Challenges and frontiers in abusive content detection

Bertie Vidgen
The Alan Turing Institute,
London, United Kingdom
bvidgen@turing.ac.uk

Dong Nguyen
The Alan Turing Institute,
London, United Kingdom
dnguyen@turing.ac.uk

Rebekah Tromble
The Alan Turing Institute,
London, United Kingdom
rtromble@turing.ac.uk

Alex Harris
The Alan Turing Institute,
London, United Kingdom
aharris@turing.ac.uk

Scott Hale
The Alan Turing Institute,
London, United Kingdom
shale@turing.ac.uk

Helen Margetts
The Alan Turing Institute,
London, United Kingdom
hmargetts@turing.ac.uk

Abstract

Online abusive content detection is an inherently difficult task. It has received considerable attention from academia, particularly within the computational linguistics community, and performance appears to have improved as the field has matured. However, considerable challenges and unaddressed frontiers remain, spanning technical, social and ethical dimensions. These issues constrain the performance, efficiency and generalizability of abusive content detection systems. In this article we delineate and clarify the main challenges and frontiers in the field, critically evaluate their implications and discuss solutions. We also highlight ways in which social scientific insights can advance research.

1 Introduction

Developing robust systems to detect abuse is a crucial part of online content moderation and plays a fundamental role in creating an open, safe and accessible Internet. It is of growing interest to both host platforms and regulators, in light of recent public pressure (HM Government, 2019). Detection systems are also important for social scientific analyses, such as understanding the temporal and geographic dynamics of abuse.

Advances in machine learning and NLP have led to marked improvements in abusive content detection systems' performance (Fortuna & Nunes, 2018; Schmidt & Wiegand, 2017). For instance, in 2018 Pitsilis et al. trained a classification system on Waseem and Hovy's 16,000 tweet dataset and achieved an F-Score of 0.932, compared against Waseem and Hovy's original 0.739; a 20-point increase (Pitsilis, Ramampiaro, & Langseth, 2018; Waseem & Hovy, 2016). Key innovations include the use of deep learning and ensemble architectures, using contextual word embeddings, applying dependency parsing, and the inclusion of user-level variables within models (Badjatiya, Gupta, Gupta, & Varma, 2017; Zhang et al., 2018). Researchers have also addressed numerous tasks beyond binary abusive content classification, including identifying the target of abuse and its strength as well as automatically moderating content (Burnap & Williams, 2016; Davidson, Warmsley, Macy, & Weber, 2017; Santos, Melnyk, & Padhi, 2018). However, considerable challenges and unaddressed frontiers remain, spanning technical, social and ethical dimensions. These issues constrain abusive content detection research, limiting its impact on the development of real-world detection systems.

We offer critical insights into the challenges and frontiers facing the use of computational methods to detect abusive content. We differ from most previous research by taking an interdisciplinary approach, routed in both the computational and social sciences. Broadly, we advocate that social science should be used in a complementary way to advance research in this field. We also highlight the lack of support given to researchers and provide guidelines for working with abusive content.

The paper is structured as follows. First, we outline three reasons why, from a research perspective, abusive content detection poses such a challenge (Section 2). Second, we identify challenges facing

Proceedings of the Third Workshop on Abusive Language Online, pages 80–93
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

the abusive content detection research community (Section 3). Third, we identify research frontiers; un- and under- addressed areas which would benefit from further investigation (Section 4).

2 Research Challenges

2.1 Categorizing abusive content

The categorization of abusive content refers to the criteria, and process, by which content is identified as abusive and, secondly, what *type* of abusive content it is identified as. This is a social and theoretical task: there is no objectively 'correct' definition or single set of pre-established criteria which can be applied. The determination of whether something is abusive is also irreducible to legal definitions as these are usually minimalistic (HM Government, 2019). Similarly, using the host platforms' guidelines is often inappropriate as they are typically reactive and vague. More generally, academia should not just accept how platforms frame and define issues as this might be influenced by their commercial interests.

Clarity in sub-tasks. Detecting abusive content generically is an important aspiration for the field. However, it is very difficult because abusive content is so varied. Research which purports to address the generic task of detecting abuse is typically actually addressing something much more specific. This can often be discerned from the datasets, which may contain systematic biases towards certain types and targets of abuse. For instance, the dataset by Davidson et al. is used widely for tasks described generically as abusive content detection yet it is highly skewed towards racism and sexism (Davidson et al., 2017).

Sartori's work in political science on the 'ladder of abstraction' can be used to understand this issue (Sartori, 1970). He argues that all concepts can be defined and described with varying degrees of abstraction. For instance, 'democracy' can be defined very broadly in relation to how 'the people' is represented or very narrowly as a set of specific institutions and procedures. The degree of abstraction should be chosen by considering the goals and nature of the research – otherwise we risk 'swim[ming] in a sea of empirical and theoretical messiness.' (Sartori, 1970, p. 1053)

Abusive content detection research is currently marked by too much of what Sartori labels 'high' and 'low' level abstraction. Some researchers use highly abstract terms to describe tasks, such as detection of 'abuse' or 'flagged' content. These terms are not very informative, and it is difficult to know exactly what sub-task is being addressed. For instance, flagged content may be abusive but is likely to also include other forms of non-abusive (albeit prohibited) content. On the other side, some research uses very narrow terms which are at an overly 'low' level of abstraction. For instance, 'hate' denotes a specific aggressive and emotional behavior, excluding other varieties of abuse, such as dismissal, insult, mistrust and belittling.

