample sizes, with larger samples having an outsized impact on effect-size (despite how small they are) (Gabrielatos, 2018). Subsetting the sample by year reduces this problem. Secondly, it was decided to compute keyness analysis by year on Twitter (e.g. 2020), whilst data from Stormfront was also paired with data the prior year (e.g. 2019 + 2020). Studies have exemplified how language from far-right websites infiltrated the mainstream media ecosystem over-time (Zannettou *et al.*, 2018) . Therefore, language prevalent on Stormfront is not expected to arise on Twitter in perfect unison, but rather as a reverberation. Furthermore, initial keyness plots indicated that across the 10 years, each year only had 5 out of the top-50 terms in common ("white", "jew", "non-whit", "negro", "anti-whit"). Meanwhile, each consecutive year of Stormfront data had 38% - 76% similarity in top-terms (Appendix G).

Subsetting Tweets which Language Mirrors Toxic Posts on Stormfront

Prior to modeling the data, addressing the second research question, the subset of Tweets with highest lexical resemblance to the toxic stormfront posts were extracted. First, key terms underwent a filtering process, and next a weighting procedure was applied.

The keyness analysis provided 8,000 - 15,000 key terms from each year ($p \leq 0.01$), in turn reduced to act as a filter to extract Tweets expected to contain similar language to the toxic posts on Stormfront. Narrowing the filter was done by 1) only considering terms with positive chi-squared values (as negative values indicate an absence from the target text - being 'key' for Twitter, not Stormfront) and 2) each term having 2-5 mentions in the targeted text and 1-3 mentions in the reference text (dependent on corpus size for that year, larger corpora having a larger threshold). This method allowed for each final 'key term filter' to be reduced to 2000 3000 terms after which small numbers of filler terms (e.g. "etc") were removed as it is encouraged to first use threshold based extraction and then edited using human judgement (King *et al.*, 2010). Other approaches considered for term reduction was 'part of speech tagging' (Wong *et al.*, 2015; Saleem *et al.*, 2017) and, rather simplistically, only selecting the top 100 terms (Gabrielatos, 2007).

Next, the terms were weighted by a fraction of their respective chi-square, given the chi-squared statistic provides a metric of how discriminating a term is for the target corpus. This was done as a list of terms in themselves are not considered a robust way to accuratly subset corpora. Thus, term-weights are applied, as inspired by 'weighted dictionaries'. Following Jegadeesh and Wu (2011), this research applies feature weighting based on the rationale that some 'key features' are more distinctive than others, meaning that this approach can assign an overall 'document score' by taking the relative importance of each word into account. By way of example, sentiment analysis uses weighting to distinguish between terms that are more positive than others, for example to indicate that 'excellent' is more positive than 'great' (Taboada *et al*., 2011; Liao *et al*., 2014). In their work Jegadeesh & Wu (2011) even test their weighting scheme and find that the term weighting is *"as least as important, and perhaps more important than, a complete and accurate complication of the word list"* (ibid: 726). As the research sets out to explore a discourse about which there is little prior knowledge of, the filter consists of a less fine-grained set of features, hence extra importance is placed on having a weighting scheme in place.

Based on this weighting scheme, each Tweet received a score reflecting to what degree it contained the 'key' terms identified on Stormfront (see Figure 2 for example). Finally, the top 1% of Tweets with the highest score each year were extracted for topic modelling (N = 25,961 - summary statistics in Table 3).

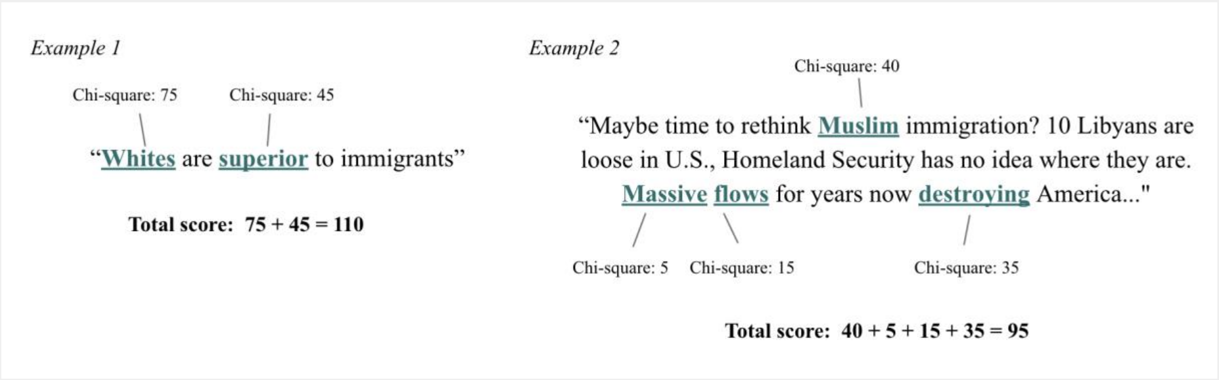

Example 1
Chi-square: 75
Chi-square: 45
"Whites are superior to immigrants"
Total score: 75 + 45 = 110
Example 2
Chi-square: 40
"Maybe time to rethink Muslim immigration? 10 Libyans are loose in U.S., Homeland Security has no idea where they are. Massive flows for years now destroying America..."
Chi-square: 5
Chi-square: 15
Chi-square: 35
Total score: 40 + 5 + 15 + 35 = 95

Figure 2: Hypothetical exemplification chi-square based document scoring

To gauge that this process in fact captured Tweets using language similar to the toxic Stormfront posts, a sample of the tweets were reviewed.

Corpus	# of documents	# of features	# types	# tokens	Avg. tokens per document
Filtered_Twitter	25,961	229,165	643999	667525	26

Table 3: Summary statistics for Twitter subset

Structural Topic Model

The STM's two outputs are posterior probabilities of 1) topic prevalence over each document p(z|d) , and word distribution over each topic p(w|z). The generative process for estimating topical prevalence p(w|z) and topical content are a complex function of document metadata (Roberts *et al.,* 2013) (see Appendix H for details including STM plate notation and posterior distribution).

The goal when topic modeling is to strike a balance between two contradictory goals to obtain distinctive topics: that words from each document should occur in as few topics as possible, and that each topic must contain various words with different probabilities to form a coherent topic.

When applying a STM the only parameter provided by the researcher is the number of topics (K), set *a priori* (ibid). The number of topics is evaluated based on qualitative and quantitative insights. Model diagnostics were run on 40 iterations of the model to systematically assess a sensible number of topics for the research objective. First models were run to compare the held-out likelihood, residuals, semantic coherence and lower bound of models with 10, 20, 30, 40, 50 and 60 topics, next 40 models with k= 2:40 were tested. Upon reviewing these diagnostics (Appendix I), it seems that the residuals quickly declined between 10 - 20 topics, and marginal increases in held-out likelihood after 20 topics. Therefore, K in range 10-20 topics were explored qualitatively by viewing the documents most representative for each topic, assessing the most prevalent words in each topic, and making an overall judgement about topic coherence and and term exclusivity (these models can be found in Appendix J. This is a critical step when topic-modeling, as the diagnostic values have been shown to be unrelated or even negatively correlated with topics' semantic coherence (Chang *et al.*, 2009).

K=15 was settled on as it modeled qualitatively interpretable $p(z|d)$ and $p(w|z)$, but also based on the diagnostics having equally high semantic coherence (≈ 175) as a model with 19 topics but with lower held-out likelihood.

As a final step the topics were labeled. Despite being computed quantitatively, the analysis and findings of a topic model are highly qualitative and ultimately a subjective task (Benoit, 2019). Labels were assigned based on an interpretation of the 1) frequently used terms 2) the terms that are most exclusive to the topic and 3) viewing the documents that were seen to have the highest topic prevalence (Roberts *et al.*, 2013). Alternatively, topics are labeled rather reductively, simplifying them to their most frequent terms. This research opts for the more qualitative approach thus interpreting topics based on the aforementioned steps, although being wary of that relies more on the researcher's understanding of the discourse.

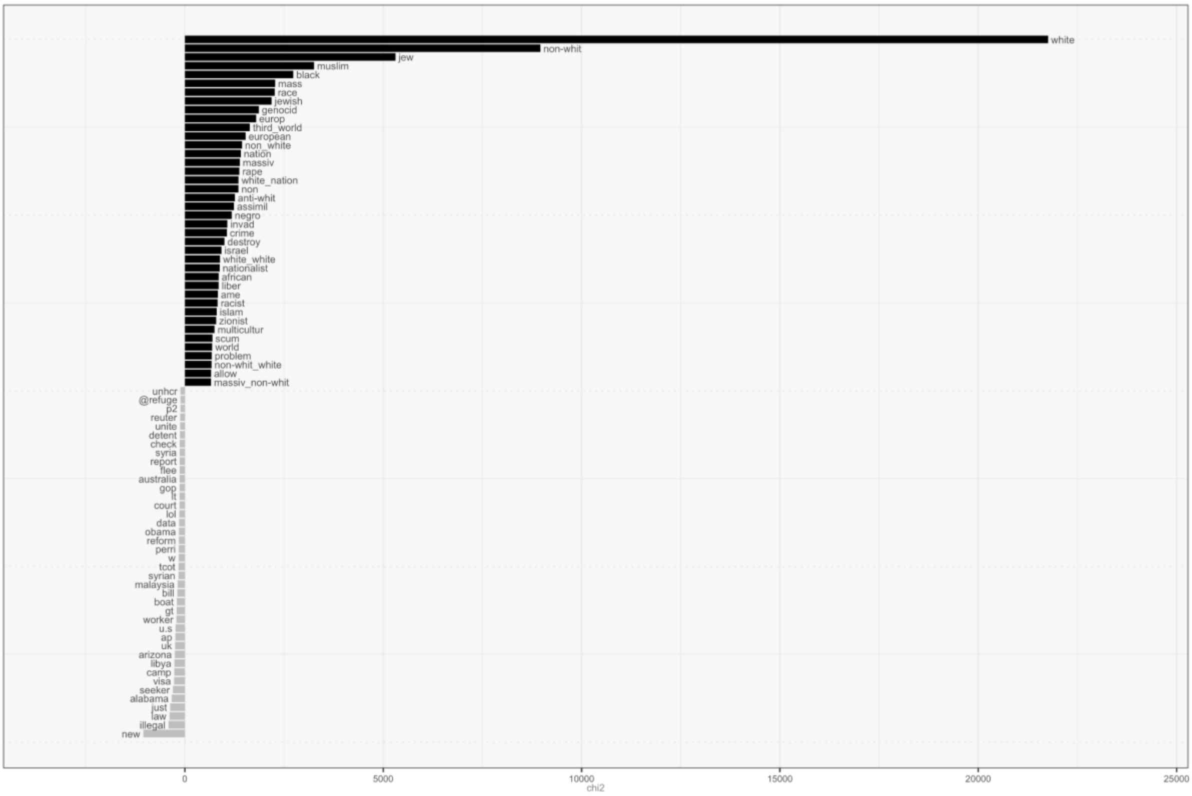
RESULTS

The two sections outlining results are each related to the posed research questions. The first section presents findings specifically pertaining to the proposed framework for surfacing toxic content using keyness. The latter presents the identified latent topics within toxic Twitter discourse, and addresses how these have changed in the past decade.

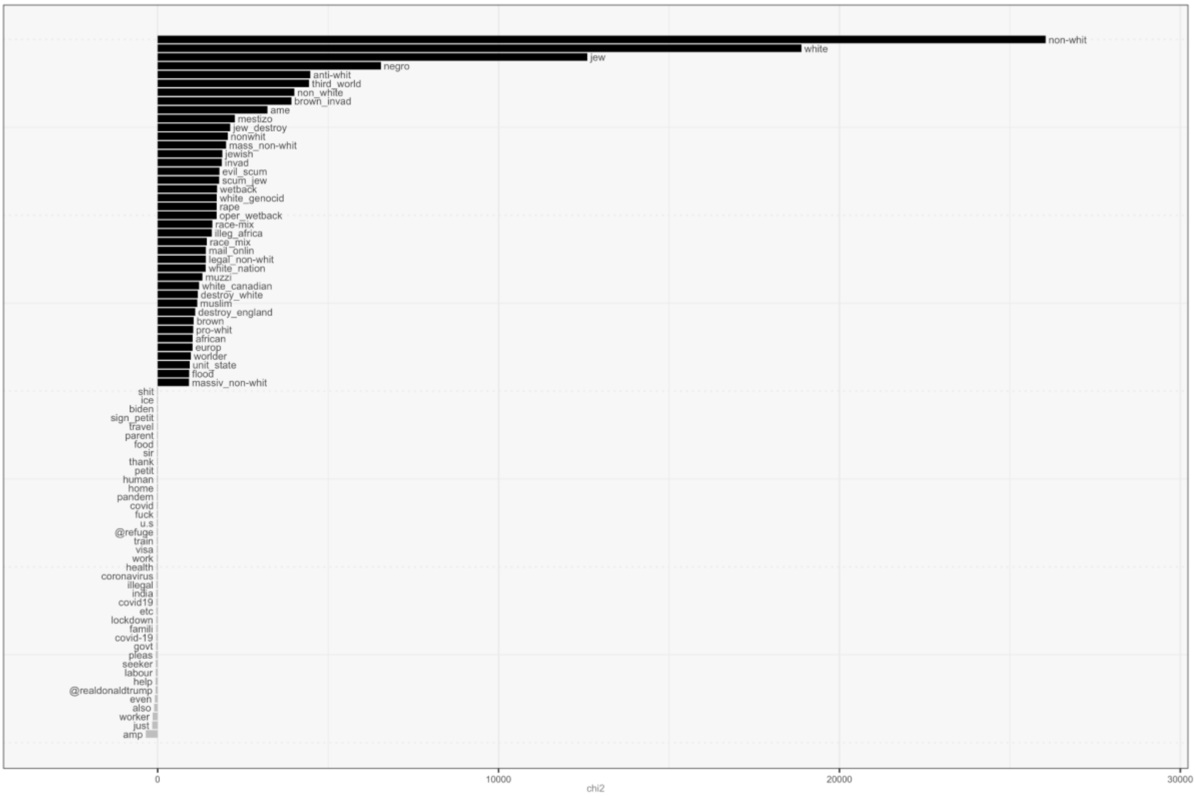
Empirical Evidence of Utilizing a 'Keyness-driven' Framework to Surface Toxic Content

The initial stage to surface toxic speech similar to that of white supremisists using the 'keyness-driven' framework was to conduct one keyness analysis per year, representing posts on Twitter and Stormfront from 2011 to 2020. This step provided an analysis of 'key terms' which had significant differences in observed vs. expected frequency on Stormfront, compared to

Twitter each year. The Top-40 'key terms' for 2011 (Plot 3) and 2020 (Plot 4) can be seen below (2012-2019 keyness plots found in see Appendix K).



Plot 3: 40 terms most and least distinctive to Stormfront corpus compared to Twitter corpus in 2011. The top section (black) represents the terms used in toxic Stormfront posts that are most divergent in their frequencies compared to those used on Twitter. The second set of terms (gray) are the most divergent in terms of not appearing in the Stormfront corpus at the frequency that would be expected based on the Twitter corpus.



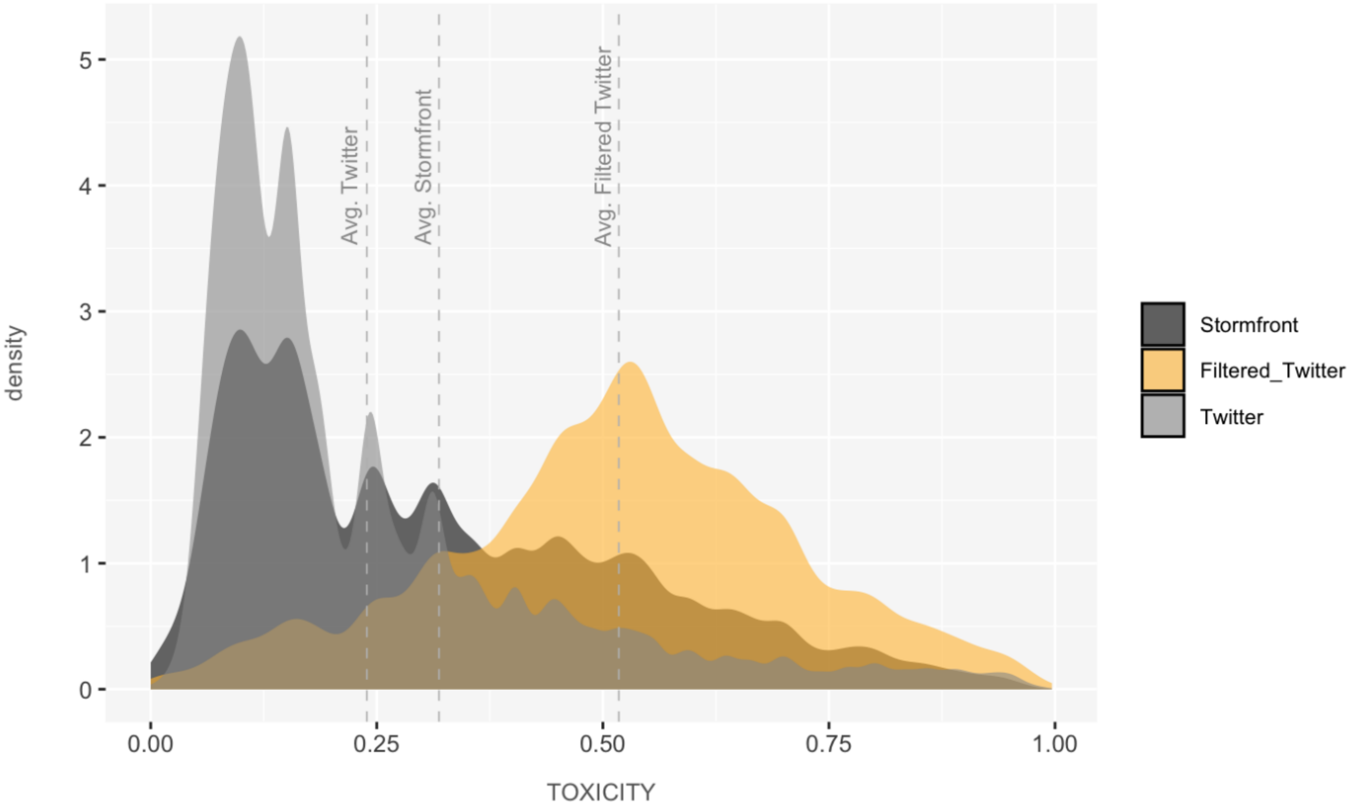
Plot 4: 40 terms most and least distinctive to Stormfront corpus compared to Twitter corpus in 2020. See Plot 3 for interpretation.