Addressing an appropriate level of abstraction is important for creating useable detection systems. It requires that subtasks are clearly disambiguated and labelled. This is a much-discussed but still unresolved problem in existing research (Waseem, Davidson, Warmsley, & Weber, 2017). Waseem et al. suggest that one of the main differences between subtasks is whether content is 'directed towards a specific entity or is directed towards a generalized group' (Waseem et al., 2017). This distinction has been widely adopted (Zampieri et al., 2019). We propose that subtasks are further disambiguated into three types of directed abuse:

Individual-directed abuse. Abuse directed against an individual. This may involve tagging the individual (e.g. '@Username you are a f*cking id*ot) or just referring to them (e.g. 'I think Tom W. is a tw*t') These two varieties can be called 'tagged individual-directed' and 'referenced individual-directed' respectively. Most research in this area falls under cyberbullying (Sugandhi, Pande, Chawla, Agrawal, & Bhagat, 2016) although there are notable exceptions (Wulczyn, Thain, & Dixon, 2017).

Identity-directed abuse. Abuse directed against an identity, such as a social group, demographic or affiliation (e.g. 'I hate Conservatives' or 'Ban Muslims') (Silva, Mondal, Correa, Benevenuto, & Weber, 2016). This can be hard to separate from individual-directed abuse as, in some cases, individuals receive abuse *because* of their identity. This might be reasonably obvious (e.g. 'You stupid b*tch', indicating misogyny) but in other cases it is hard to discern as the content alone does not reveal prejudice. Establishing when abuse is truly individual-directed compared with identity-directed needs to be investigated further, especially given evidence that some identities receive more individual-directed abuse (Gorrell, Greenwood, Roberts, Maynard, & Bontcheva, 2018).

Concept-directed abuse: abuse which is directed against a concept or entity, such as a belief system, country or ideology, e.g. 'Capitalism sucks *ss.'. Concept-directed abuse may not be considered a form of abuse in all cases as it can be very similar to simply expressing criticism of something. We include it here because there are deep links between hatred of a concept and hatred of those who embody that concept. For instance, there are cross-overs between anti-Islamic and anti-Muslim abuse, whereby abuse of the concept (Islam) is used as a proxy for abusing the associated identity (Muslims) (Allen, 2011). At the same time, we caution against automatically moderating concept-directed abuse as this could have concerning implications for freedom of expression.

This typology can be integrated with other dimensions of abuse to create additional subtasks. One consideration is who the system detects abuse for; that is, who actually receives abuse (Salminen et al., 2018). Within identity-directed abuse, this can be separated into different identities and affiliations (e.g. Muslims or the Republican party). Within individual-directed abuse, this includes different roles (such as content producers vs. moderators) and social relations (such as friends vs. strangers). Either one or several recipients of abuse can be studied within any model. Specifying the recipient not only makes tasks tractable, but also helps build social scientific and policy-relevant knowledge.

A further consideration is how abuse is articulated, which can include hatefulness, aggression, insults, derogation, untruths, stereotypes, accusations and undermining comments. Detecting different articulations of abuse within a single system involves multi-label or multi-class modelling and can be computationally difficult. However, it also leads to more nuanced outcomes. A key distinction is whether abuse is explicit or implicit (Waseem et al., 2017; Zampieri et al., 2019). Other articulations of abuse can also be addressed. For instance, Anzovino et al. develop a system that not only detects misogyny but also whether it consists of stereotypes, discrediting, objectification, harassment, dominance, derailing or threats of violence (Anzovino, Fersini, & Rosso, 2018).

Drawing these points together, we propose that researchers consider at least three dimensions of abusive content. They can be incorporated in various ways to produce different tasks.

1. What the abuse is directed against
2. Who receives the abuse
3. How the abuse is articulated

Clarity in terminology. Clarifying terminology will help delineate the scope and goals of research and enable better communication and collaboration. Some of the main problems are (1) researchers use terms which are not well-defined, (2) different concepts and terms are used across the field for similar work, and (3) the terms which are used are theoretically problematic. Specifically, three aspects of existing terminology have considerable social scientific limitations.

The intention of the speaker. Abusive content is often defined in reference to, and focuses on, the speakers' intentions. In particular, it is central in the notion of 'hate', which suggests a specific orientation of the speaker. For instance, Pitsilis et al. describe hate as 'all published text that is used to express hatred towards some particular group with the intention to humiliate its members' (Pitsilis et al., 2018). Elsewhere, Kumar et al. distinguish 'overt' from 'covert' hate (Kumar, Ojha, Malmasi, & Zampieri, 2018). The implication of 'covert' is that speakers are behaving surreptitiously to hide their abusive intentions. However, the intention of speakers is difficult to discern using socially-generated data and may not directly correspond with their actions (Crawford & Gillespie, 2016; Margetts, John, Hale, & Yasseri, 2015). The way in which meaning is 'encoded' in online contexts cannot be easily ascertained, particularly given the anonymity of many users and the role of 'context collapse'. (Marwick & boyd, 2010). The true audience which speakers address may be different from the ones that they imagine they are addressing (Ibid.). As such, little should be assumed about speakers' intentions, and it can be considered an unsuitable basis for definitions of abuse.

The effect of abuse. Many definitions pre-empt the effects of abusive language. For instance, Lee et al. describe abusive language as 'any type of insult, vulgarity, or profanity that debases the target; it also can be anything that causes aggravation' (Lee, Yoon, & Jung, 2018). Similarly, in Wulczyn et al.'s dataset of over 100,000 Wikipedia comments, 'toxicity' is defined in relation to how likely it is to make individuals leave a discussion (Wulczyn et al., 2017). These definitions are not only very subjective but they also risk conflating distinct types of abuse: first, content which expresses abuse and, second, content which has an abusive effect. These two aspects often coincide, but not always, as shown in sociological studies of prejudice. In relation to Islamophobia, Allen distinguishes between 'Islamophobia-as-process' and 'Islamophobia-as-product', whereby the first refers to actions which can be considered Islamophobic (intrinsically) and the second to outcomes which can be considered Islamophobic (extrinsically) (Allen, 2011). This distinction should also be used to understand abusive content: language which does not express an inherently abusive viewpoint but is experienced as abusive is very different and, as such, should be addressed separately, to language which is intrinsically abusive.

The sensibilities of the audience. Online audiences are hugely varied and attempts to discern their sensibilities are fundamentally flawed: inevitably, some proportion of the audience will be mischaracterized. This is reflected by research into inter-annotator agreement, whereby annotators often vary considerably in what they consider to be hateful or abusive, even with training and guidance (Salminen & Almerekhi, 2018). Binns et al. show that male and female annotators have different perceptions of what is considered toxic (Binns, Veale, Van Kleek, & Shadbolt, 2017). Assumptions about the sensibilities of the audience are entailed by the widely-used term 'offensiveness' (Davidson et al., 2017), which is intrinsically subject-oriented: it begs the question, *offensive for whom?* What is considered offensive by one audience, or in one context, might not be offensive elsewhere. As such, we advocate avoiding definitions of abuse which make strong assumptions about the audience without in-depth empirical analysis.

2.2 Recognizing abusive content

We identify five linguistic difficulties which increase the challenge of detecting abusive content. They have all been associated with classification errors in previous work. However, they are not always discussed and handled systematically, and their impact is hard to assess as they are often discussed qualitatively rather than measured.

Humor, irony and sarcasm. Supposedly humorous, ironic or sarcastic abusive content is often viewed as a source of classification error (Nobata, Thomas, Mehdad, Chang, & Tetreault, 2016; van Aken, Risch, Krestel, & Löser, 2018). However, drawing on critical studies of prejudice and hate, we propose that such content is still abusive (Weaver, 2010). There are three reason for this. First, these rhetorical devices have been shown to serve as ways of hiding, spreading and legitimating genuine abuse (Ji Hoon Park, Gabbadon, & Chernin, 2006). Second, individuals who view such content may be unaware of who the author is and the broader context, and as such not recognize that it is humorous – as discussed above, intentions are hard to discern online. A supposedly ironic comment which is intended to lampoon abuse may be indistinguishable from genuine abuse (LaMarre, Landreville, & Beam, 2009). Third, purportedly ironic, satirical and humorous abusive content usually relies on a kernel of prejudice: the lynchpin of the rhetorical strategy is that the audience recognizes, and perhaps implicitly accepts, the negative tropes and ideas associated with the targeted group (Ma, 2014). Thus, whilst humor, irony and sarcasm are often seen as being non-abusive, we recommend that they are re-evaluated.

Spelling variations. Spelling variations are ubiquitous, especially in social media (Eisenstein, 2013). Examples of spelling variation include the elongation of words (e.g., 'oh' to 'ohh') and use of alternatives (e.g., 'kewl' instead of 'cool'). Spelling variation is often socially significant, reflecting expressions of identity and culture (Sabba, 2009). At the same time, some variations reflect semantically near-identical content (e.g. 'whaaaaa?' and 'whaaa?'). Spelling variations are also sometimes used adversarially to obfuscate and avoid detection (e.g. by using unusual punctuation or additional spaces) (Eger et al., 2019; Gröndahl, Pajola, Juuti, Conti, & Asokan, 2018). In most contexts, it is hard to identify why spelling varies.

Spelling variations increase the likelihood of errors as they create many 'out of vocabulary' terms which have to be handled (Serrà et al., 2017). Text normalization has been proposed as a solution, however this risks losing meaningful social information (Eisenstein, 2013). Using larger and more diverse datasets will only partly mitigate this problem as no dataset will ever account for all variations, and language use changes over time. A more promising way of addressing this is to model language at the character or subword level (Devlin, Chang, Lee, & Toutanova, 2018; Mehdad & Tetreault, 2016). More empirical research into why particular spelling variations occur would also be useful.

Polysemy. This is when a word with a single spelling has multiple meanings. Which meaning is elicited depends on the context. Magu and Luo describe how 'euphemistic' code words, such as 'Skype' or 'Bing', are used to derogate particular groups (Magu, Joshi, & Luo, 2017). Similarly, Palmer et al. describe how adjectival nominalization (e.g. changing 'Mexicans' to 'the Mexicans') can transform otherwise neutral terms into derogations (Palmer, Robinson, & Phillips, 2017). Polysemy is a particular challenge with abusive content as many users avoid obvious and overt forms of hate (which are likely to be automatically removed by platforms) and instead express hate more subtly (Daniels, 2013). Word representations which explicitly take into account context are one way of overcoming this issue (Devlin et al., 2018).

Long range dependencies. Much existing research is focused on short posts, such as Tweets (Schmidt & Wiegand, 2017). However, socially generated content can cross over multiple sentences and paragraphs. Abuse may also only be captured through conversational dynamics, such as multi-user threads (Raisi & Huang, 2016). This has been well-addressed within studies of cyberbullying, but is also highly relevant for the field of abusive content detection more widely (Van Hee et al., 2018). Creation of more varied datasets will help to address this problem, such as using data taken from Reddit or Wikipedia.

Language change. The syntax, grammar, and lexicons of language change over time, often in unexpected and uneven ways. This is particularly true with informal forms of 'everyday' language, which proliferate in most online spaces (Eisenstein, O'Connor, Smith, & Xing, 2014). One implication is that the performance of systems trained on older datasets degrades over time as they cannot account for new linguistic traits. Using multiple temporally separated datasets to evaluate systems will help to address this, as well as further research into the impact of time on language.

2.3 Accounting for context

Meaning is inherently ambiguous, depending upon the subjective outlook of both speaker and audience, as well as the specific situation and power dynamics (Benesch, 2012). These factors have long been given insufficient attention in the study of online abuse, which has mostly focused on just the content alone. This has clear limitations. For instance, in most cases, the term "N***a" has an almost opposite meaning if uttered by a white compared to a black person.