The use-value of keyness analysis was demonstrated by the fact that the majority of terms which occurred consistently across years were seen to be central to the white supremacist ideology, exemplified by terms often used to attribute labels to specific populations, with those (e.g. "white", "jew", "black", "anti-white", "african", "nationalist", "liberal", "wetback"). Overall, 'key terms' showed remarkable consistency across years (Table 4).

Terms present across the full sample (9 or 10 years)	**Terms present in many years (8 or 7 years)**	**Terms present in majority of years (5 or 6 years)**
"white", "jew", "muslim", "jewish", "rape", "invad", "non-whit", "black", "non_white", "white_nation", "anti-whit",	"mass", "genocid", "third world", "african", "ame", "race", "non", "nationalist", "liber", "scum", "wetback"	"europ", "destroy", "israel", "zionist", "white_race", "nonwhit", "european", "massiv", "white_white", "world", "illeg", "mestizo", "mass_non_white"

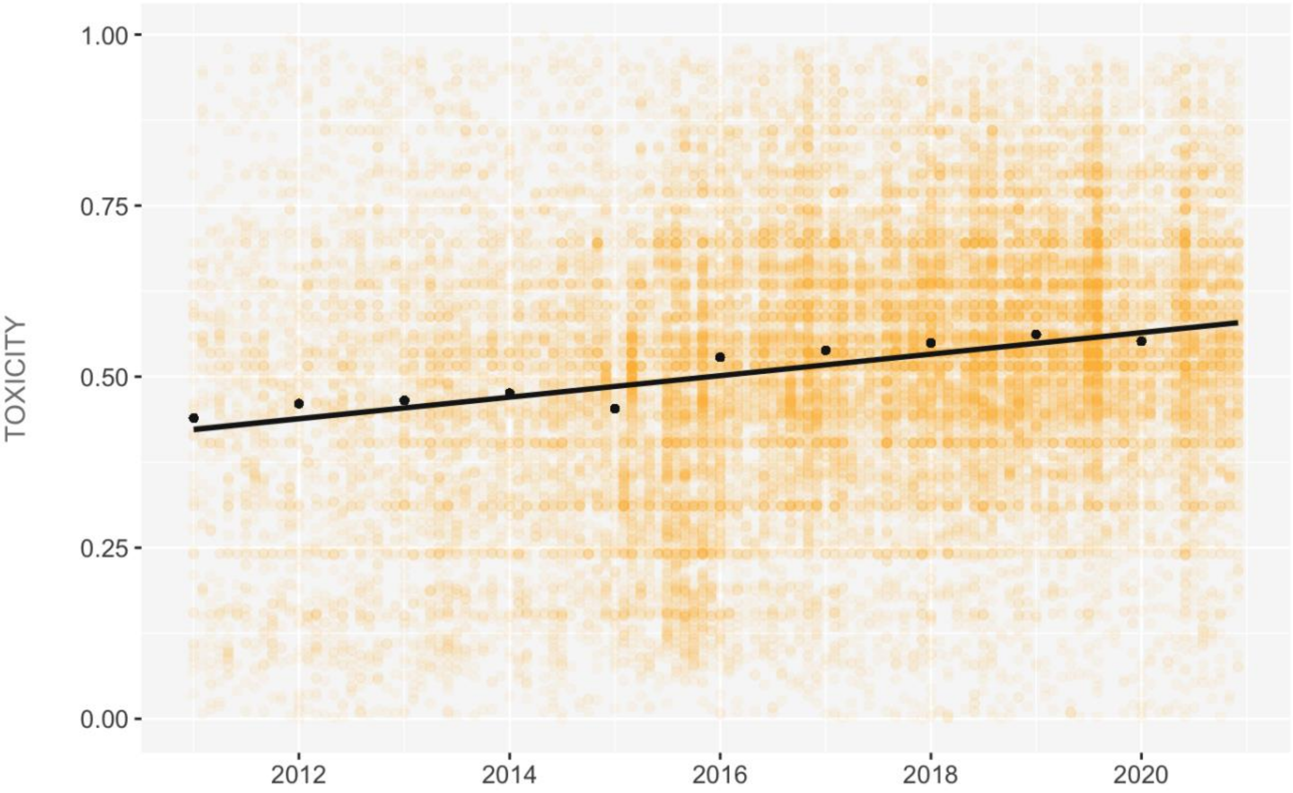
Table 4: Top 'key terms' (by chi-squared), organized by number of years each term was present in

As seen in Plot 5, the Tweets surfaced using this 'keyness-driven framework' had an average 'toxicity' score +117% higher than the Twitter benchmark for any term matching the immigration taxonomy, as well as showed +62% higher toxicity than the posts matching the same taxonomy on Stormfront - a forum known to be contain particularly toxic language.



Plot 5: Density plot showing the average Toxicity Scores for Twitter, Stormfront and the Tweets surfaced using the 'keyness driven' framework

Furthermore, plotting the toxicity scores for the 26,000 surfaced tweets over time indicates that toxicity has increased considerably in the past decade - observing a 28% increase in average toxicity since 2011 (Plot 6). Results of each individual year's average indicate a steady increase, and that no individual year is having an outsized impact on the observed increase.



Plot 6: Average Toxicity Scores per year of tweets surfaced using the 'keyness driven' framework. Each yellow dot represents a tweet and its toxicity score. Black dots indicate annual toxicity averages. Trendline plotted for clarity. (N= 25,961)

The Tweets surfaced using the 'keyness-driven' framework were also qualitatively inspected, showing examples of how toxic language which resembles that of white supremacist manifest on Twitter (Table 5). As should be expected when working with natural language modeling, sampling also showed a number of 'false-positives' (Table 6) as well as instances of counterspeech (Table 7). A supplementary qualitative analysis comparing the use of the key terms "illegal", "breed", "blacks", "black", "genocide", "destroy" and "flood" and their contextual use across Stormfront and Twitter, conducted to gain richer insight into the corpus, can be found in Appendix L.

Examples of toxic tweets resembling a use of language seen on Stormfront
"This is why the small Country of **Sweden**, is now the **rape** capital of **Europe**. **Muslim** migrants have managed to subjugate that Nation.."
"I guess that means I've got **white** privilege. I'm not ashamed of it. Btw - Not all cultures are equal, Some cultures are better than others - How else do you explain the immigrant **flows** TO Western countries and not the other way around?"
"This **non-White immigration**, combined with **forced** "**diversity**" of **White** areas intended to turn **White** people into a **minority**."
"If a **white** nationalist would've **killed** an **illegal immigrant** all of the media would've have gone crazy. Now that an **illegal immigrant killed** an innocent girl, they don't cover it too much. They got to keep the narrative that **illegals** are "great" people. Total bs"
"Both CIS and FAIR believe that certain immigrant groups are engaged in competitive **breeding** to **diminish** American **White Majority**."
"Based on this **racial** consciousness, **whites** must counter the demographic **threat** they face from immigration and **non-white fertility** and **whites** own infertility. This means (a) an absolute halt to all future **legal** immigration into the United States, deployment of the armed forces."
"And the **white race** is an **endangered** rare **species** of great ape. It can be argued that **whites** are being extra **tolerant** at this point in time and the posterity of the **white race** depends on not overflowing immigration with migrants from **non-white** countries."
"The questions we have to asked our elites, why they would so like refugees, they like them culture, them attitude, they believes, they religion despite they are so antagonist against **Europeans**. They even encourage **white womens** to **sex** with them."
"Zero **whites** get political asylum in USA, zero **white anchor babies**. Means only one **white** for every 20 **non-whites** gets to immigrate to America."

Table 5: Sample of Tweets surfaced using the 'keyness driven' framework, terms bolded which are seen to be statistically 'key' for toxic white supremisict speech.

Examples of counterspeech on Twitter
"If the "White man" has the right to kick a non-white immigrant, who has lived here for 30 yrs, out of this country, just because the White man has been here for 400 yrs, then shouldn't the Native Americans be able to kick the "White Man" out, since they've been here 15,000 years?"
"@realDonaldTrump Really? Crime has increased because of immigration of young Muslim men? Let's see...Vegas mass shooting, White US Male, Parkland mass shooting, White US male, Sante Fe,Texas mass shooting, White US male, Marshall high school shooting, White US male, Sutherland Springs, TX…"
"You know the reason @USERNAMEREMOVED! The refugees aren't white and our white government wants to keep Whites in the majority and in power. The last Census told them White numbers are falling behind by comparison to non whites! So much for racial tolerance!!!"
"@realDonaldTrump It has nothing to do with illegal immigration. A white citizen could've done the same thing. This was a question concerning premeditated murder beyond a reasonable doubt. Was that proven? No. Our justice system works properly. This is not a political issue."
"Pls stop saying white people. I am an American. We come in many colors and ethnic backgrounds. There are many white people from other nations that can face immigration policies as well."
"My grandma has been an illegal immigrant for over 50 years but no on gives a shit because she's white. Don't tell me it's not a race thing.

Table 6: Sample of counterspeech surfaced using the 'keyness driven' framework

Examples of 'false-positives' on Twitter
"EU to speed up deportations to tackle migrant crisis - EUROPE http://t.co/DnAEXUc6Tl via @HDNER" "Illegal immigration is not a new problem. Native Americans used to call it 'White People'" "Israel Eyes European Jewish Immigration After Denmark Attack http://t.co/5ZgnJbn05t"

Table 7: Sample of 'false-positives' surfaced using the 'keyness-driven' framework

Latent Topics on Twitter Using Language which Mirrors that of White Supremacists

The following results were obtained from the topic model which modeled the 26,000 tweets containing language mirroring that used by white supremasicts, therby addressing the second research question.

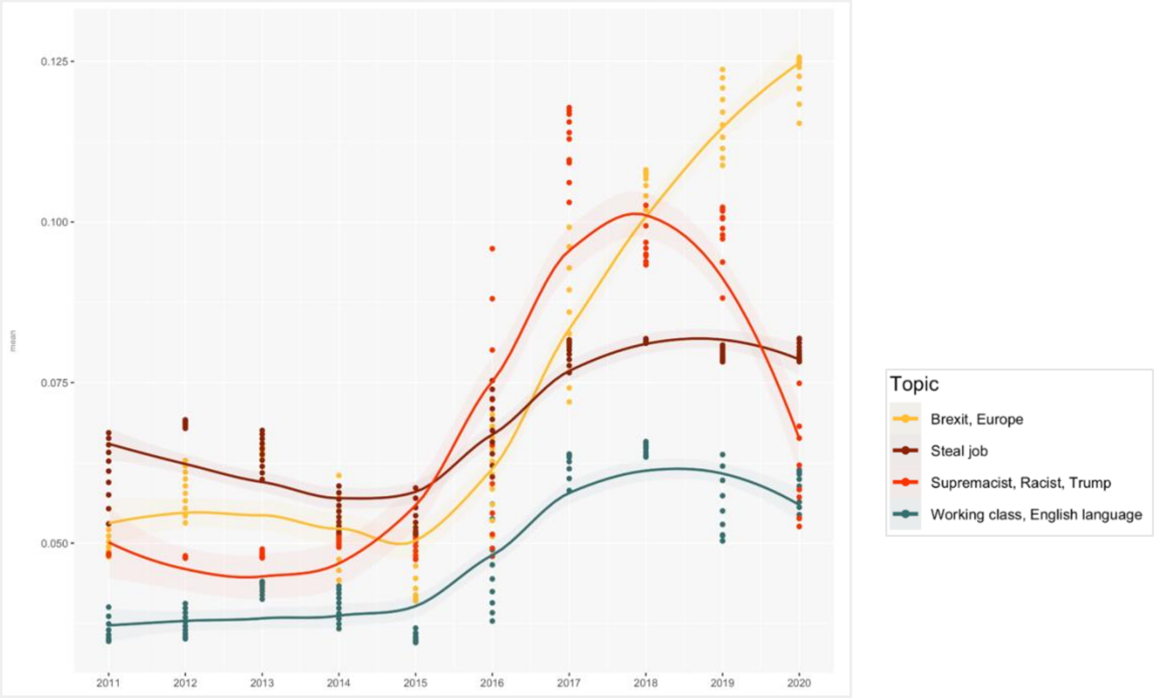
Using the previously outlined interpretative approach to labeling topics, eleven out of fifteen topics were labeled. Four topics were too ambiguous for labeling -- as should be expected when topic modeling -- and were therefore discarded. Each of the labeled topics are presented in Table 8 alongside terms which describe them, as well a short description of the topic.

Topic Label	**Description**	**Topic Terms**
English language, Middle class	Discussions on whether or not immigrants speak english to the working middle class.	"work", "man", "class", "speak", "work_class", "white_work", "middl_class", "hard_work", "speak_english", "american_heritage", "american_english"
Europe, Brexit	Immigration to Europe, free movement in the the EU, Brexit vote	"european", "vote", "issu", "white_eu", "commonwealth", "leav_eu", "eu", "white_european", "free_movement"
Jewish, Israel	Jewish immigration to other countries, discussions on Israel and Palestine ´.	"jew", "israel", "nazi", "palestinian", "jew_christian", "jew_murder", "state_israel", "creat_israel"

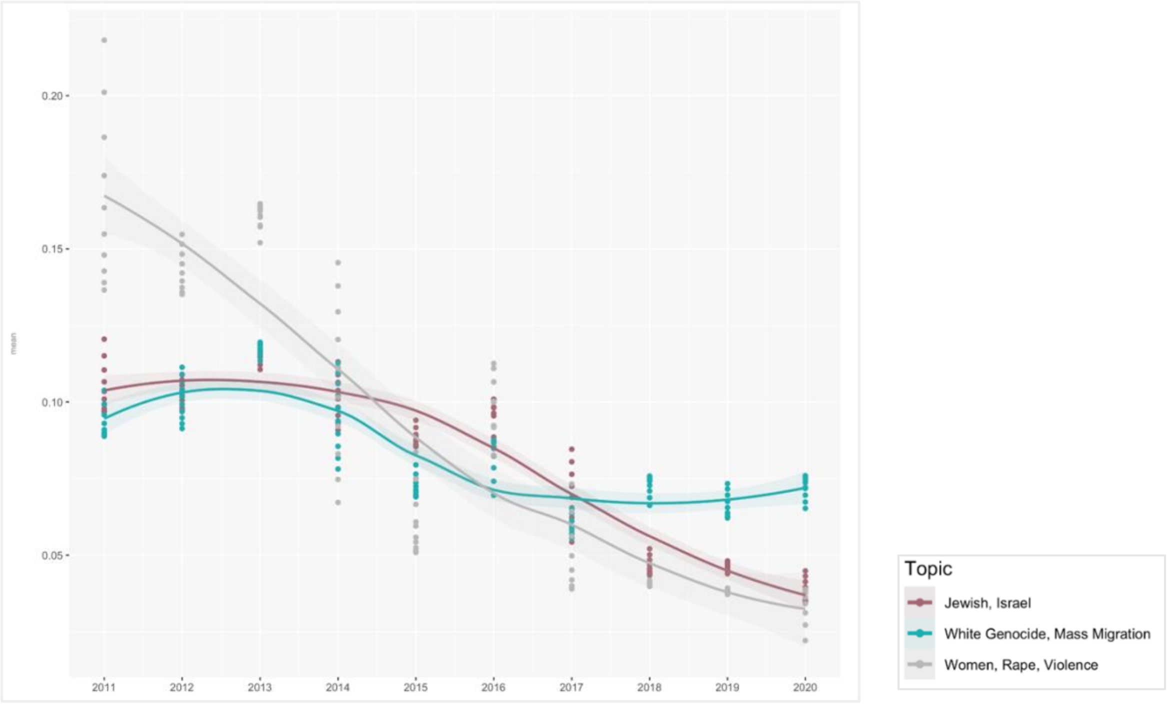
South Africa	Race and integration in South Africa	"south_african", "expat", "farmer", "accept_sa", "white_south", "white_expat", "africa", "folk"
Americans, Natives, Ancestors	Discussions on who are "true" Americans, settlers in America	"nativ", "go", "first", "invade", "nativ_american", "land", "home", "ancestor_white", "came", "back"
White genocide, Mass migration	The 'risk' of white genocide, immigration as a force to destroy the white race	"white", "mass", "migr", "cultur", "genocide", "white_countri", "assimil", "white_genocide", "destroy", "white_flight", "destroy_white",
Supremacist, Rasist, Trump	Discussions on racism, white supremacists, gun control and Donald Trump	"white", "supremacist", "trump", "male", "racist", "whilte_male", "shoot", "american_still", "shoot_white",
Women, Rape, Violence	Islam and violence against women, statements on a correlation between crime and immigration	"girl", "women", "muslim", "gang_rape", "muslim_invad", "white_girl", "sweden", "asylum", "women_muslim"
Skin color	A multitude of references to skin color - slightly ambiguous	"america", "skin", "black_brown", "white_america", "brown", "white_brown", "chain", "skin_color", "brown_white", "brown_skin", "straight_white", "color" , "white_famili"
Steal jobs	Immigrants stealing jobs, immigration and the economy	"white", "illeg", "take", "job", "illega_take", "hate_white", "take_job", "job_white", "steal_job", "illeg_come", "border jumper", "problem, "talk", "border_go"

Table 8: Overview of the ten labeled topics

In response to the latter half of the second research question, pertaining to how topics changed - the topics were explored in two ways. First, the topic prevalence across tweets were visualized to observe how the topics evolved over time. Four topics, *'Europe, 'Brexit'*, *'Steal jobs'*, *'Supremacist, Racist, Trump'* and *'Working class, English language'*, were seen to increase in topic proportion over time (Plot 7). The topics seen to have average topic proportions decreasing over time were *'Jewish, Israel'*, *'White genocide, Mass migration'* and *'Women, rape,* *violence'* (Plot 8). The two remaining topics, *'Skin color'* and *'South Africa'* did not indicate any change over time.