Some recent work has started to explicitly account for context by including user-level variables in classification systems. Unsvåg and Gambäck evaluate a system on three datasets and find that, compared with a baseline using logistic regression with n-grams, inclusion of individual-level features, such as gender, social network, profile metadata and geolocation, improves performance (Unsvåg & Gambäck, 2018). Other studies report similar results, using both local and global social network features, noticeably through incorporating the node2vec algorithm (Papegnies, Labatut, Dufour, & Linarès, 2017; Raisi & Huang, 2017). The use of network representations is supported by social science research which shows evidence of homophily online; it is likely that abusive users are connected to other abusive users (Caiani & Wagemann, 2009; Tien, Eisenberg, Cherng, & Porter, 2019). We propose that anonymity should also be explicitly modelled in future work as it has disinhibiting effects (Amichai-hamburger & McKenna, 2006) and is empirically associated with users posting abuse (Hine et al., 2017).

The inclusion of user-level features helps to drive improvements in classification performance and should be welcomed as an important step towards more nuanced and contextually-aware models. That said, we offer four warnings. First, it may make temporal or network analysis difficult as the classification of users' content is based on these features, creating clear risk of confounding. Second, it may lead to new types of unfairness and bias whereby the content of certain network topologies or certain nodes are more likely to be detected as hateful – which may, in turn, be related to meaningful social characteristics, such as gender or age. Third, these systems are largely trained on a snapshot of data and do not explicitly take into account temporality. It is unclear how much data is required for them to be trained. Fourth, models may be biased by the training data. Wiegand et al. show that if most abusive content in a dataset comes from only a few users then including user-level information risks overfitting: the classifier just picks up on those authors' linguistic traits (Wiegand, Ruppenhofer, & Kleinbauer, 2019).

Context goes beyond just the identity of the speaker. It also includes the social environment in which they operate, which in most cases comprises both the platform and the specific group or community, such as the subreddit or Facebook page. Existing research can be leveraged to address this: Qian et al. report a model which identifies the origins of posts from 40 far right hate groups on Twitter (Qian, ElSherief, Belding, & Wang, 2018) Chandrasekharan et al similarly build a model that identifies whether content is from 9 different communities on niche social media platforms (Chandrasekharan, Samory, Srinivasan, & Gilbert, 2017). This is promising research which should be integrated into the detection of online abuse, thereby accounting explicitly for the social environment in which content is shared. To more fully address the role of context we also need more empirical analysis of which aspects most greatly impact perceptions of abuse.

3 Community Challenges

Abusive content detection is a relatively new field of study; in only 2016, Waseem and Hovy wrote 'NLP research on hate speech has been very limited' (Waseem & Hovy, 2016). Since then it has expanded propitiously. Noticeably, a recent shared task had over 800 teams enter of which 115 reported results (Zampieri et al., 2019). The creation of a research community is fundamental for advancing knowledge by enabling collaboration and resource sharing. However, the abusive content detection community currently faces several challenges which potentially constrain the development of new and more efficient methods.

3.1 Creating and sharing datasets

Creating appropriate datasets for training hate detection systems is a crucial but time-consuming task (Golbeck et al., 2017). Currently available datasets have several limitations.

Degradation. With many datasets, including those from Twitter, content cannot be shared directly but, instead, IDs are shared and the dataset recreated each time. This can lead to considerable degradations in the quality of datasets over time. For instance, Founta et al. shared a dataset of 80,000 tweets but soon after this was reduced to 70,000 (Founta et al., 2018; Lee et al., 2018). This not only decreases the quantity of data, reducing variety, but also the class distribution changes. This makes it difficult to compare performance of different models on even one dataset. To address this issue, we encourage more collaborations with online platforms to make datasets available. A successful example of this is Twitter's release of the IRA disinformation dataset (Twitter, 2018).

Annotation. Annotation is a notoriously difficult task, reflected in the low levels of inter-annotator agreement reported by most publications, particularly on more complex multi-class tasks (Sanguinetti, Poletto, Bosco, Patti, & Stranisci, 2018). Noticeably, van Aken suggests that Davidson et al.'s widely used hate and offensive language dataset has up to 10% of its data mislabeled (van Aken et al., 2018). Few publications provide details of their annotation process or annotation guidelines. Providing such information is the norm in social scientific research and is viewed as an integral part of verifying others' findings and robustness (Bucy & Holbert, 2013). In line with the recommendations of Sabou et al., we advocate that annotation guidelines and processes are shared where possible (Sabou, Bontcheva, Derczynski, & Scharl, 2014) and that the field also works to develop best practices.

Dataset variety. The quality, size and class balance of datasets varies considerably. Understanding the decisions behind dataset creation is crucial for identifying the biases and limitations of systems trained on them. When creating datasets, researchers need to weigh up ensuring there are sufficient instances of abuse (by biased sampling through e.g. using abusive keywords) with making sure the variety of non-abusive content is great enough for the system to be applied in 'the wild' and avoid overfitting. Wiegand et al. measure the impact of biased sampling on several widely used datasets (Wiegand et al., 2019). They find it can lead to confounding whereby non-abusive terms serve as signals for identifying abuse as they are highly correlated – but such signals are unlikely to exist in the real world. To enable greater research transparency, sampling methods should always be reported in accessible dataset documentation.

At present, the main goal of biased sampling is to increase the incidence of abusive content. We propose that this should be adjusted to focus on dataset *variety*. Datasets could be curated to include linguistically difficult instances, as well as 'edge cases': content which is non-abusive but very similar to abuse. Three examples are:

1. *Non-abusive profanities*. Most detection systems use the existence of profanities (also known as 'obscenities') as an input feature. However, profanities are not inherently abusive and can be used to express other emotions.
2. *Abusive reporting*. Content which reports/comments on the abuse of others or aims to challenge/counter such abuse.
3. *Same topic but non-abusive*. Content which is on the same topic as the abusive content but is non-abusive. For instance, if the classification system detects xenophobia, then a suitable edge case is non-abusive content about foreigners.

3.2 Research ethics

The ethics of social scientific and socially-relevant computational research has received considerable scrutiny in recent times (Buchanan, 2017). Most abusive content detection systems are presented as neutral classifiers which merely aim to achieve a well-defined task. However, it is difficult to separate the descriptive from normative aspects of any social system. Academic research can be used to not only monitor and capture social behaviors but also influence and manipulate them (Ruppert, Law, & Savage, 2013). As such, given the sensitivity of this area, ethics should be at the forefront of all research.

Impact on users. Individuals and groups suffer considerably from online abuse, and there is evidence that online abuse is linked with offline attacks (Müller & Schwarz, 2017). Political science research also suggests that any form of extremist behavior, such as online hate, could fuel social antagonisms and even reprisals (Eatwell, 2006). As such, the ethical case for moderating online content is strong. However, at present, research is unevenly distributed, with far more attention paid to abuse in English as well as abuse directed against certain targets, such as racism and sexism rather than anti-Semitism, transphobia or anti-disability prejudice. This is partly due to how research is organized. For example, much research has focused on detecting abuse in Hindi-English – primarily because of a shared competition with a publicly available dataset (Kumar, Reganti, Bhatia, & Maheshwari, 2018). The uneven nature of existing research has unintended harmful consequences as certain targets of abuse receive more focus and as such are better protected. Researchers should aim to diversify the types and targets of abuse which are studied, where possible.

Impact on researchers. Researching online abuse inevitably involves viewing and thinking about abusive content, often for prolonged periods. This can inflict considerable emotional harm on researchers, particularly through vicarious trauma. Social and mental health support is necessary to protect the wellbeing of researchers and to ensure that research is sustainable in the long-term. In our online appendix, we provide a checklist of actions to help reduce the harmful impacts of viewing, annotating and researching abusive content.[1]

Researchers conducting work around sensitive topics are increasingly at risk of receiving online abuse themselves, which can range from spreading false information to 'doxing' (where identifying features, such as a home address, are published online) and 'swatting' (where a false threat is reported to the police). Researchers should not have to compromise on the type of research that

[1] https://github.com/bvidgen/Challenges-and-frontiers-in-abusive-content-detection

they conduct for fear of victimization. The abuse suffered by researchers may also reflect other prejudices, whereby women and minorities are targeted more often. We encourage that best practices are shared between institutions so that individuals can work within the safest and most supportive environments possible. We also recommend that Marwick et al.'s existing guidelines for dealing with harassment are used (Marwick, Blackwell, & Lo, 2016).

4 Research frontiers

4.1 Multimedia content

Most abusive content detection research focuses on text. Little research considers other forms of content, such as images, audio, memes, GIFs, and videos – all of which can be used to spread hate. One noticeable exception is research by Zannettou et al. who create a system for detecting hateful memes by mining hateful Internet forums (Zannettou, Caulfield, Blackburn, & Cristofaro, 2018). The lack of research into non text-based abuse is a severe restriction given the multimedia nature of behavior on social media. It also means that the true recall rate for abusive content detection is potentially orders of magnitude lower than what is reported.

Multimedia content poses both technical and social challenges. Technical challenges relate to the fact that different tools are needed, such as optical character recognition (OCR), image recognition and audio translation. Social challenges relate to the fact that abuse can be expressed in different ways with multimedia. For instance, in Memes, the whole is often more than the sum of its parts: a non-abusive image and non-abusive text can be used which when combined express an abusive message. Figure 1 shows an example of such a meme. It consists of a non-hateful image (Muslims in prayer) and non-hateful text ('Australia, America, England, woken up yet?'). If the image or text were changed (e.g. to a cup of coffee or the phrase 'united in prayer'), then the meme would not be Islamophobic. This kind of abuse only emerges through the text and image combination, and as such is qualitatively different to text which is abusive on its own.

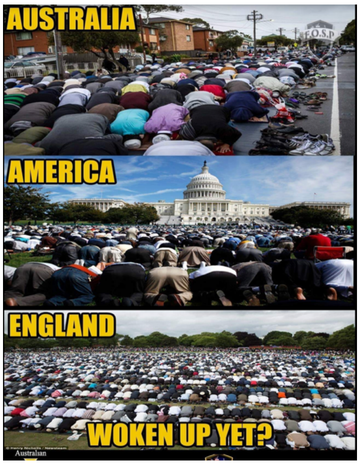

Figure 1, Islamophobic Meme

4.2 Implementation

Fairness. Fairness is a growing concern within abusive content detection. Recent research has shown that systems often perform better for content aimed against certain targets, such as women rather than men (Badjatiya, Gupta, & Varma, 2019; Ji Ho Park, Shin, & Fung, 2018). This feeds into broader research which shows that computational methods can encode and reinforce social biases – even when they are meant to ameliorate them (Garg, Schiebinger, Jurafsky, & Zou, 2017). Metrics have been developed to evaluate bias which enable post-hoc quantification of the extent of these issues (Zhang et al., 2018). However, it would be particularly valuable if detection systems were automatically debiased at the point of creation, for instance by adjusting model parameters given relevant demographic variables, as suggested by Dixon et al. (Dixon, Li, Sorensen, Thain, & Vasserman, 2018). This is important for not only measuring but also removing bias.

A social scientific challenge in this space is that, at present, only biases which are socially 'recognized' can be identified, measured and thus accounted for within models (Fraser, 1997). Potentially, there are other social biases which have not yet received considerable attention but still effect social outcomes and warrant debiasing. For instance, recognition of transphobia has increased considerably over the last ten years, despite previously not being recognized in some parts of society as an important issue (Hines et al., 2018). A related area of bias that needs further investigation is how systems perform at detecting abuse produced by different types of actors, such as those in particular linguistic communities. For instance, systems may have far more false positives when detecting abuse *from* certain types of users, whose content is thus mislabeled and may be incorrectly censored.