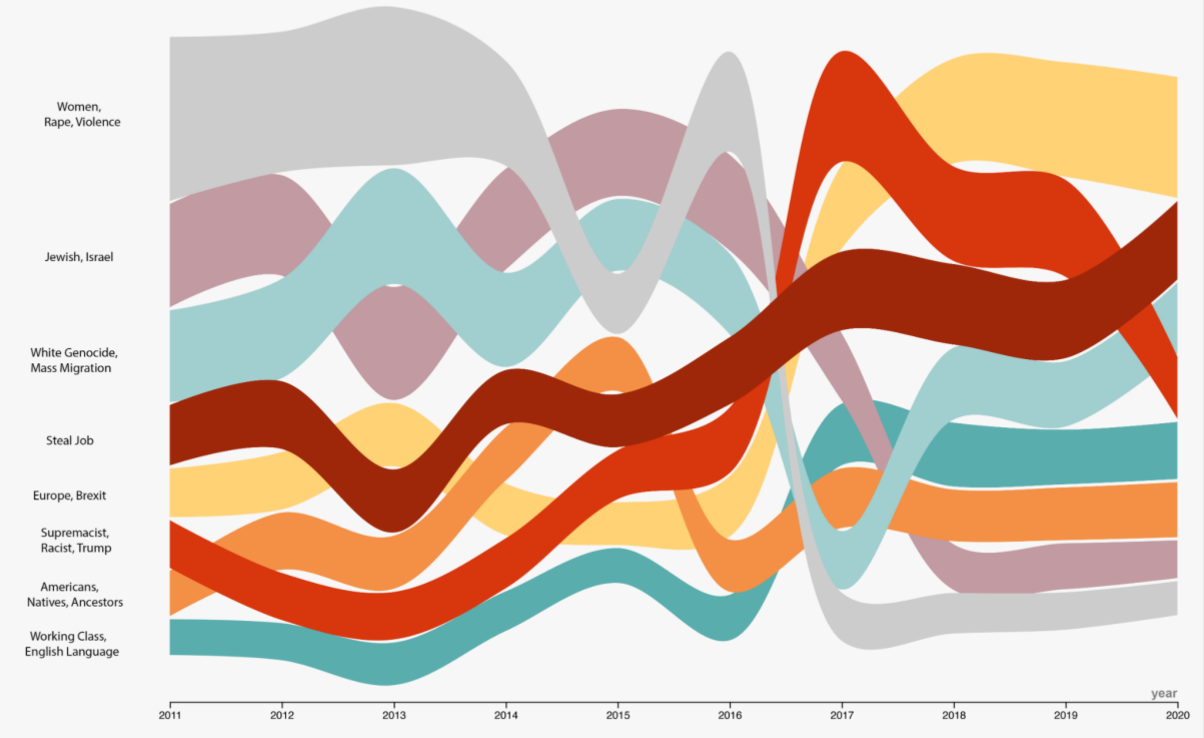


Plot 7: Topic prevalence seen to increase on Twitter 2011 – 2020. The solid line represents the aggregated average topic prevalence per year, calculated based on each of the plotted dots with the same color - these each represent a monthly average per topic.



Plot 8: Topic prevalence seen to decrease on Twitter 2011 – 2020. Interpretation same as Plot 7.

Next, to gain a holistic overview of how the conversations on Twitter had changed in the past decade, topics were also plotted to show the average topic prevalence each year relative to each of the other topics (Plot 9).



Plot 9: Topic distribution (relative volume of content) on Twitter per year. Each line represents a topic. The thickness of the line shows the volume of the discourse each year (as a proportion relative to total number of Tweets). Vertical fluctuation represents topic rank, placing the most prevalent topic across documents each year at the top of the graph. Topics that did not change over time were removed.

The topic models show that all but one of the the eight categories analysed showed a notable and often drastic change in the discourse between 2016 - 2017. At no other point in 2011 - 2020 did the topics shift in complete unison or see such drastic rate of change, indicated as a steep slope.

Entering the 2010's, discussions on immigration in relation to violence against women was the most common latent topic with language similar to that used by white supremacists; by the end of the 2010's was the topic gaining least traction. Equivalently, the typical anti-semitic discourse of white-supremacists was prevalent in the early to mid 2010's, after which it was overshadowed by discussions on racism, white supremacism, mass migration, white genocide and the 'working class'. Notably, the topic of immigrants being problematic for the workforce was seen to slowly escalate over time, being the topic with least fluctuation. Mass migration and ideas of 'white genocide' were among the three most prevelant topics for the first 5 years studied as well as it contiously increased 2018 - 2020, but saw a decrease 2017. This year, discourse was overshadowed by the discourse on Donald Trump, racism and white supremacism, which later slowly declined. The discourse on European immigration presents somewhat of an outlier, as the expected topic proportion would be assumed to be higher around the Brexit referendum, further studies should dismantling more precisely what subjects this topic contains.

DISCUSSION

The next section discusses the posed research questions, and concludes by relating the findings to scholarship which calls for a new paradigm of anticipatory governance for the internet.

Framework Evaluation

To evaluate the extent to which a 'keyness-driven' framework can be used to retroactively surface toxic content from Twitter, the framework was judged on three criteria: that it successfully surfaces toxic content from each year, by an evaluation of the content it is seen to surface, and the impact of the methods limitations. Based on these grounds, applying the 'keyness-driven' framework to retroactively surface Tweets shows astounding promise.

The 'keyness-driven' framework was successful in its retroactive surfacing of toxic content from each year, based on the average toxicity over time being relatively consistent. The consistency is an indicator that the 'key terms' extracted each year also reflected changes in discourse, as dramatic year-on-year fluctuations in language are unlikely without an explainable intervention, such as changes to the content moderation policy on Twitter.

The question of whether the model is surfacing the desired genre of texts is assessed both quantitatively and qualitatively. As presented in the findings, the model extracts texts that are substantially more toxic than both the Twitter reference corpus, as well as scoring an average toxicity well above the posts from Stormfront. The fact that the toxicity score is higher than those extracted from a Stormfront -- posts that are known to be pernicious -- is a solid indicator that the novel framework proves a valuable method for surfacing toxic content from Twitter. Moreover, that the language extracted from Twitter not *only mirrors, but scores much higher* for toxicity than the Stormfront posts empirically may give insight to Klein's (2012) hypothesis that the far-right's intolerant and dogmatic attitudes may have in fact *"slipped into the mainstream"*. Although requiring further study, this method shows promise for initiating research of toxic discourses with help of the 'keyness-driven' framework and contributes with the automatized surfacing of nearly 30,000 instances of toxic language with high average toxicity, by contrast to one of the largest toxic speech corpora which provide 1034 instances (Kolhatkar, 2019).

Next, examining the output from the framework qualitatively, samples of the Tweets suggest that the language captured shares sentiment with the posts from Stormfront. This is exemplified through the use of references to 'forced-assimilation', the endangerment of the 'white race', a rejection of diversity, as well as xenophobic connotations on subjects of fertility and reproduction. However, the framework was also seen to surface false positives. These implications need to be weighed against the purpose of the framework: to surface toxic language for early stage research and guide studies on toxic speech. With this in mind, false positives would not create significant issues as the data is assumed to undergo further processing or expert annotation. That being said, a false positive rate would be needed to understand the breadth of the problem for further studies using similar approaches.

Finally, the key limitation of this method is considered: its heavy reliance on the chi-squared statistic. Before adapting this method efforts should be directed in two areas relating to the sensitivity of the chi-square statistic. First, the effects of various corpus sizes should be examined as the statistic is known to be sensitive to N; this is especially pressing for large-N, big-data research (Gries, 2010). This means that the smallest observed differences between corpora show small p-values, based on a large N. Second, the statistic is also sensitive due to the power-law distribution of the chi-square for keyness analysis (Appendix M), as non-linear statistical relationships amplify results (O'Sullivan, 2016). The implications are that a small drop in rank, especially among the highest ranking terms can also mean a disproportionately large drop in chi-square, and therefore also each term's weighting. Future studies should consider other test statistics for measuring keyness, such as the likelihood-ratio statistic - as advocated for by Gabrielatos (2018).

In sum, the assessed criteria show room for improvement - but none of the raised concerns suggest fundamental flaws in the framework. Rather, they indicate issues known to all data scientists, prevalent in all methods dealing with automating the analysis of natural language. Researchers therefore emphasize how, ultimately, these methodologies rely on thoughtful analysis by the researcher, computers in this context not replacing humans, but *"amplifying human abilities"*, which come at the cost of some error (Grimmer & Stewart, 2017: 270). Given quantitative indicators of success, qualitative and critical reading of individual texts, and assessing the key methodology underpinning the framework and in response to the posed first research question, it is considered to successfully retroactively surface toxic content from Twitter.

Topic Insights - Political Shifts and Changes in the Discourse on Immigration

In response to the second research question, the findings presented an overview of ten topics prevalent in Tweets extracted through the 'keyness-driven' framework, shown to use similar language to that of white supremacists. The following section discusses the way in which these topics have changed over time - thereby exemplifying the framework's potential to contribute to early stage analysis of toxic discourses.

In STM, examining topic fluctuation and model validation go hand-in-hand, as apart from addressing the semantic validity by iterating dozens of times over K and qualitatively assessing semantic coherence, the second validation step is to assess predictive validity (Dehler-Holand *et al*., 2021). This is done by comparing topic time series against real-world events, examining if topics change in accordance with events that are expected to have a large impact on topic-volume.

A striking trend is that all but one of the the eight topics analysed showed a notable, often dramatic, change in the discourse between 2016 - 2017. At no other point during the studied decade did the topics shift in unison or with such velocity.

Many events took place during these years that could be expected to impact the narrative on immigration. The 2010's were a time characterized by the Arab spring uprisings and subsequent power vacuums, Barack Obama serving as the United States first biracial president, and the rapid expansion of the Islamic State of Iraq and Syria. By mid-2016 there had been extremist terrorist attacks in Boston, San Bernardino, Paris and Nice and 1.3 million refugees had entered Europe in the so-called 'migrant crisis'. Populist powers rose to power in the UK, Italy, Hungry and France. Meanwhile, Britain voted to leave the EU, and shortly after Donald Trump was elected president of the United States. Finally, the 'Unite the Right' rally took place in Charlottesville, Virginia; an event marking the 2000's first large-scale public manifestation of the White Nationalist and neo-Nazi allegiances which in turn the heighted the public interest of white supremacist ideologies.

None of these events should be viewed in isolation, but initial findings indicate a dramatic discursive shift on the topic of immigration from those expressing views using language similar to that of white supremacists in 2016 - 2017. The conversation was seen to change from the anti-semitic language of "jews" and the explicit problematizing of a "white genocide" to an emphasis on the working middle class, ideas of jobs being 'stolen' and territorial protectionsm. These preliminary findings deepen the existing knowledge on far-right discourse, which has previously been conceptually described as "blurred" with mainstream politics (Brown *et al*, 2021), but without providing empirical suggestions of how this more concretely is manifested. Furthermore, it provides an exciting avenue for continued interdisciplinary research on the association between harmful language and hate incidents, which so far has been limited to hate-speech (Williams *et al*, 2020). Moving forward, comparativly monitoring changes across discourses' toxicity levels may allow for proactive intervention, as called for by Jurgens *et al*. (2019), which would have significant effect, online and offline, for those currently harmed by to toxic narratives.

CONCLUSION

This research has detailed a new framework for surfacing toxic language retroactively. The surfaced content was also analysed within the well-established STM topic model, illustrating the framework's relevance for understanding how discourses change over time. The construction of this framework was underpinned by the idea that thoughtful and robust keyword selection, appropriate weighting, corpus selection and model application could be creatively synthesized to surface toxic language without using state-of-art language models, which have been challenging for researchers to apply for the study of toxic speech.

Initial findings using the 'keyness-driven' framework suggest that toxicity in the context of immigration increased on Twitter by 28% from 2011 to 2020, despite the advances in language modeling and introduced content moderation efforts - empirically signaling that the degree to which toxic language is dispersed on mainstream platforms over time requires further study. The other finding which warrants specific attention is the remarkable shift in discourse observed between 2016 - 2017, shown to dramatically alter the discursive landscape on immigration and related topics. Notably, these were also years in which the United Kingdom voted to leave the EU, and Donald Trump was elected U.S. President.

To reap the true value of retroactive language modeling, a continuation of this study would focus specifically on examining the drivers of the observed discursive shifts, both by close readings of texts, and situating the changes within the broader socio-political context. This research would also gain from understanding the potential effect caused by changes in the technology and platforms which host these discussions, granting an increasingly critical angle towards tech platforms by, for instance, taking into account how algorithms affect the dissemination of content, which scholars have nodded to have a polarizing effect (Feezell *et al*. 2021). This research charts the path to better understanding the normalisation of toxic discourses, making a unique contribution to the complex field of language modeling - a field that in due course can provide empirical evidence supporting a governing of the internet which not only seeks the absence of hate, but instead takes action in anticipation of harmful narratives taking root.

SUPPLEMENTARY MATERIAL

All data analysis for this research was executed in R. To facilitate an understanding of how the framework was created, a sample of the code used for this research is publically accessible on GitHub: https://bityl.co/8IPB.

REFERENCES

Artwick, C. G. (2013) News sourcing and gender on Twitter, *Journalism 15*(8): 1111-1127.

Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020) Feature selection using an Improved Chi-square for Arabic text classification, *Journal of King Saud University - Computer and Information Sciences 32*(2): 225-231.

Benoit, K. (2020). Text as data: An overview, *The SAGE Handbook of Research Methods in Political Science and International Relations,* 461-497.

Biber, D., Connor, U., & Upton, T. A. (2007) *Discourse on the move using corpus analysis to describe discourse structure*, Amsterdam: John Benjamins Pub.

Blei, D. M. (2012) Probabilistic topic models, *Communications of the ACM 55*(4): 77-84.

Bonilla-Silva, E. (2002) The linguistics of color blind Racism: How to TALK nasty about blacks without Sounding "Racist", *Critical Sociology 28*(1): 41-64.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019) Nuanced metrics for measuring unintended bias with real data for text classification, *Companion Proceedings of The 2019 World Wide Web Conference*.

Burnap: , & Williams, M. L. (2016) Us and them: Identifying cyber hate on twitter across multiple protected characteristics, *EPJ Data Science 5*(1).

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017) You can't stay here, *Proceedings of the ACM on Human-Computer Interaction 1*(CSCW).

Coats, S. (2021) 'Bad language' in the NORDICS: Profanity and gender in a social media corpus, *Acta Linguistica Hafniensia 53*(1): 22-57.

Combating political hate speech online. (2020) *Online Political Hate Speech in Europe,* 196-212.

Davidson, T., Warmsley, D., Macy, M., & Weber4, I. (2017) *Automated Hate Speech Detection and the Problem of Offensive Language.*

Deng, Q., Hine, M., Ji, S., & Sur, S. (n.d.) Inside the Black Box of Dictionary Building for Text Analytics: A Design Science Approach, *Journal of International Technology and Information Management 27*(3).

Dentice, D. (2019) "So much for darwin" - an analysis of stormfront discussions on race, *Journal of Hate Studies 15*(1): 133.

Fairclough, N. (1989) *Language and power*, Harlow: Longman.

Fersini, E., Nozza, D., & Rosso (2018) Overview of the evalita 2018 task on automatic misogyny identification (ami), *EVALITA Evaluation of NLP and Speech Tools for Italian,* 59-66.

Fišer, D., Erjavec, T., & Ljubešić, N. (2017) Legal framework, dataset and Annotation schema for socially Unacceptable online discourse practices in Slovene, *Proceedings of the First Workshop on Abusive Language Online*.

Fortuna: , & Nunes, S. (n.d.). A Survey on Automatic Detection of Hate Speech in Text, *ACM Computing Surveys 51*(4).

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Kourtellis, N. (2018) Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior, *Twelfth International AAAI Conference on Web and Social Media*.

Gabrielatos, C. (2018) Keyness analysis. *Corpus Approaches To Discourse,* 225-258.

Gilda, S., M. S., L. G., & Oliviera, D. (2021). Predicting Different Types of Subtle Toxicity in Unhealthy Online Conversations, *ArXiv*.

Grimmer, J., & Stewart, B. M. (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts, *Political Analysis 21*(3): 267-297.

Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., & Margetts, H. (2021) An expert annotated dataset for the detection of online misogyny, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Hammer, H. L. (2017) Automatic detection of hateful comments in online discussion. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering,* 164-173.

Han, X., & Tsvetkov, Y. (2020) Fortifying toxic speech detectors against veiled toxicity, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Hino, A., & Fahey, R. A. (2019) Representing the Twittersphere: Archiving a representative sample of Twitter data under resource constraints, *International Journal of Information Management 48*: 175-184.

Holgate, E., Cachola, I., Preoţiuc-Pietro, D., & Li, J. J. (2018) Why swear? analyzing and inferring the intentions of vulgar expressions, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Holt, T. J., Freilich, J. D., & Chermak, S. M. (2020) Examining the online expression of ideology among far-right extremist forum users, *Terrorism and Political Violence,* 1-21.

Holt, T. J., Freilich, J. D., & Chermak, S. M. (2020) Examining the online expression of ideology among far-right extremist forum users, *Terrorism and Political Violence,* 1-21.

Holtz: , Kronberger, N., & Wagner, W. (2012) Analyzing internet forums, *Journal of Media Psychology 24*(2): 55-66.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science, *American Journal of Political Science 54*(1): 229-247.

Immigration. (n.d.) URL, https://dictionary.cambridge.org/dictionary/english/immigration [Last consulted August 19, 2021]

Jegadeesh, N., & Wu, A. D. (2011) Word power: A new approach for content analysis, *SSRN Electronic Journal*.

Jurgens, D., Hemphill, L., & Chandrasekharan, E. (2019) A just and comprehensive strategy for using nlp to address online abuse, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Kennedy, H. (2016). *Post, mine, REPEAT: Social media data mining becomes ordinary*, Basingstoke: Palgrave Macmillan.

Kim, H., Jang, S. M., Kim, S., & Wan, A. (2018) Evaluating sampling methods for content analysis of twitter data, *Social Media + Society 4*(2).

Klein, A. (2012) Slipping racism into the mainstream: A theory of information laundering, *Communication Theory 22*(4): 427-448.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2019) The sfu opinion and comments corpus: A corpus for the analysis of online news comments, *Corpus Pragmatics 4*(2): 155-190.

Krippendorff, K. (1980) *Content analysis an introduction to its methodology*, Beverly Hills: SAGE.

Lebryk, T. (2021, April 18) Introduction to the Structural topic Model (stm), URL: https://towardsdatascience.com/introduction-to-the-structural-topic-model-stm-34ec4bd5383 [Last consulted August 19, 2021]

Lewis, S. C., Zamith, R., & Hermida, A. (2013) Content analysis in an ERA of big data: A hybrid approach to computational and manual methods, *Journal of Broadcasting & Electronic Media 57*(1): 34-52.

Liao, X., Chen, H., Wei, J., Yu, Z., & Chen, G. (2014) A weighted lexicon-based generative model for opinion retrieval, 2014 International Conference on Machine Learning and Cybernetics.

Lucas, B. (n.d.) *Methods for monitoring and mapping online hate speech* (Rep.), GSDRC.

Mahmud, A., Ahmed, K., & Khan, M. (2008) *Detecting flames and insults in text*.

Malmasi, S., & Zampieri, M. (2017). Detecting hate speech in social media, *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*.

Marantz, A. (2020) *Antisocial: How ONLINE extremists broke America*, Basingstoke: Picador.

Marantz, A. (2020, October 09) Why facebook can't fix itself. URL: https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself [Last consulted August 15, 2021]

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2015) Using twitter for demographic and social science research: Tools for data collection and processing, *Sociological Methods & Research 46*(3): 390-421.

Meddaugh: M., & Kay, J. (2009). Hate speech OR "Reasonable Racism?" the other in Stormfront, *Journal of Mass Media Ethics 24*(4): 251-268.

Milutinović, V., & Kotlar, M. (2021) *Handbook of research on methodologies and applications of supercomputing*, Hershey, PA: Engineering Science Reference, an imprint of IGI Global.

Mishra, S. (2017, May 21). Unsupervised learning and data clustering. URL: https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a [Last consulted August 19, 2021]

Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive language detection on arabic social media, *Proceedings of the First Workshop on Abusive Language Online*.

Munezero, M., Mozgovoy, M., Kakkonen, T., & Klyuev, V. (2013) *Antisocial behavior corpus for harmful language detection*.

Munger, K. (2016). Tweetment effects on THE TWEETED: Experimentally Reducing Racist harassment, *Political Behavior 39*(3): 629-649.

Nario, J., Lees, A., Kivlichan, I., Borkan, D., & Goyal, N. (2021) Capturing Covertly Toxic Speech via Crowdsourcing, *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content, *Proceedings of the 25th International Conference on World Wide Web*.

Noever, D. (2018) *Machine Learning Suites for Online Toxicity Detection*.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: A systematic review, *Language Resources and Evaluation 55*(2): 477-523.

Price, I., Gifford-Moore, J., Flemming, J., Musker, S., Roichman, M., Sylvain, G., Sorensen, J. (2020) Six attributes of UNHEALTHY CONVERSATIONS, *Proceedings of the Fourth Workshop on Online Abuse and Harms*.

Refugees, asylum-seekers and migrants. (2021, June 01), URL: https://www.amnesty.org/en/what-we-do/refugees-asylum-seekers-and-migrants/ [Last consulted August 19, 2021].

Risch, J., Schmidt: , & Krestel, R. (2021) Toxic comment COLLECTION: Making more than 30 datasets easily accessible in one unified format, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*.

Roberts, M., Stewart, B., Tingley, D., & Airoldi, E. (2013) *The Structural Topic Model and Applied Social Science*.

Röttger: , Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021) HateCheck: Functional tests for hate Speech detection models, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Saleem, H., Dillon, K., Benesch, S., & Ruths, D. (2017) A Web of Hate: Tackling Hateful Speech in Online Social Spaces, *Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)*.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020) Social bias FRAMES: Reasoning about social and Power implications of language, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Schmidt, A., & Wiegand, M. (2017) A survey on hate speech detection using natural language processing, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*.

Scott, M. (1997). PC analysis of key words — and key key words, *System 25*(2): 233-245.

Silge, J., & Robinson, D. (2016) Tidytext: Text mining and analysis using tidy data principles in r, *The Journal of Open Source Software 1*(3): 37.

Sood, S., Antin, J., & Churchill, E. (2012) Profanity use in online communities, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Soroka, S., Young, L., & Balmas, M. (2015) Bad news or Mad News? Sentiment scoring of negativity, fear, and anger in news content, *The ANNALS of the American Academy of Political and Social Science 659*(1): 108-121.

Spertus, E. (1997) Smokey: Automatic recognition of hostile messages, *In Proceedings of the Four- Teenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*.

STUBBS, M. (1994) Grammar, text, and IDEOLOGY: Computer-assisted methods in the linguistics of representation, *Applied Linguistics 15*(2): 201-223.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011) Lexicon-based methods for sentiment analysis, *Computational Linguistics 37*(2): 267-307.

Takahashi, B., Tandoc, E. C., & Carmichael, C. (2015) Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines, *Computers in Human Behavior 50*: 392-398.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., & Daelemans, W. (2016) The Automated Detection of Racist Discourse in Dutch Social Media, *Computational Linguistics in the Netherlands Journal* 6: 3-20

Törnberg, A., & Törnberg: (2016) Muslims in social MEDIA discourse: Combining topic modeling and critical discourse analysis, *Discourse, Context & Media, 13*: 132-142.

United nations office on GENOCIDE prevention and the responsibility to protect. (n.d.) URL: https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml [August 15, 2021]

Using machine learning to reduce toxicity online. (n.d.) URL: https://www.perspectiveapi.com/how-it-works/ [Last consulted August 15, 2021]

Van Dijk, T. A. (1993) Principles of critical discourse analysis, *Discourse & Society 4*(2): 249-283.

Vidgen, B., & Derczynski, L. (2020) Directions in abusive language training data, a systematic review: Garbage in, garbage out, *PLOS ONE 15*(12).

Vidgen, B., & Yasseri, T. (2019) Detecting weak and strong islamophobic hate speech on social media, *Journal of Information Technology & Politics 17*(1): 66-78.

Vidgen, B., Hale, S., Staton, S., Melham, T., Margetts, H., Kammar, O., & Szymczak, M. (2020) Recalibrating classifiers for interpretable abusive content detection, Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science.

Waqas, A., Salminen, J., Jung, S., Almerekhi, H., & Jansen, B. J. (2019) Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate, *PLOS ONE 14*(9).

Warner, W., & Hirschberg, J. (2010) *Detecting Hate Speech on the World Wide Web*.

Waseem, Z., & Hovy, D. (2016) Hateful symbols or hateful people? Predictive features for hate speech detection on twitter, *Proceedings of the NAACL Student Research Workshop*.

Williams, M. L., Burnap: , Javed, A., Liu, H., & Ozalp, S. (2019) Hate in the Machine: Anti-Black and Anti-muslim social media posts as predictors of Offline racially and religiously AGGRAVATED Crime, *The British Journal of Criminology*.

Williams, M. L., Burnap: , Javed, A., Liu, H., & Ozalp, S. (2020) Hate in the machine: Anti-black and Anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime, *The British Journal of Criminology*.

Wong, M. A., Frank, R., & Allsup, R. (2015) The supremacy of online white supremacists – an analysis of online discussions by white supremacists, *Information & Communications Technology Law 24*(1): 41-73.

Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012) Detecting offensive tweets via topical feature discovery over a large scale twitter corpus, *Proceedings of the 21st ACM International Conference on Information and Knowledge Management - CIKM '12*.

Xu, J., Jun, K., Zhu, X., & Bellmore, A. (n.d.) Association for Computational Linguistics, in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*.

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Suarez-Tangil, G. (2018) On the origins of memes by means of Fringe web communities, *Proceedings of the Internet Measurement Conference 2018*.

APPENDICES

Appendix A - methodological details on existing studies of discourse prevalent on Stormfront

Brown, 2009 (qualitative) - cherry-picks single comments which are used for analysis from three different sites (one of them being Stormfront).

Meddaugh & Kay 2009 (qualitative) - present a theoretical argument based on a rhetorical analysis of messages on Stormfront, thus not systematic or provide insight into the language more generally or trends overtime.

Figea et al., 2016 (quantitative) - study of narrow scope as they only look at the use of affect (worries, aggression, racism) in three subforums

Dentice, 2018 (qualitative) - research specifically focuses on the narrow scope of Stormfront members' discussion of the Trump presidency.

Gibert et al., 2018 (quantitative) - annotate 10,000 sentences into binary classification hate/no hate, use a large sample spanning many years but do not use year as a variable in itself to depict changes.

Appendix B - Visual illustration of "information laundering" (Klein, 2012: 435)

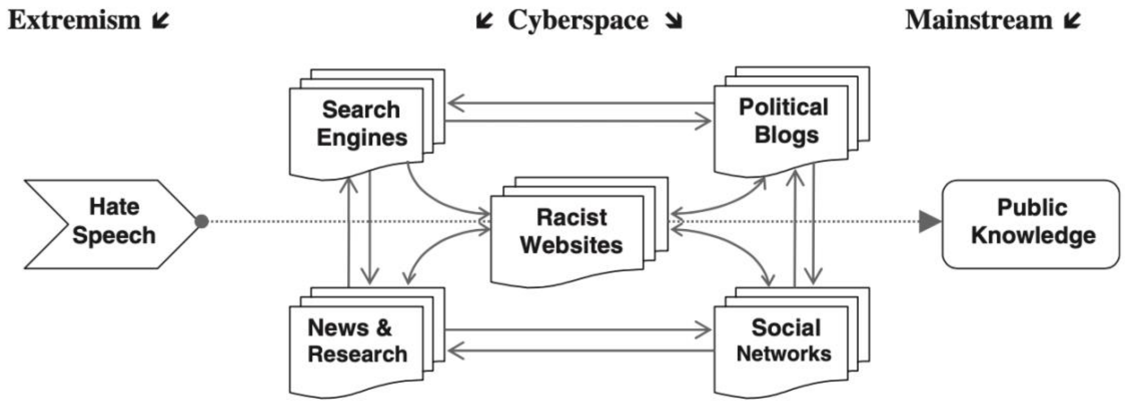


Figure 1 Model of information laundering.

Appendix C - Taxonomy created to capture 'immigration'

```
("refug*", "immig*", "migra*", "immagrants", "immagrant", "immogrants",
"immogrants", "immirgration", "immirgrate", "immirgants", "imagration",
"immgiration", "imergration", "immgration", "immegration", "immigrtion",
"imigrent", "imigrents", "imigrants", "imigrant", "migrent", "rufugees",
"yellow invaders", "yellow invader", "brown invader", "brown invaders",
"border bunnies", "border hoppers", "border jumpers", "black invaders",
"black invader", "anchor babies", "reffo", "border nigger", "cab nigger",
"anchor baby", "border hopper", "wetback")
```

Appendix D - Sample of key term matches using wildcards

IMMIGR*	REFUGE*	MIGRA*
[1] "immigration"	[1] "refugees"	[1]"migration"
[2] "immigrant"	[2] "refugee"	[2]"migrations"
[3] "immigrants"	[3] "refugee-shy"	[3]"migrationwatchuk"
[4] "immigrated"	[4] "refuge"	[4]"migrated"
[5] "immigrant-wise"	[5] "refugees-their"	[5]"migrate"
[6] "immigrate"	[6] "refugeees"	[6]"migrationpolicy.org"
[7] "immigratsiooni"	[7] "refugee-centered"	[7]"migrating"
[8] "immigrating"	[8] "refuges"	[8]"migrates"
[9] "immigrant87"	[9] "refugee's"	[9]"migrationsverket"
[10] "immigrant's"	[10] "refuge-"	[10]"migratory"
[11] "immigration-gumball"	[11] "refugeerush"	[11]"migration.check"
[12] "immigrationist"	[12] "refugeeing"	[12]"migrationwatch"
[13] "immigratio"	[13] "refugee-kicking"	[13]"migrators"
[14] "immigration.is"	[14] "refugees-"	[14]"migration-watch"
[15] "immigrant-loving"	[15] "refugee-ism"	
[16] "immigrant-target"	[16] "refugeeresettlementwatch"	
[17] "immigrants-to-be"	[17] "refugee-crisis"	
[18] "immigration-related"	[18] "refugee-free"	
[19] "immigration-"	[19] "refugee_scheme"	
[20] "immigrant71"	[20] "refugee-terrorist"	
[21] "immigrations"	[21] "refugee-leeches"	
[22] "immigrant-owned"	[22] "refugess"	
[23] "immigrunts"	[23] "refugees.look"	
[24] "immigrats"	[24] "refugee'-madness"	
[25] "immigratiion"	[25] "refugeess"	

Appendix E - Examples of highest and lowest scored Stormfront posts

Text	**Toxicity Score**
"you people bitch about us using the words ******, kike, spick, wetback, the list goes on and on but its ok for you cal everyone on here a nazi, a honkey, or a cracker."	0.99
"refugees my ass."	0.98
"with every new low paid immigrant hire, with every stupid wetback who buys their imported chink junk, their "sufferink" increases exponentially."	0.98
"then came the first wave of "mass immigration" (black commonwealth) who initially used public resources such as the health and education systems but who usually worked and saw to their own housing requirements."	**> 0.01**
"as tensions rise, immigrants in athens are faced with resentment in graffiti on building walls and from the fluttering red and black golden dawn flags that feature an ancient greek geometric-era symbol."	**> 0.01**
"the group, which included germans, italians and austrians as well as french activists, all wore blue padded jackets marked "defend europe" and were backed by two rented helicopters which scoured the area for migrants."	**> 0.01**

Appendix F - Extended list of stopwords

```
"abaft, aboard, aboon, about, above, across, adown, aff, afore, after, against,
alongside, along, amidst, amid, amongst, among, anear, anenst, anent, around,
aslant, astraddle, astride, as, athwart, atop, atween, at, a, barring, bar, bating,
before, behind, below, beneath, ben, besides, beside, between, betwixt, beyond,
but, by, chez, circa, concerning, considering, contra, cum, d', despite, des, de,
di, down, during, ere, excepting, except, ex, failing, fer, forby, fore, fornenst,
fornent, forth, for, frae, from, inby, inside, into, in, lacking, less, like,
maugre, midst, mid, minus, natheless, near-hand, near, neath, next, nigh,
notwithstanding, o', o'er, off, of, onto, on, opposite, opuscule, outshout,
outside, outwith, out, over, pace, past, pending, per, plus, pro, qua, reference,
regarding, respecting, re, roundabout, round, sans, save, saving, secundum, senza,
since, sine, sith, sur, syne, tae, than, thro', throughout, through, thro, thru,
thwart, till, touching, towards, toward, to, underneath, under, unless, unlike,
until, unto, up-and-down, upon, up, versus, via, vice, visard, wantage, wanting,
while, withal, within, without, with, albeit, although, and/or, and, an, because,
both, directly, either, ergo, et, forasmuch, forwhy, hence, howbeit, howe'er,
however, howsoever, if, immediately, lest, moreover, neither, nevertheless,
nonetheless, nor, now, once, ophiuchus, otherwise, provided, providing, quoties,
seeing, sobeit, so, that, therefore, though, tho, ubi, vel, whenas, whencesoever,
whene'er, whenever, whensoever, when, where'er, where's, whereas, wheresoever,
wherethrough, whereunto, whereupon, wherever, where, whether, whiles, whilst,
whithersoever, whither, why, yet, ain, all, ane, another, any, billion, certain,
divers, dozen, each, eighteen, eighty, eight, eleven, else, enough, every, few,
fifteen, fifty, five, forty, fourscore, fourteen, four, galore, half, her, his,
hundred, its, least, littler, littlest, little, many, million, mine, more, most,
much, my, nethermost, nineteen, ninety, nine, no, n, one, other, our, own, plenty,
quadrillion, seventeen, seventy, seven, several, sixty, six, some, such, sundry,
ten, their, them, these, the, thine, thirteen, thirty, this, those, thousand,
threescore, three, thy, trillion, twain, twelve, twenty, two, umpteen, various,
whatever, what, wheen, whichever, which, whose, ye, yonder, yon, your, zillion,
allyou, anybody, anyone, anything, aught, baith, couple, everybody, everyone,
everything, fewer, haec, her'n, herself, hers, he, himself, him, his'n, hisself,
hoc, hoo, idem, ilka, itself, it, i, lot, me, myself, nane, no-one, noblewoman,
nobody, none, nothing, oneself, our'n, ourself, ourselves, ours, owt, quibus, self,
she, somebody, someone, something, succussion, thae, thee, theirself, theirselves,
theirs, themselves, there, they, thir, thou, thyself, tother, un, us, we, whate'er,
whatsoe'er, whatsoever, whence, whereby, wherefrom, whereinto, wherein, whereof,
whereon, whereto, wherewithal, wherewith, whichsoever, whoever, whomever,
whomsoever, whom, whosesoever, whosoever, whoso, who, ya, you-all, your'n,
yourself, yours, youse, yous, you"
```