Explainability. Closely linked to the notion of fairness is explainability. Abuse detection systems should be explainable to those whose content has been classified and they should avoid becoming 'black boxes'. This is particularly important given the contentious nature of online content moderation and its intersection with issues of censorship, free speech and privacy. One challenge here is that 'explainability' is itself a contested term and what it entails is not well stipulated (Lipton, 2016). Some have also criticized the idea of building secondary post-hoc explanative models as they can be misleading and unreliable. Rudin argues that a better approach is to 'design models that are inherently interpretable' (Rudin, 2018, p. 1). This would also be beneficial from a research perspective, reflecting the scientific process. If we can understand and explain what aspects of a system drive the classifications, then we are more likely to make advances and correct errors. As such, we encourage researchers to develop interpretable models. Nonetheless, given the utility of even hard-to-explain models, such as those using deep learning, post-hoc explanations should also be used where appropriate.

Efficiency. Few publications focus specifically on the challenge of implementing abusive content detection systems at scale and in a timely manner, although there are exceptions (Robinson, Zhang, & Tepper, 2018; Yao, Chelmis, & Zois, 2018). Ensuring that systems can be implemented efficiently is crucial if the research community is to meaningfully impact wider society.

4.3 Cross domain applications

Ensuring that abusive content detection systems can be applied across different domains is one of the most difficult but also important frontiers in existing research. Thus far, efforts to address this has been unsuccessful. Burnap and Williams train systems on one type of hate speech (e.g. racism) and apply them to another (e.g. sexism) and find that performance drops considerably (Burnap & Williams, 2016). Karan and Šnajder use a simple methodology to show the huge differences in performance when applying classifiers on different datasets without domain-specific tuning (Karan & Šnajder, 2018). Noticeably, in the EVALITA hate speech detection shared task, participants were asked to (1) train and test a system on Twitter data, (2) on Facebook data and (3) to train on Twitter and test on Facebook (and vice versa). Even the best performing teams reported their systems scored around 10 to 15 F1 points fewer on the cross-domain task. Part of the challenge is that domains vary across many characteristics, including: type of platforms, linguistic practices and dialects of users, how content is created, length of content, social context and the subtask (see above). Accounting for all these sources of variation is a considerable task.

Potential solutions are available to address this issue, such as transfer learning. Initial studies show this can help improve performance by leveraging existing datasets when there is little training data available (Agrawal & Awekar, 2018; Karan & Šnajder, 2018). However, a key challenge in transfer learning is that systems may develop 'bad' learning habits and as such newly created transfer-based models could be more simplistic and unfair (Pan & Fellow, 2009). Thus, whilst transfer learning is a promising avenue for future research, the implications need to be fully investigated.

5 Conclusion

Abusive content detection is a pressing social challenge for which computational methods can have a hugely positive impact. The field has matured considerably and in recent times there have been many advances, particularly in the development of technically sophisticated methods. However, several critical challenges are unsolved, including both those which are longstanding (such as the lack of dataset sharing) and those which have only recently received attention (such as classification biases). There are also many unaddressed frontiers of research. In this paper we have summarized and critically discussed these issues and proposed and discussed possible solutions. We have also demonstrated the utility of social scientific insights for clarifying issues.

Acknowledgements

We thank the anonymous reviewers. Thanks to Ohad Kammar for discussing the Memes example.

Agrawal, S., & Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In *ECIR: Advances in Information Retrieval* (pp. 141–153). https://doi.org/10.1007/978-3-319-76941-7_11

Allen, C. (2011). *Islamophobia*. Surrey: Ashgate.

Amichai-hamburger, Y., & McKenna, K. (2006). The Contact Hypothesis Reconsidered: Interacting via the Internet. *Journal of Computer-Mediated Communication*, *11*(1), 825–843. https://doi.org/10.1111/j.1083-6101.2006.00037.x

Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on Twitter. In *NLDB* (pp. 57–64). https://doi.org/10.1007/978-3-319-91947-8_6

Badjatiya, P., Gupta, M., & Varma, V. (2019). Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations. In *World Wide Web* (pp. 49–59).

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. In *World Wide Web* (pp. 759–760). https://doi.org/10.1145/3041021.3054223

Benesch, S. (2012). *Dangerous Speech: A Proposal to Prevent Group Violence*. New York.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Lecture Notes in Computer Science* (pp. 1–12). https://doi.org/10.1007/978-3-319-67256-4_32

Buchanan, E. (2017). Considering the ethics of big data research: A case of Twitter and ISIS / ISIL. *PLoS ONE*, *12*(12), 1–6.

Bucy, E., & Holbert, L. (2013). *Sourcebook for political communication research*. London: Routledge.

Burnap, P., & Williams, M. (2016). Us and Them: Identifying Cyber Hate on Twitter across Multiple Protected Characteristics. *EPJ Data Science*, *5*(1), 1–15. https://doi.org/10.1140/epjds/s13688-016-0072-6

Caiani, M., & Wagemann, C. (2009). Online networks of the Italian and German Extreme Right. *Information, Communication & Society*, 66–109. https://doi.org/10.1080/13691180802158482

Chandrasekharan, E., Samory, M., Srinivasan, A., & Gilbert, E. (2017). The Bag of Communities. In *CHI* (pp. 3175–3187). https://doi.org/10.1145/3025453.3026018

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media and Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Daniels, J. (2013). Race and racism in Internet Studies: A review and critique. *New Media and Society*, *15*(5), 695–719. https://doi.org/10.1177/1461444812462849

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM* (pp. 1–4).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805v2*, 1–16. Retrieved from http://arxiv.org/abs/1810.04805

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and Mitigating Unintended Bias in Text Classification. In *AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67–73). https://doi.org/10.1145/3278721.3278729

Eatwell, R. (2006). Community Cohesion and Cumulative Extremism in Contemporary Britain. *Political Quarterly*, *77*(2), 204–216. https://doi.org/10.1111/j.1467-923X.2006.00763.x

Eger, S., Şahin, G. G., Rücklé, A., Lee, J.-U., Schulz, C., Mesgar, M., … Gurevych, I. (2019). Text Processing Like Humans Do: Visually Attacking and Shielding NLP Systems. *ArXiv:1903.11508v1*, 1–14.