Appendix G - Terms in common for Stormfront corpus paired by year

```
2011 - 2012 (39 terms)

[1] "white"           "non-whit"        "jew"             "white_countri"   "black"
"negro"           "countri"

 [8] "muslim"          "rape"            "anti-whit"       "third_world"     "scum"
"race"            "white_nation"

[15] "non_white"       "genocid"         "white_peopl"     "traitor"
"jewish"          "third"           "mass"

[22] "white_race"      "destroy"         "ame"             "u."
"parasit"         "nonwhit"         "countri_white"

[29] "non-whit_white"  "non"             "massiv"          "mass_non-whit"
"immigr"          "everi_white"     "liber"

[36] "assimil"         "massiv_non-whit" "islam"           "oper_wetback"



2013 - 2014 (38 in common)


 [1] "non-whit"          "white"             "jew"               "white_countri"
"anti-whit"         "negro"

 [7] "genocid"           "scum"              "white_nation"      "muslim"
"third_world"       "white_peopl"

[13] "non_white"         "black"             "white_genocid"     "countri"
"rape"              "white_race"

[19] "everi_white"       "invad"             "immigr"            "parasit"
"nonwhit"           "race"

[25] "jewish"            "massiv_non-whit"   "youtub_ame"        "3rd_world"
"non"               "mestizo"

[31] "traitor"           "countri_white"     "non-whit_white"    "race_mix"
"non-whit_forc"     "non-whit_countri"

[37] "u."                "brown_invad"
```

2015- 2016 (24 in common)

```
 [1] "non-whit"      "white"          "ame"            "negro"          "nonwhit"
"jew"            "white_countri"

 [8] "youtub_ame"    "anti-whit"      "third_world"    "white_nation"   "non_white"
"e2_u"           "u."

[15] "white_race"    "u_u"            "e2"             "rape"           "u"
"mail_onlin"     "u_s"

[22] "flood_white"   "u_t"            "news_daili"

>
```

2017 - 2018 (19 in common)

```
[1] "non-whit"      "white"          "negro"          "ame"            "jew"
"white_countri" "mail_onlin"

 [8] "brown_invad"   "youtub_ame"     "nonwhit"        "u."             "anti-whit"
"oper_wetback"   "wetback"

[15] "third_world"   "u"              "u_u"            "mass_non-whit" "u_t"
```

2019 - 2020 (19 in common)

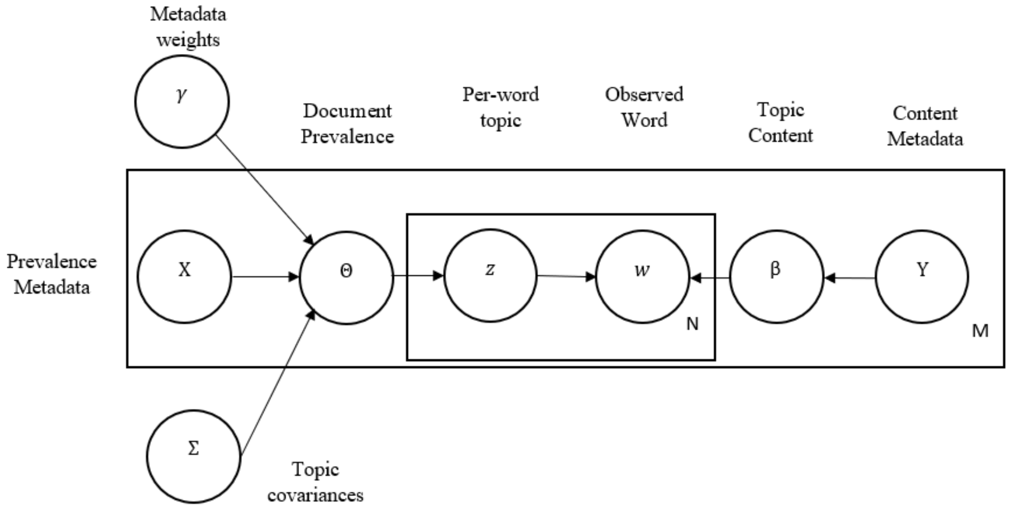
```
[1] "non-whit"        "negro"           "u_t"             "jew_destroy"
"scum_jew"        "evil_scum"       "anti-whit"

 [8] "destroy_england" "oper_wetback"    "england_flood"   "jew"
"white"           "brown_invad"     "u_s"

[15] "illeg_africa"    "don_u"           "wetback_ii"      "flood_black"
```

Comparing across the 10 years of data, instead of pairwise:

```
Only "white"     "non-whit"  "jew"       "negro"     "anti-whit" are the only terms
```

Appendix H - STM plate notation and posterior distribution (Roberts et al., 2013)

Plate notation

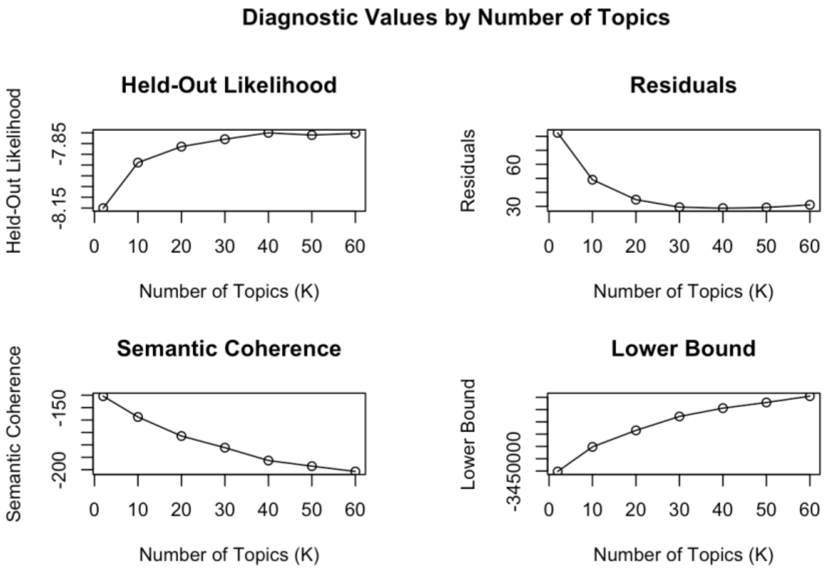


<u>*Posterior distribution*</u>

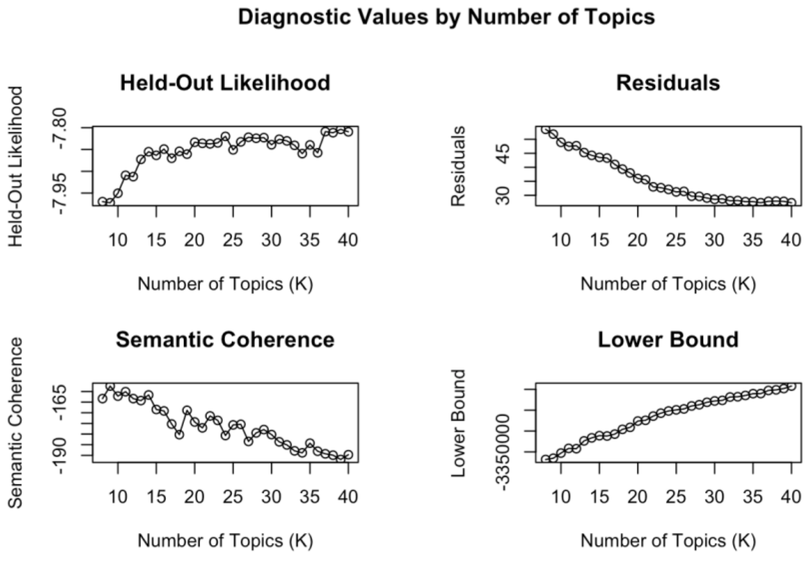
$$P(\eta, z, \kappa', \gamma, \Sigma | W, X, Y, k, \mathrm{m}) \propto$$
$$[\prod_{\mathrm{i}=1}^{D} Normal(\eta_{\mathrm{i}} | X_{\mathrm{i}}, \gamma, \Sigma)\left(\prod_{n=1}^{N} Multinomial\left(\mathrm{z}_{n,\mathrm{i}} \middle| \Theta_{\mathrm{i}}\right) \mathrm{x} Multinomial\left(w_n | \beta_{\mathrm{i},k=z_{d,n}}\right)\right)] \mathrm{x} \prod \mathrm{p}(\kappa) \prod \mathrm{p}(\tau)$$