Eisenstein, J. (2013). What to do about bad language on the Internet. *NAACL HLT*, 359–369.

Eisenstein, J., O'Connor, B., Smith, N., & Xing, E. (2014). Diffusion of lexical change in social media. *PLoS ONE*, *9*(11), 1–13. https://doi.org/10.1371/journal.pone.0113114

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., … Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *ICWSM* (pp. 1–11).

Fraser, N. (1997). *Justice Interruptus: Critical Reflections on the "Postsocialist" Condition*. London: Routledge.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2017). Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *PNAS*, *115*(16), 3635–3644. https://doi.org/10.1073/pnas.1720347115

Golbeck, J., Geller, A. A., Thanki, J., Naik, S., Hoffman, K. M., Wu, D. M., … Jienjitlert, V. (2017). A Large Labeled Corpus for Online Harassment Research. In *WebSci* (pp. 229–233). https://doi.org/10.1145/3091478.3091509

Gorrell, G., Greenwood, M., Roberts, I., Maynard, D., & Bontcheva, K. (2018). Twits, Twats and Twaddle: Trends in Online Abuse towards UK Politicians. In *ICWSM* (pp. 600–603).

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All You Need is "Love": Evading Hate-speech Detection. *ArXiv:1808.09115v2*, 1–11.

Hine, G. E., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., … Blackburn, J. (2017). Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM* (pp. 92–101).

Hines, S., Davy, Z., Monro, S., Motmans, J., Santos, A. C., & Van Der Ros, J. (2018). Introduction to the themed issue: Trans* policy, practice and lived experience within a European context. *Critical Social Policy*, *38*(1), 5–12. https://doi.org/10.1177/0261018317732879

HM Government. (2019). *Online Harms White Paper*. London: Department of Digital, Culture, Media and Society.

Karan, M., & Šnajder, J. (2018). Cross-Domain Detection of Abusive Language Online. In *2nd Workshop on Abusive Language Online* (pp. 132–137). Retrieved from http://takelab.fer.hr/alfeda

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. In *1st Workshop on Abusive Language Online* (pp. 1–11).

Kumar, R., Reganti, A., Bhatia, A., & Maheshwari, T. (2018). Aggression-annotated Corpus of Hindi-English Code-mixed Data. In *LREC* (pp. 1–7). Retrieved from http://arxiv.org/abs/1803.09402

LaMarre, H., Landreville, K., & Beam, M. (2009). The irony of satire: Political ideology and the motivation to see what you want to see in The Colbert Report. *International Journal of Press/Politics*, *14*(2), 212–231. https://doi.org/10.1177/1940161208330904

Lee, Y., Yoon, S., & Jung, K. (2018). Comparative Studies of Detecting Abusive Language on Twitter. In *2nd Workshop on Abusive Language Online* (pp. 101–106).

Lipton, Z. (2016). The Mythos of Model Interpretability. In *ICML Workshop on Human Interpretability in Machine Learning* (pp. 1–9). New York.

Ma, C. (2014). *What are you laughing at? A social semiotic analysis of ironic racial stereotypes in Chappelle's Show*. London.

Magu, R., Joshi, K., & Luo, J. (2017). Detecting the Hate Code on Social Media. In *ICWSM* (pp. 608–611). https://doi.org/10.1016/j.vetimm.2017.02.003

Margetts, H., John, P., Hale, S., & Yasseri, T. (2015). *Political Turbluence: How Social Media Shape Collective Action*. New Jersey: Princeton University Press.

Marwick, A., Blackwell, L., & Lo, K. (2016). *Best practices for conducting risky* *research and protecting yourself from online harassment*. New York: Data & Society.

Marwick, A., & Boyd, D. (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media and Society*, *13*(1), 114–133. https://doi.org/10.1177/1461444810365313

Mehdad, Y., & Tetreault, J. (2016). Do Characters Abuse More Than Words? In *SIGDAL* (pp. 299–303). https://doi.org/10.18653/v1/w16-3638

Müller, K., & Schwarz, C. (2017). Fanning the Flames of Hate: Social Media and Hate Crime. *CAGE Working Paper Series*, 1–82. https://doi.org/10.2139/ssrn.3082972

Nobata, C., Thomas, A., Mehdad, Y., Chang, Y., & Tetreault, J. (2016). Abusive Language Detection in Online User Content. In *World Wide Web* (pp. 145–153). https://doi.org/10.1145/2872427.2883062

Palmer, A., Robinson, M., & Phillips, K. (2017). Illegal is not a Noun: Linguistic Form for Detection of Pejorative Nominalizations. In *1st Workshop on Abusive Language Online* (pp. 91–100). https://doi.org/10.18653/v1/w17-3014

Pan, S. J., & Fellow, Q. Y. (2009). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 1–15.

Papegnies, E., Labatut, V., Dufour, R., & Linarès, G. (2017). Graph-based features for automatic online abuse detection. In *SLSP* (pp. 70–81). https://doi.org/10.1007/978-3-319-68456-7_6

Park, Ji Ho, Shin, J., & Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. In *EMNLP* (pp. 2799–2804).

Park, Ji Hoon, Gabbadon, N. G., & Chernin, A. R. (2006). Naturalizing racial differences through comedy: Asian, black, and white views on racial stereotypes in rush hour 2. *Journal of Communication*, *56*(1), 157–177. https://doi.org/10.1111/j.1460-2466.2006.00008.x

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Detecting Offensive Language in Tweets Using Deep Learning. *ArXiv:1801.04433v1*, 1–17.