APPENDIX I - Model diagnostics for variations in k

I. Estimating *k* = 10, 20, 30, 40, 50, 60

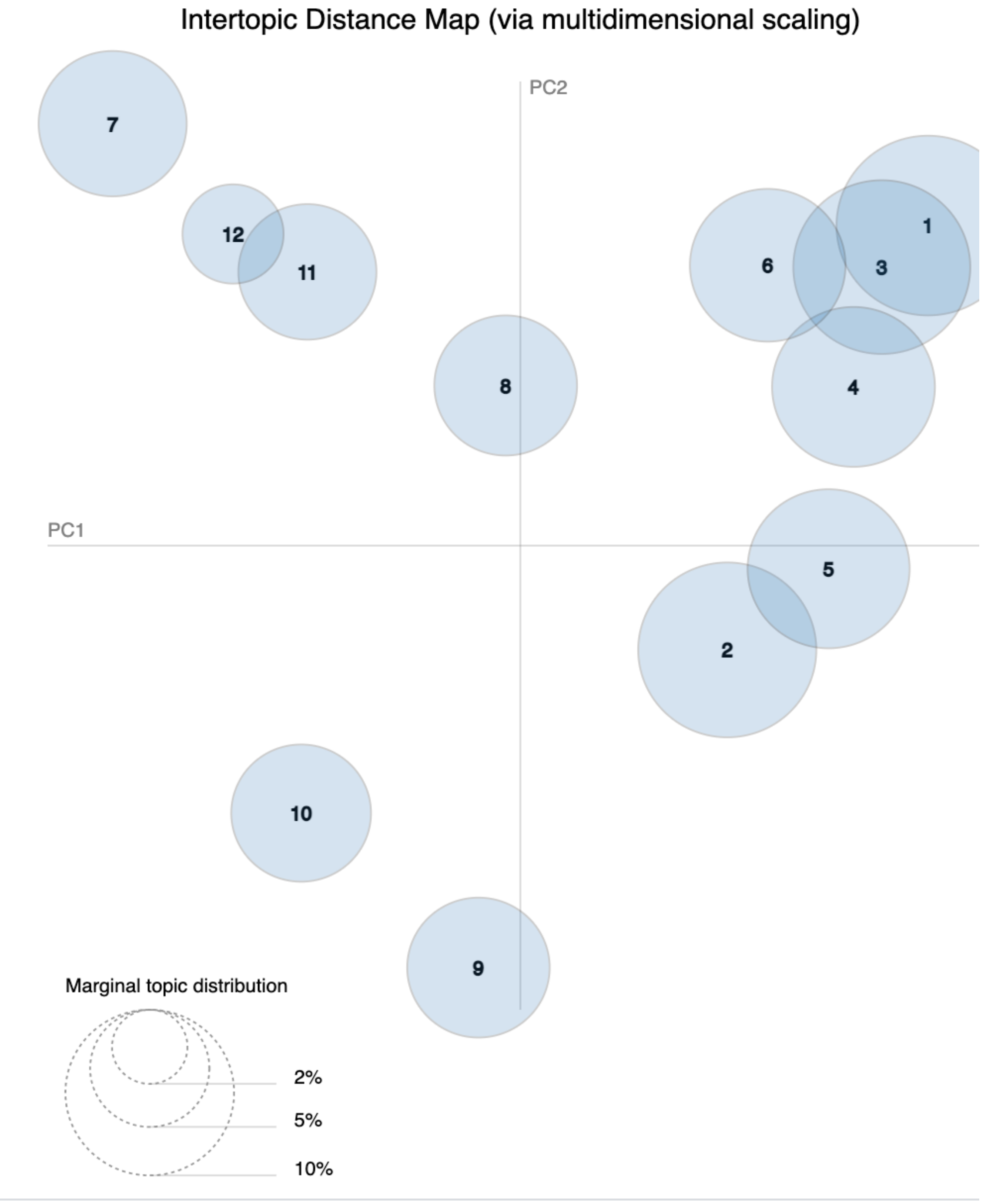


II. Topics ranging from *k* = 1 - 40

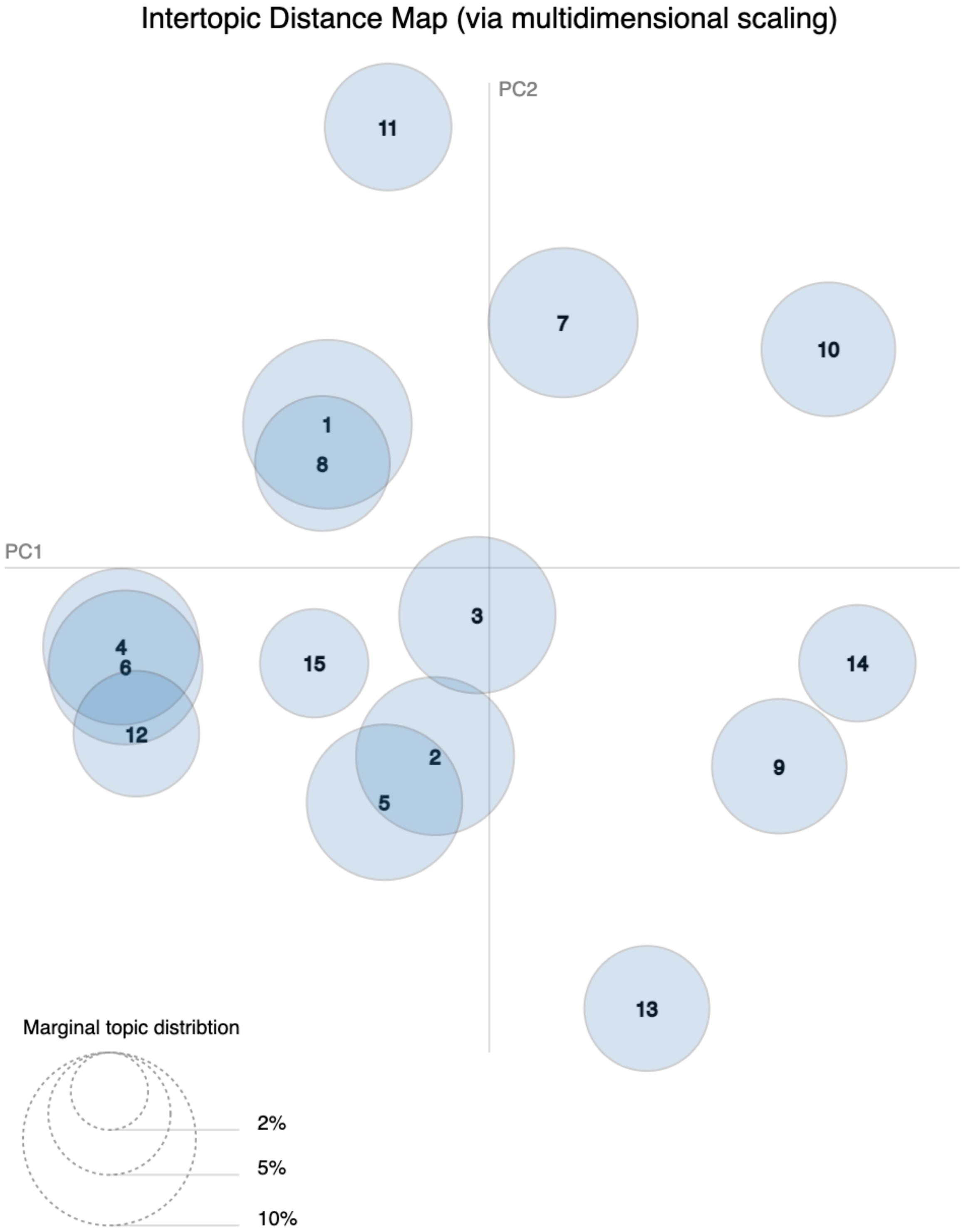


Appendix K - Iterations of Topic Models using various *k*

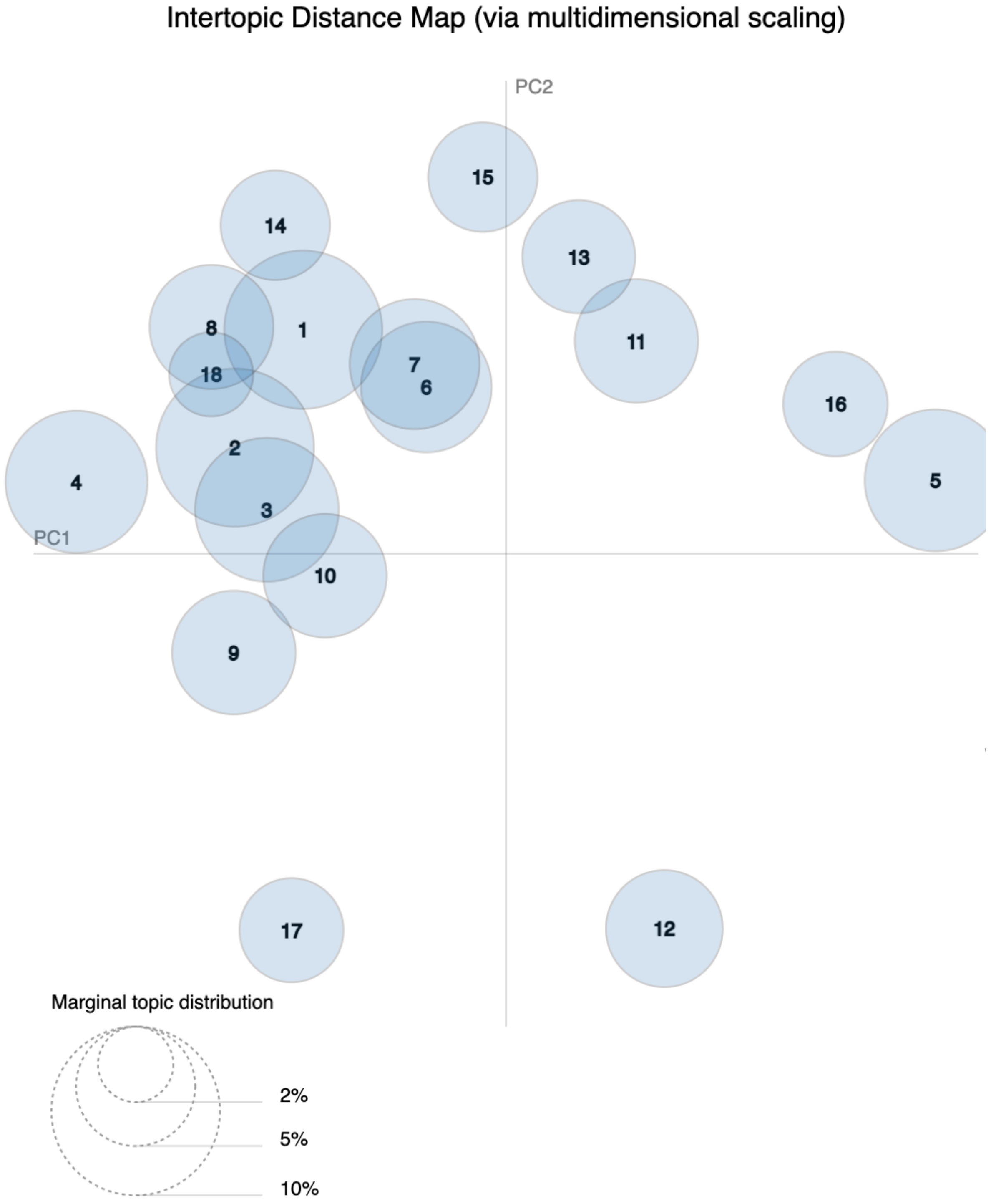
TOPIC MODEL WITH 12 TOPICS



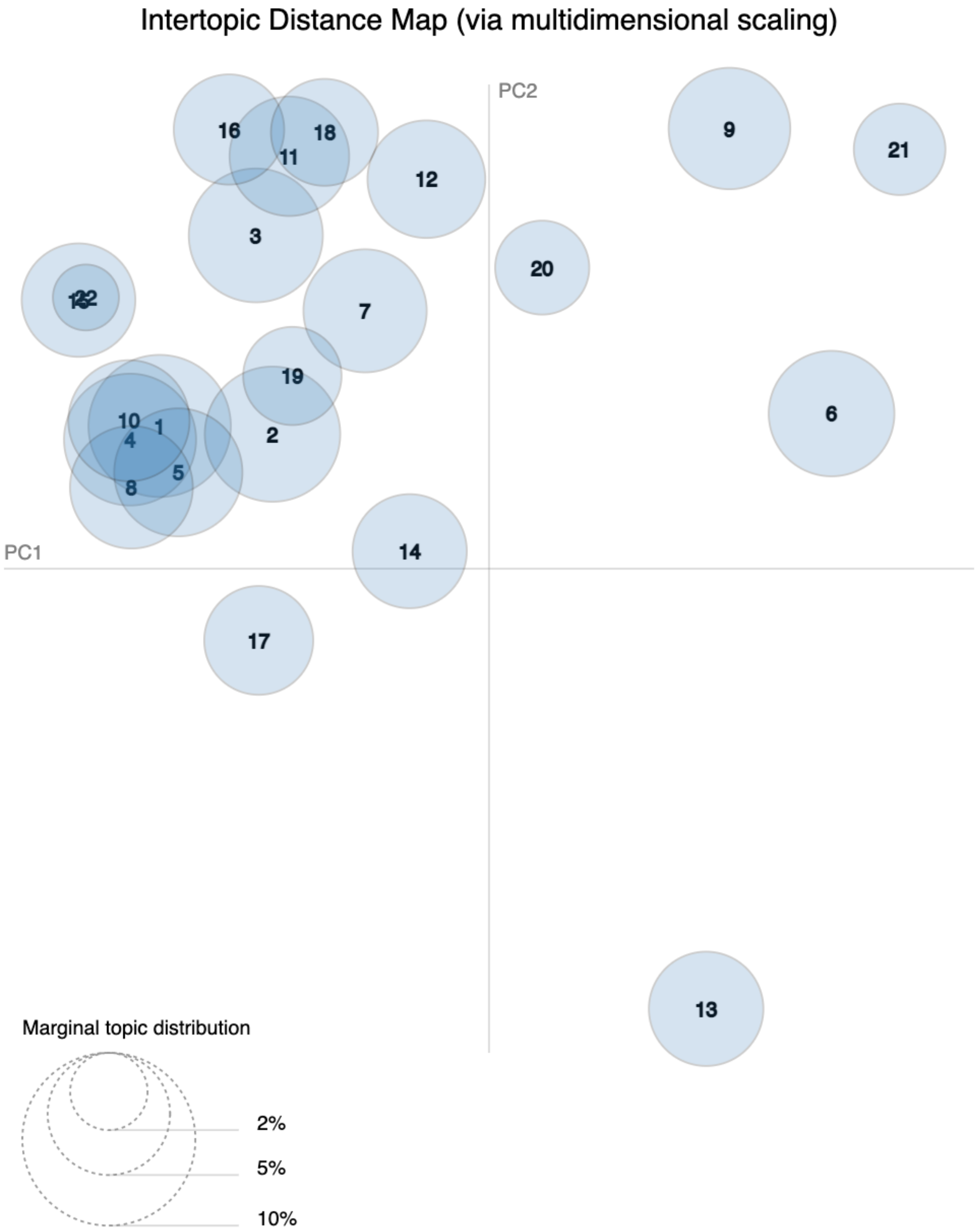
TOPIC MODEL WITH 15 TOPICS



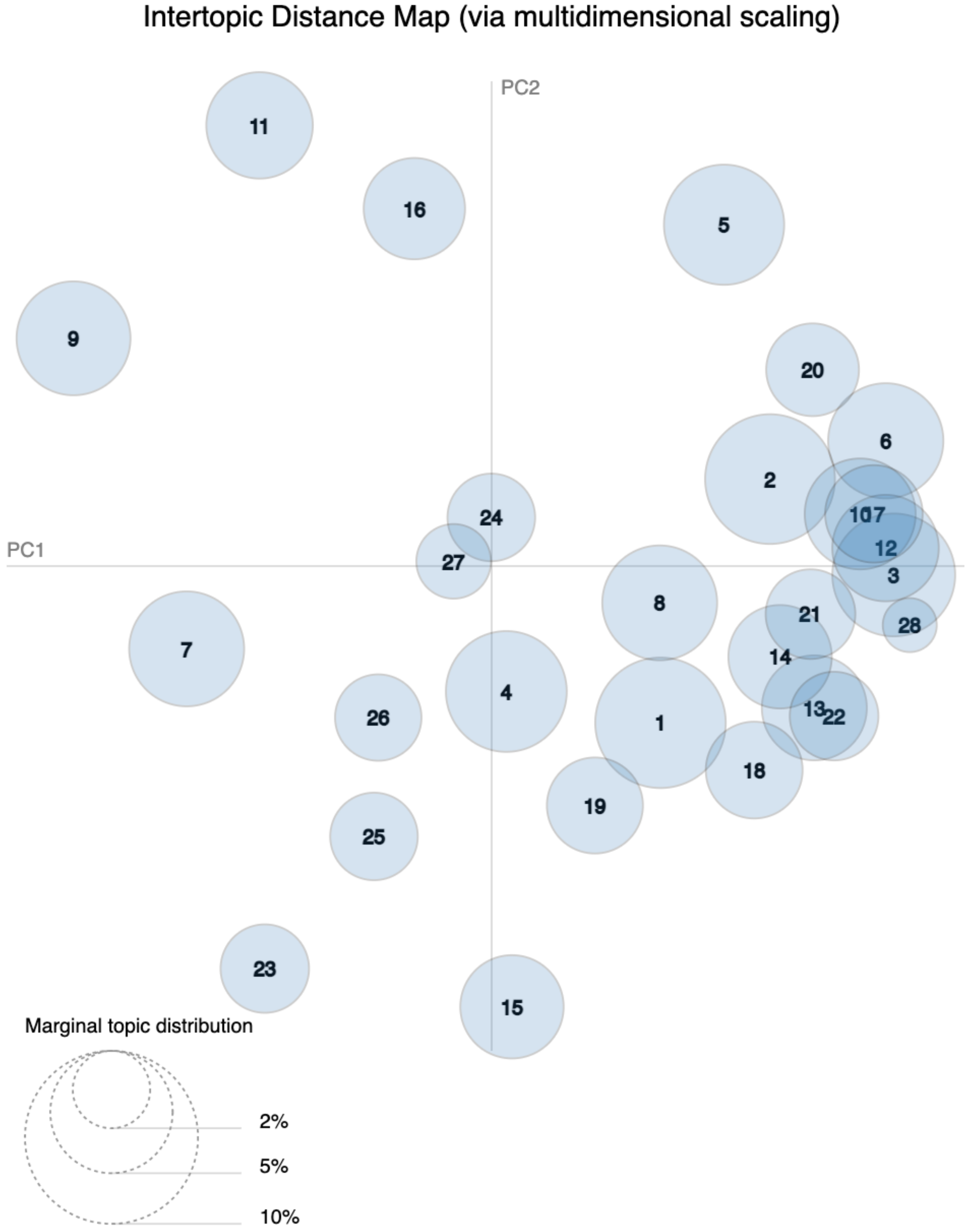
TOPIC MODEL WITH 18 TOPICS



TOPIC MODEL WITH 22 TOPICS

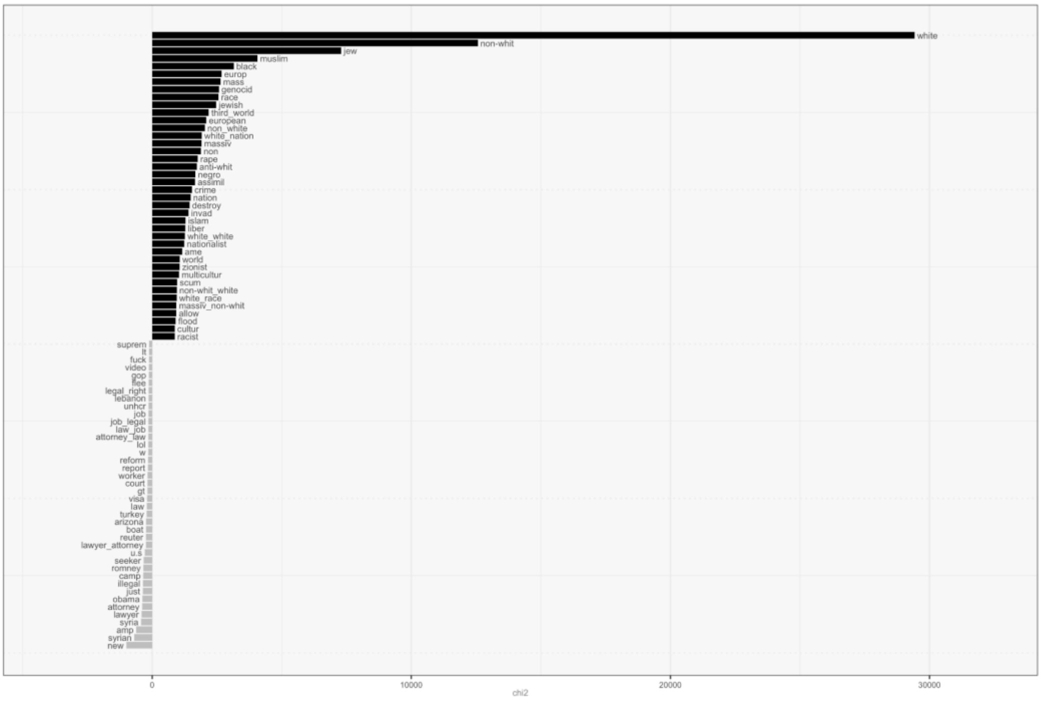


TOPIC MODEL WITH 28 TOPICS

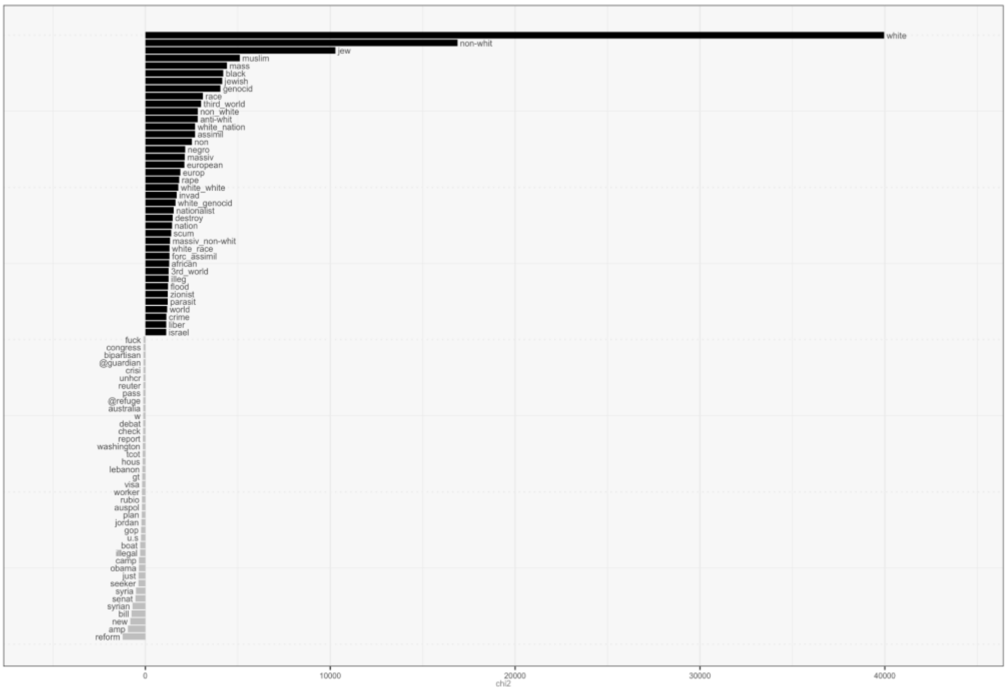


Appendix J - KEYNESS ANALYSIS 2012 - 2019:

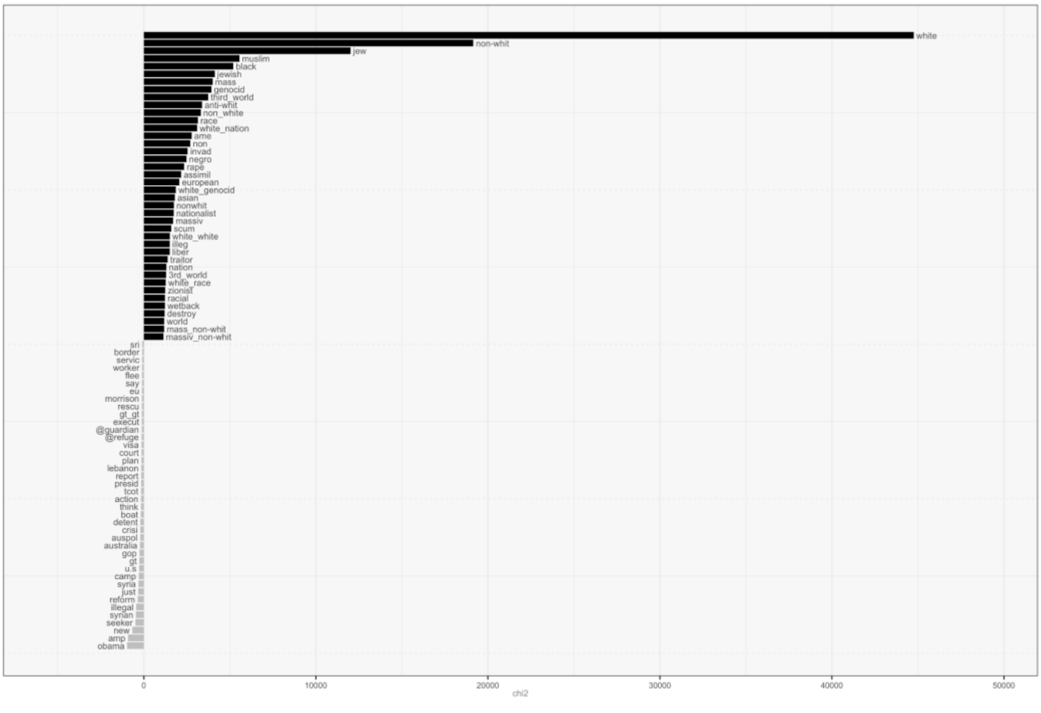
Keyness Analysis 2012 [Target: Toxic posts on Stormfront, Reference: Twitter]



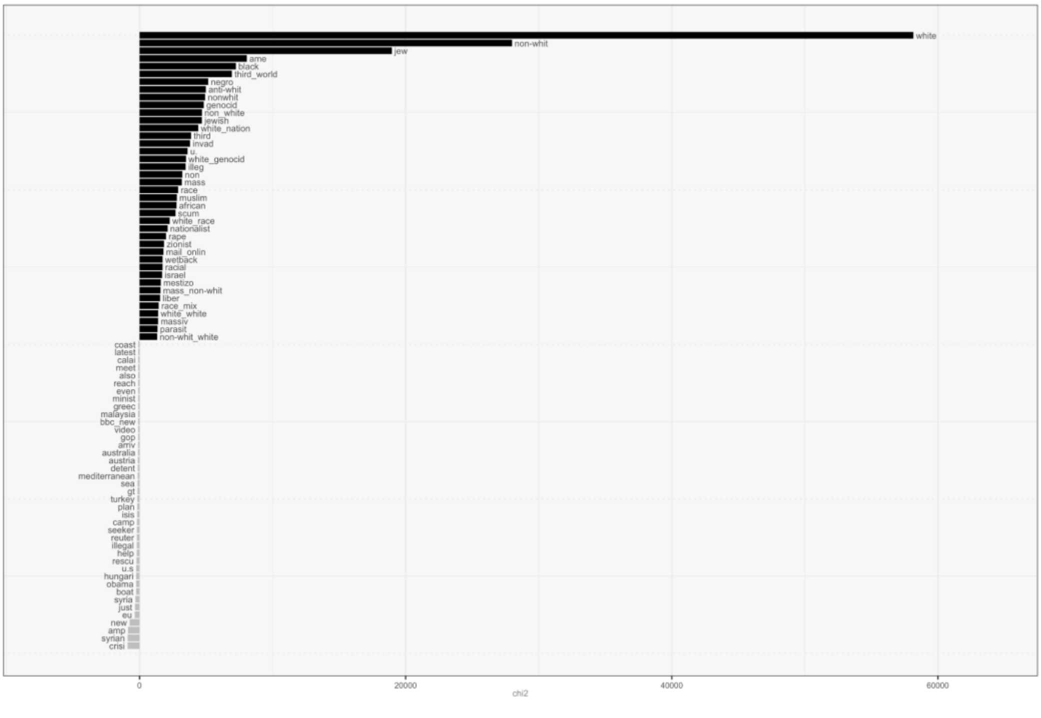
Keyness Analysis 2013 [Target: Toxic posts on Stormfront, Reference: Twitter]



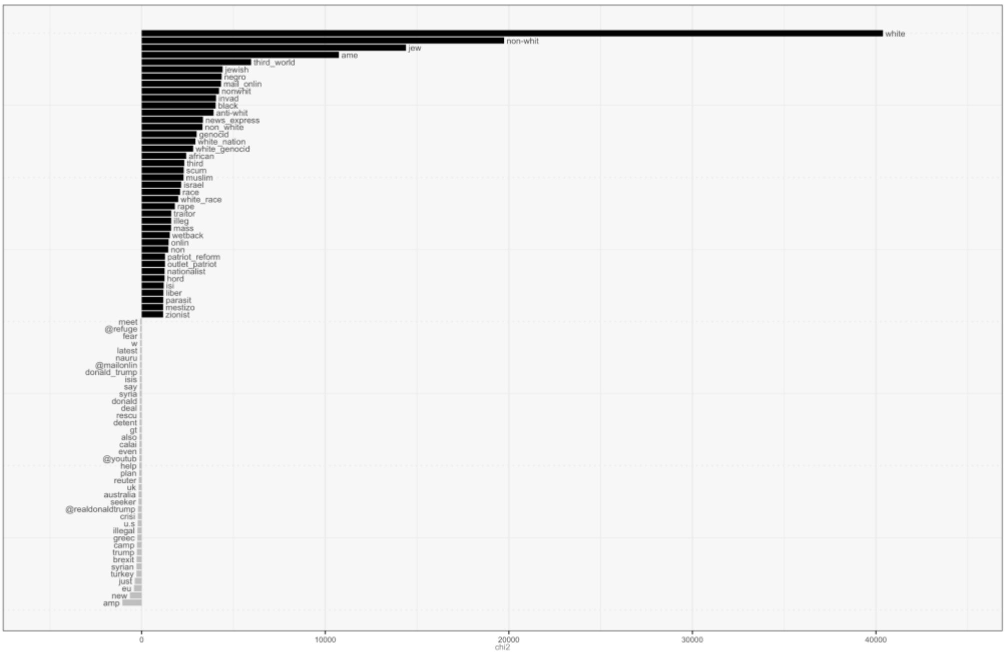
Keyness Analysis 2014 [Target: Toxic posts on Stormfront, Reference: Twitter]



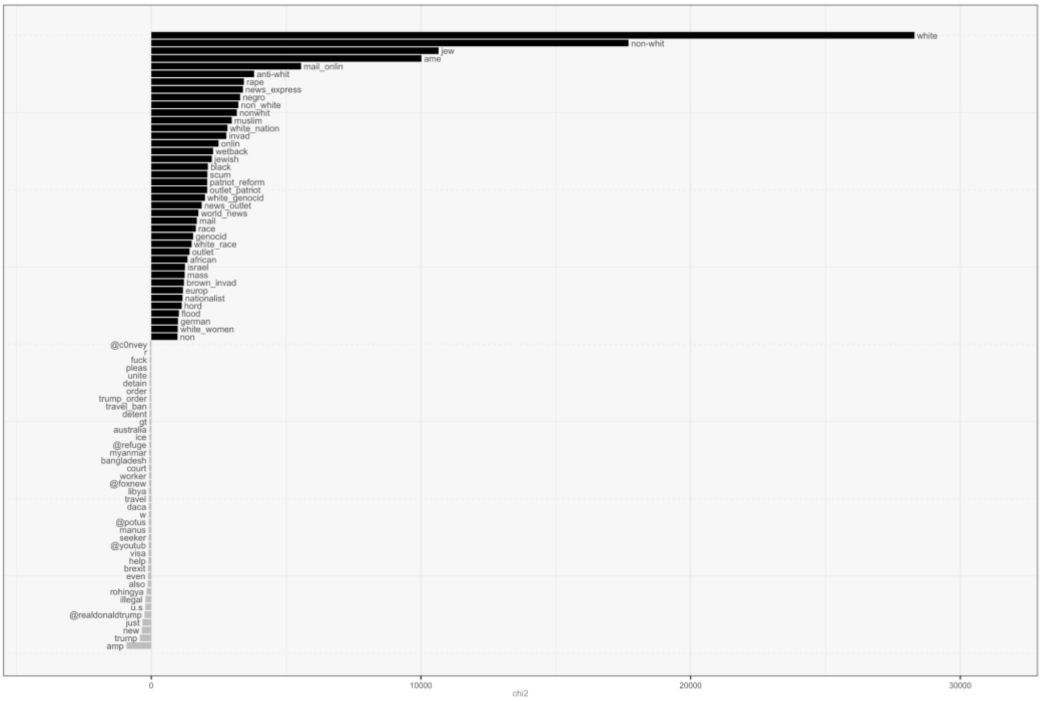
Keyness Analysis 2015 [Target: Toxic posts on Stormfront, Reference: Twitter]



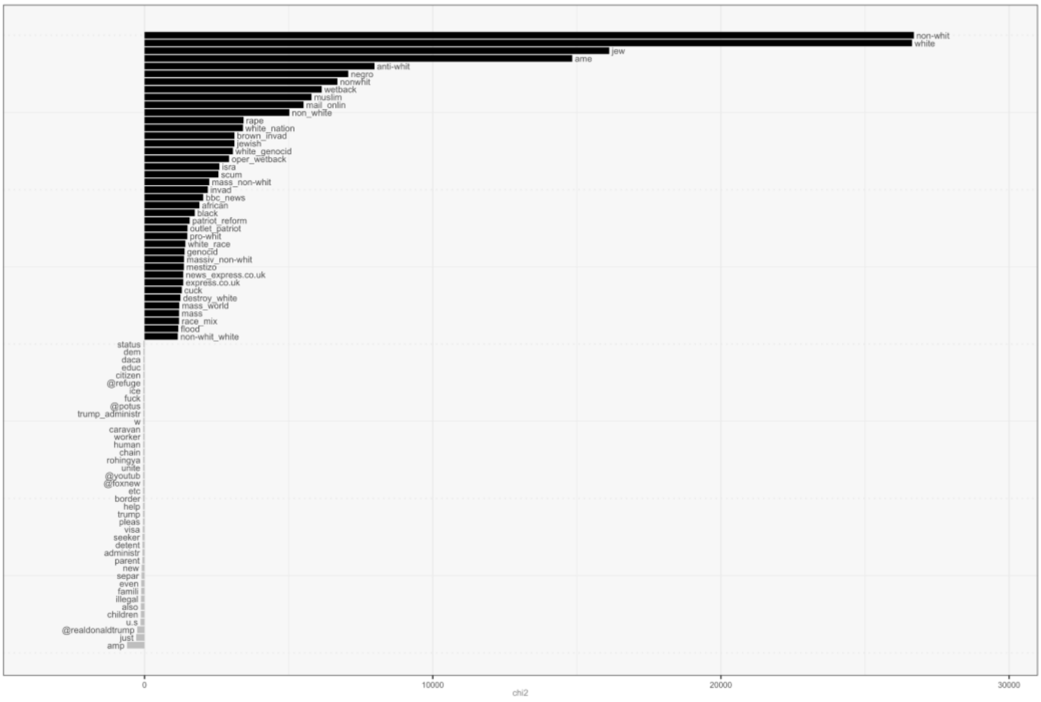
Keyness Analysis 2016 [Target: Toxic posts on Stormfront, Reference: Twitter]



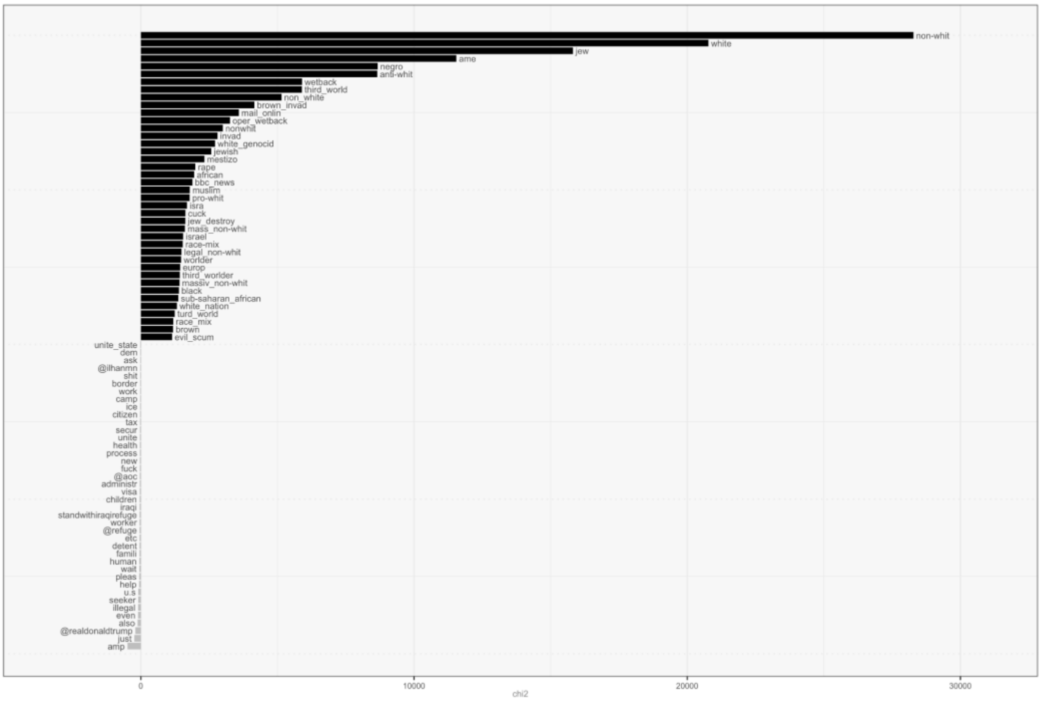
Keyness Analysis 2017 [Target: Toxic posts on Stormfront, Reference: Twitter]



Keyness Analysis 2018 [Target: Toxic posts on Stormfront, Reference: Twitter]

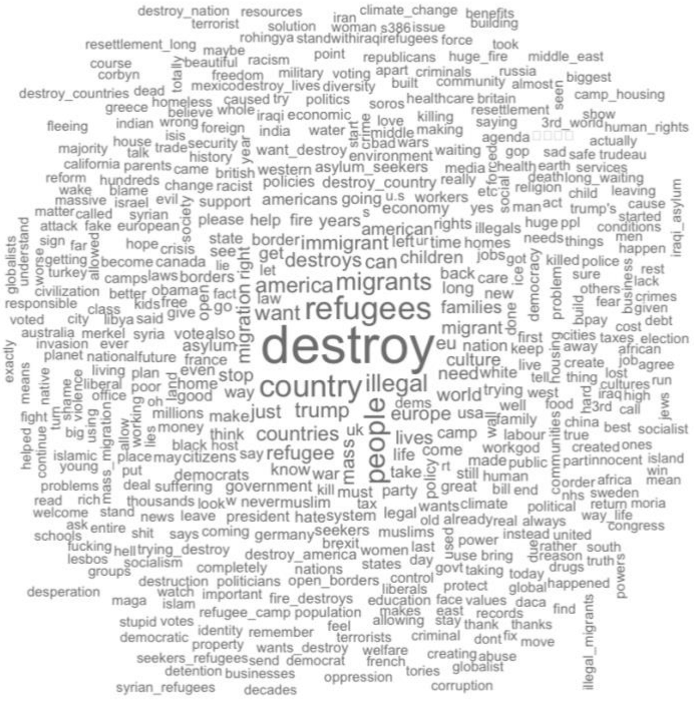


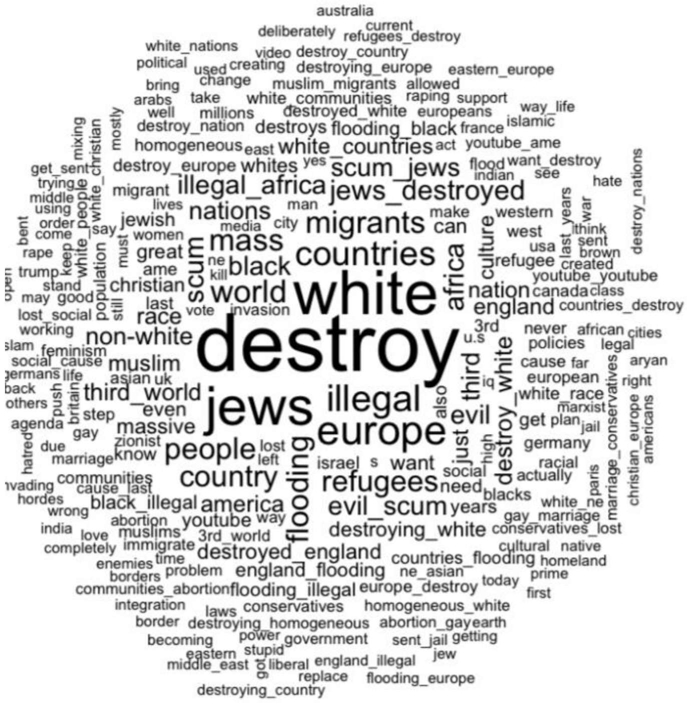
Keyness Analysis 2019 [Target: Toxic posts on Stormfront, Reference: Twitter]



Appendix L - Supplementary qualitative analysis comparing the use of key terms and their contextual use across Stormfront and Twitter.