Qian, J., ElSherief, M., Belding, E., & Wang, W. Y. (2018). Hierarchical CVAE for Fine-Grained Hate Speech Classification. In *EMNLP* (pp. 3550–3559). https://doi.org/arXiv:1809.00088v1

Raisi, E., & Huang, B. (2016). Cyberbullying Identification Using Participant-Vocabulary Consistency. In *ICML Workshop on #Data4Good* (pp. 46–50). Retrieved from http://arxiv.org/abs/1606.08084

Raisi, E., & Huang, B. (2017). Cyberbullying Detection with Weakly Supervised Machine Learning. In *ASONAM* (pp. 1–8). https://doi.org/10.1145/3110025.3110049

Robinson, D., Zhang, Z., & Tepper, J. (2018). Hate speech detection on Twitter: Feature engineering v.s. feature selection. *ESWC*, 46–49. https://doi.org/10.1007/978-3-319-98192-5_9

Rudin, C. (2018). Please Stop Explaining Black Box Models for High Stakes Decisions. In *NIPS* (pp. 1–15). Retrieved from http://arxiv.org/abs/1811.10154

Ruppert, E., Law, J., & Savage, M. (2013). Reassembling Social Science Methods: The Challenge of Digital Devices. *Theory, Culture & Society*, *30*(4), 22–46. https://doi.org/10.1177/0263276413484941

Sabba, M. (2009). Spelling as a social practice. In J. Maybin & J. Swann (Eds.), *Routledge Companion to English Language Studies* (pp. 243–257). London: Routledge.

Sabou, M., Bontcheva, K., Derczynski, L., & Scharl, A. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. *LREC*, 859–866.

Salminen, J., & Almerekhi, H. (2018). Online Hate Interpretation Varies by Country, But More by Individual. In *SNAMS* (pp. 1–7). https://doi.org/10.1109/SNAMS.2018.8554954

Salminen, J., Almerekhi, H., Milenković, M., Jung, S.-G. G., An, J., Kwak, H., & Jansen, B. (2018). Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models. In *ICWSM* (pp. 330–339).

https://doi.org/10.1109/SNAMS.2018.8554954

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian Twitter Corpus of Hate Speech against Immigrants. In *LREC* (pp. 2798–2805). https://doi.org/10.1561/1500000001

Santos, C. N. dos, Melnyk, I., & Padhi, I. (2018). Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. In *Meeting for ACL* (pp. 189–194). Retrieved from http://arxiv.org/abs/1805.07685

Sartori, G. (1970). Concept misinformation in comparative politics. *American Political Science Review*, *64*(4), 1033–1053. https://doi.org/10.2307/1958356

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In *International Workshop on NLP for Social Media* (pp. 1–10). Valencia, Spain. https://doi.org/10.18653/v1/w17-1101

Serrà, J., Leontiadis, I., Spathis, D., Stringhini, G., Blackburn, J., & Vakali, A. (2017). Class-based Prediction Errors to Detect Hate Speech with Out-of-vocabulary Words. In *1st Workshop on Abusive Language Online* (pp. 36–40). Vancouver, Canada. https://doi.org/10.18653/v1/w17-3005

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the Targets of Hate in Online Social Media. In *IWCSM* (pp. 687–690).

Sugandhi, R., Pande, A., Chawla, S., Agrawal, A., & Bhagat, H. (2016). Methods for detection of cyberbullying: a survey. In *International Conference on Intelligent Systems Design and Applications* (pp. 173–177). https://doi.org/10.1109/ISDA.2015.7489220

Tien, J. H., Eisenberg, M. C., Cherng, S. T., & Porter, M. A. (2019). *Online reactions to the 2017 'Unite the Right' rally in Charlottesville: Measuring polarization in Twitter networks using media followership.*

Twitter. (2018). Enabling further research of information operations on Twitter. Retrieved May 30, 2019, from https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html

Unsvåg, E., & Gambäck, B. (2018). The Effects of User Features on Twitter Hate Speech Detection. In *2nd Workshop on Abusive Language Online* (pp. 75–85). Retrieved from http://aclweb.org/anthology/W18-5110

van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *2nd Workshop on Abusive Language Online* (pp. 33–42).

Van Hee, C., Lefever, E., De Pauw, G., Daelemans, W., Hoste, V., Jacobs, G., … Verhoeven, B. (2018). Automatic detection of cyberbullying in social media text. *PLoS ONE*, *13*(10), 1–21. https://doi.org/10.1371/journal.pone.0203794

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *1st Workshop on Abusive Language Online* (pp. 78–84). https://doi.org/10.1080/17421770903114687

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *NAACL-HLT* (pp. 88–93). https://doi.org/10.18653/v1/n16-2013

Weaver, S. (2010). Developing a rhetorical analysis of racist humour: examining anti-black jokes on the Internet. *Social Semiotics*, *20*(5), 537–555. https://doi.org/10.1080/10350330.2010.513188

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. In *NAACL-HLT* (pp. 602–608).

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. In *World Wide Web* (pp. 1391–1399). Perth, Australia. https://doi.org/10.1039/C5CC05843K

Yao, M., Chelmis, C., & Zois, D. S. (2018). Cyberbullying detection on Instagram with optimal online feature selection. In

ASONAM (pp. 401–408). https://doi.org/10.1109/ASONAM.2018.8508329

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). *ArXiv:1903.08983v3*, 1–12.

Zannettou, S., Caulfield, T., Blackburn, J., & Cristofaro, E. De. (2018). On the Origins of Memes by Means of Fringe Web Communities. In *18th ACM Internet Measurement Conference* (pp. 1–23).

Zhang, Z., Robinson, D., Tepper, J., Gangemi, A., Navigli, R., Vidal, M. E., … Alam, M. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *ESWC* (pp. 745–760). https://doi.org/10.1007/978-3-319-93417-4_48