"Destroy" word clouds of the most co-occurring words on Twitter, left (N = 6991),
Stormfront, right (N = 263)





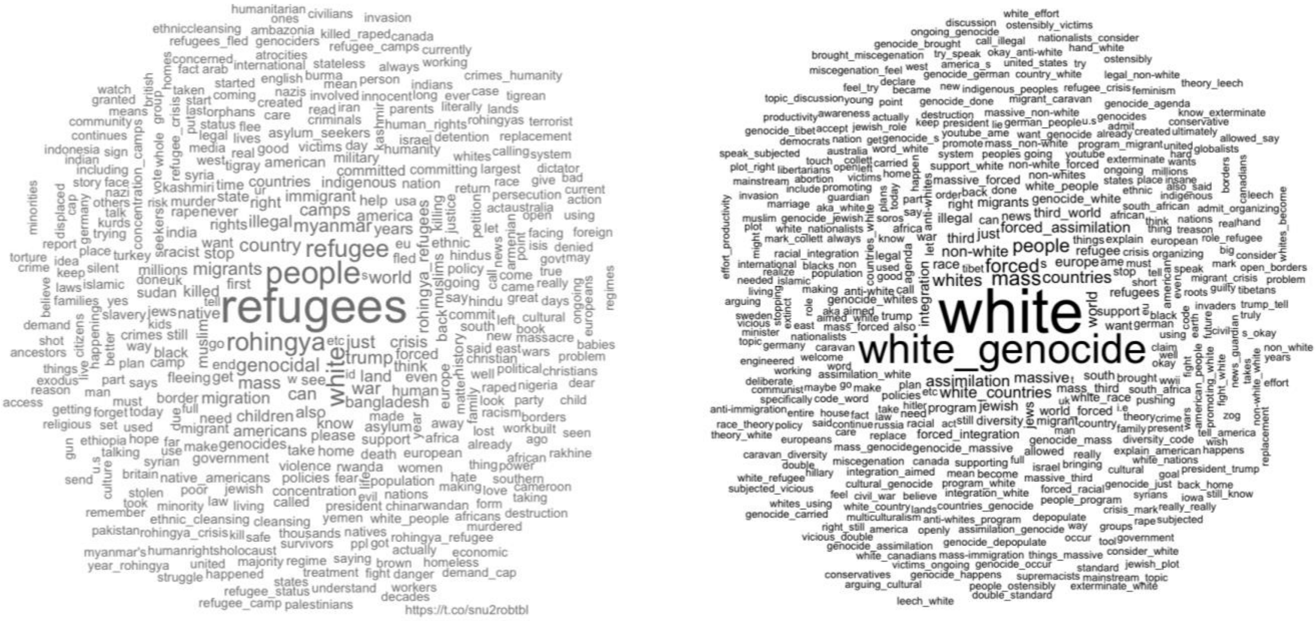
Twitter - examples
"In 1965 immigration law was changed to **destroy** the white population. Only explanation when you have 3 whites for every 15 non whites." "Globalist policy. look around the Western world. Open borders, illegal migrant voting. They'll vote for more open borders, more free stuff. It's aim? To **destroy** the West, to bring in a new order. One we will not recognise, nor enjoy. This is about absolute control over us."

Stormfront - examples

“nothing jews love more than to **destroy** white nations through non-white immigration.”

“yes, globalism, unlimited foreign immigration, and multiracial integration as a means to **destroy** the white working class.”

“Genocide” word clouds of the most co-occurring words on Twitter, left (N = 2171), Stormfront, right (N = 603)



Twitter - examples

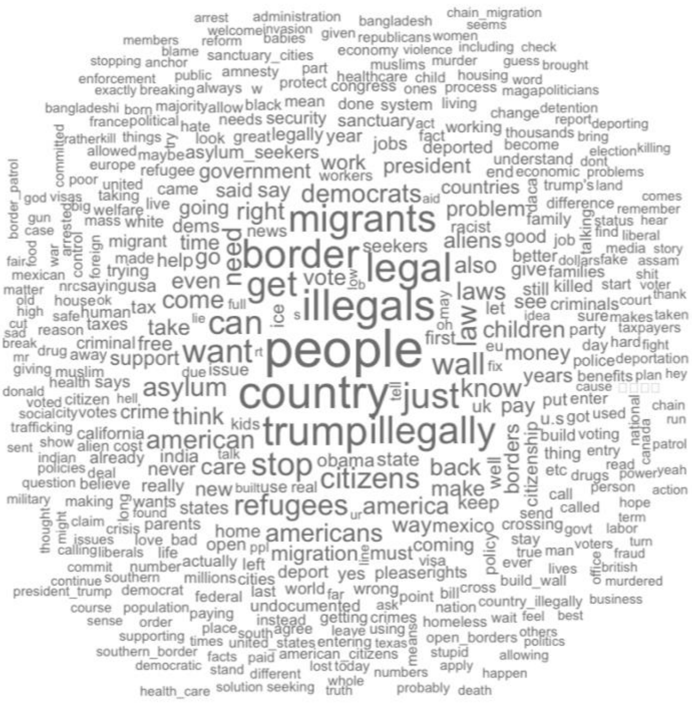
“Anti-Whites support WHITE **GENOCIDE** via mass immigration &amp; forced assimilation in ALL ONLY White countries”

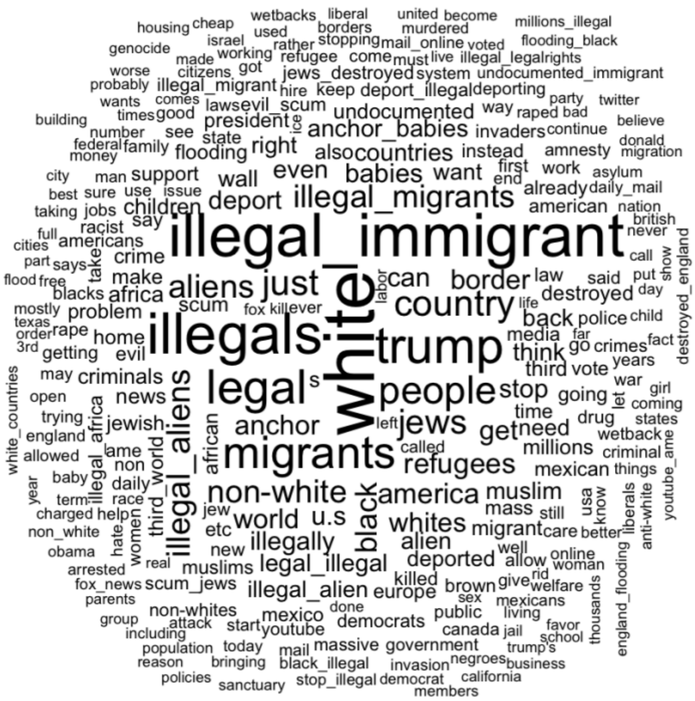
“WE DO NOT GET STRENGTH FROM DIVERSITY! Am I the only person in Britain that is concerned about the **genocide** of the indigenous English speaking people. I created this petition to demand a cap to immigration. Please sign and retweet petition. https://t.co/SNU2roBtBL”

“No it is not! We can deport them! Accepting mass immigration accepts our own **genocide** as defined by UN, also complicit. We can legally deport millions based on our constitution, illegals, criminals, terrorists, change laws to stop our cultural appropriation and welfare state!”

Stormfront - examples
“stick to things like, "stop immigration, start repatriation", or "hitler was right!", or "diversity = white **genocide**" etc." "there was no obama, no queer 'marriage', no war on police officers, no mass immigration and appeasement of mooslims, and liberals were still at least attempting to mask their designs on the **genocide** of white people."

“Illegal” word clouds of the most co-occurring words on Twitter, left (N = 100488), Stormfront, right (N = 1258)





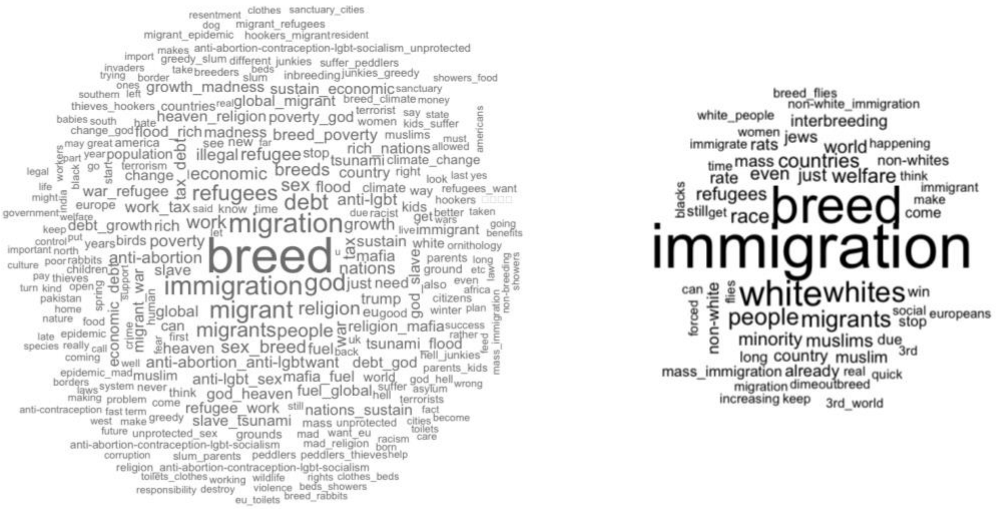
Twitter - examples
"There's no one to fix this problem of massive **illegal** migration into this country, than us the masses. We need to take the battle to the streets coz there's no willing from the gov or politicians. Let's fix this problem ourselves" "I believe if you start your life in America by breaking the law the likelyhood of you committing a high offense crime is very strong and we see this as **illegal** aliens are 3x more likely to break the law than a legal resident BorderCrisis BuildTheWall https://t.co/u83aBYMpbk"

Stormfront - examples

"he was alwats pro jewish, pro negro, pro non white immigration (so far as it's not **illegal** immigration), pro homosexuality."

"the u.s. court system is falling apart everything from the lance scarsella case to a negro female judge in kansas ruling for **illegal** immigrants to vote."

"Breed" word clouds of the most co-occurring words on Twitter, left (N = 964)
Stormfront, right (N = 66)



Twitter - examples

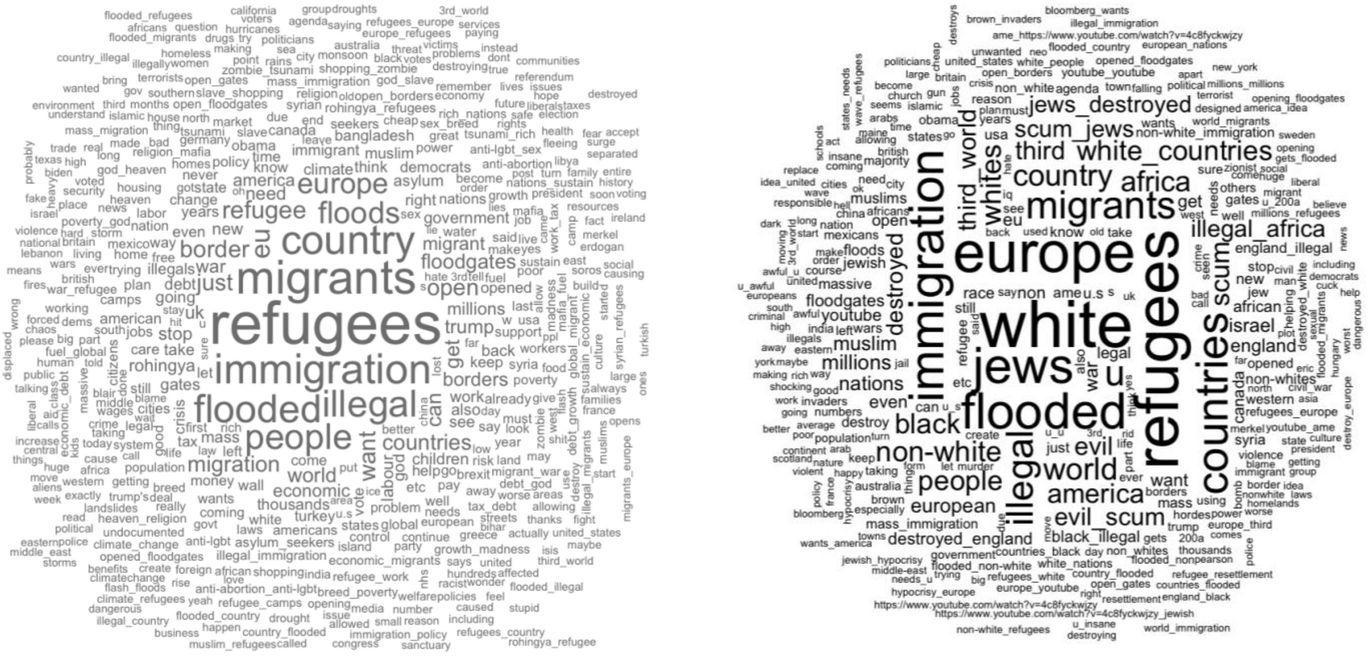
"Soft stance on immigration doesn't work. They are quite honest that their objective is to out **breed** us."

"You want to know another good way to destroy America? Have legal immigration from third world countries. People who **breed** out of control, bring in their 57 relatives through chain migration, and turn America into a hell hole. That's another good way."

"They are just handing their country over to the UN and it's mass migration. Why do they have to take migrants. Someone needs vasectomies and ovary removals. If they can't take care of their own people stop **breeding**."

Stormfront - examples
"Even if we stop all immigration, the mexicans (and muslims even) already have a **breeding** population here of some 50 million (or whatever the number); and since they collectively breed like rats, you're necessarily facing a scenario and policy of expulsion." "Third world migration is the worst immigration not just for damage to the national character but also because these people **breed** like flies unlike europeans."

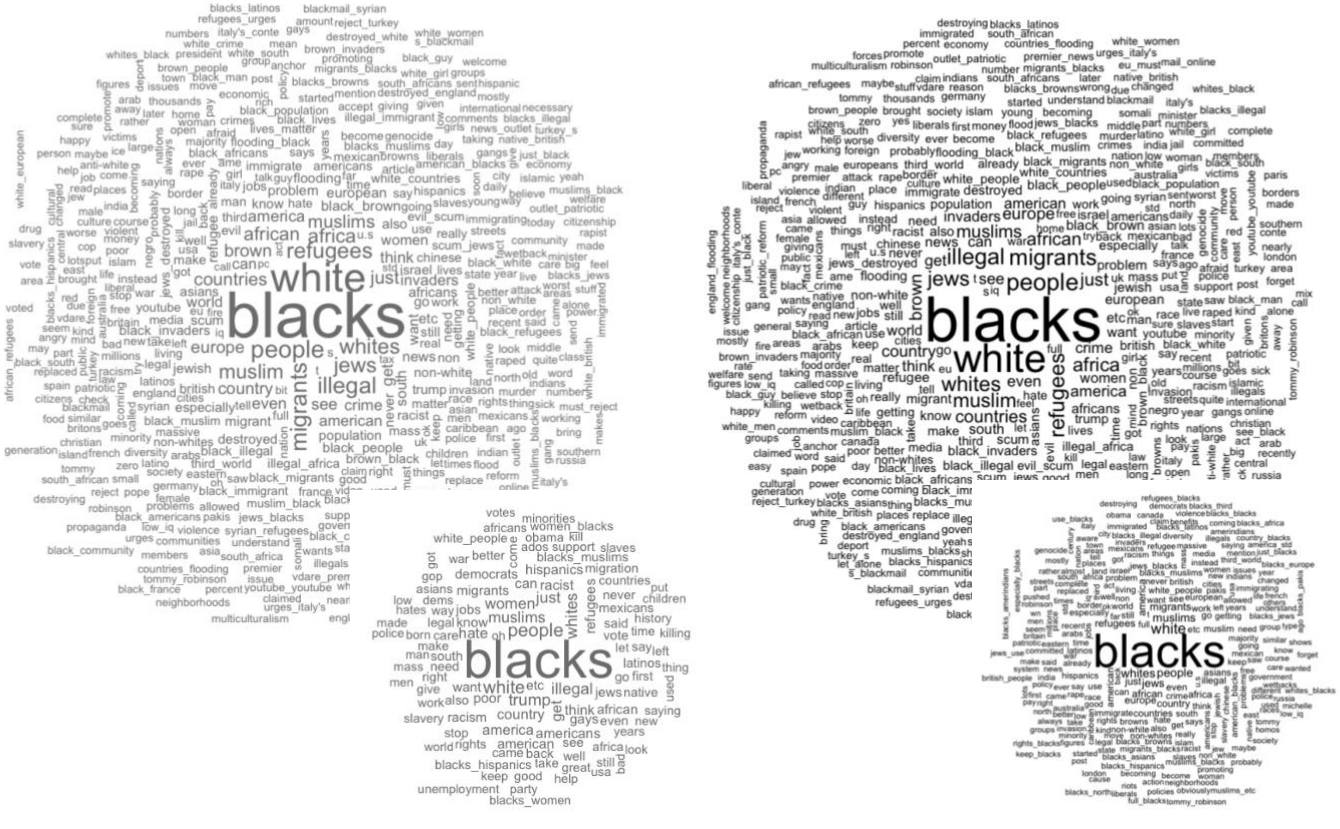
"Flood" word cloud of the most co-occurring words on Twitter, left (N = 4422), Stormfront, right (N = 287)



Twitter - examples
"Why do the EU even require an army, who are they fighting? The biggest threat to the Western World is not Russia, it's Islam. Yet, the EU seem to think it's fine to open the **floodgates** to millions of Muslim migrants and refugees. Why form an army, when the enemy is within?"
"80,000 Somali refugees **flooding** into MN and reproducing like rabbits. That's how."
"@realDonaldTrump Mr.President: Save us from Muslim Migration **flood**, caused by Merkels negligence. She rejected all warnings of ministers."

Stormfront - examples
"i believe the the liberal left knew the plan they had for the west was not moving quick enough so they **flood** our nations with anti western immigrants hoping they'll speed up our demise."
"what did brussels expect **flooding** the migrant scum in, peace and love?"
"though i can't imagine white people in north dakota wanting a **flood** of refugees coming to town."

"Black" word clouds of the most co-occurring words on Twitter (top left , N = 11191) Stormfront (top right, N = 603). "Blacks" added as both domains had the term as its most recurring one (Twitter, N = 1514; Stormfront, N = 253).



Twitter - examples
"Thank you assholes for showing your face now we know what a wetback looks like without it's sheets it shows why you guys have a 5th grade education since your so smart answer this? True or false do all you guys have **black** daddy's in Kentucky!!" "Beto saw you at rally tell a **black** dude it was alright to kneel. Did you kneel for your daddy when you had a hit n run. You suck and invaders suck too. No green card for welfare. Please stay home. Beto ur a dummy."

Stormfront - examples
"you also need to propagandize your family and close friends, be unrelenting, send them videos about **black** on white crime and the endless invasion of immigrants, make them afraid, make them angry, and keep doing it until they get the message." "it's time that high profile 'tommy robinson' and ukip openly admitted islam is not the only threat to britain, and that third world immigration in general and spiralling **black** crime is destroying once proud white communities."

ISSN: 1474-1938/1946