Can Analytics as a Service Save the Online Discussion Culture? – The Case of Comment Moderation in the Media Industry

Jens Brunk*, Marco Niemann†, Dennis M. Riehle‡

European Research Center for Information Systems (ERCIS)
University of Münster, Leonardo-Campus 3, 48149 Münster, Germany

*jens.brunk@ercis.uni-muenster.de, †marco.niemann@ercis.uni-muenster.de, ‡dennis.riehle@ercis.uni-muenster.de

***Abstract*—In recent years, online public discussions face a proliferation of racist, politically, and religiously motivated hate comments, threats, and insults. With the failure of purely manual moderation, platform operators started searching for semi-automated or even completely automated approaches for comment moderation. One promising option to (semi-) automate the moderation process is the application of Natural Language Processing and Machine Learning (ML) techniques. In this paper we describe the challenges, that currently prevent the application of these techniques and therefore the development of (semi-) and automated solutions. As most of the challenges (e.g., curation of big datasets) require huge financial investments, only big players, such as Google or Facebook, will be able to invest in them. Many of the smaller and medium-sized internet companies will fall behind. To allow this bulk of (media) companies to stay competitive, we design a novel Analytics as a Service (AaaS) offering that will also allow small and medium sized enterprises to profit from ML decision support. We then use the identified challenges to evaluate the conceptual design of the business model and highlight areas of future research to enable the instantiation of the AaaS platform.**

***Index Terms*—comment moderation, machine learning, hate speech, abusive language, moderation, business model**

I. Introduction

The internet and especially the Web 2.0 have always been intended to enhance information exchange and discussion. Despite all good intentions, the menace of abuse and abusive communication has been luring right from the beginning [1]. For example, newspapers and similar online media companies set up comment sections to allow their audiences to interact with each other and their journalists [2]–[4]. Engagement with users was considered to be central to ensure the future economic sustainability of online news outlets [4]. However, the reality is dramatically different for many comment sections. Often, comment sections are "crude, bigoted, or just vile" [2]. Others resign and urge users and fellow journalists to "don't read the comments" [2] or point out that "they're great on paper but not so much in practice" [5]. Numeric estimates for the share of such abusive contents vary from optimistic guesses around 2% on normal websites and approx. 12% on websites like 4chan [2], [6], [7] to rather pessimistic estimates of up to 30% to 80% [8]–[10]. Again others just state that abusive comments simply constitute a "great deal of extra work" [11]. The reactions of news outlets have been diverse. Concerned about potential legal and economic consequences, some outlets released their discussions to social media platforms [11]. Other newspapers lock the comment sections for sensitive articles [10], [11] while again others closed down their discussion fora as a whole [12], [13].

However, journalists usually prefer to refrain from these radical solutions. Some based on their journalist ethos, others out of fear to lose readers and again others out of pure defiance – pointing out that turning away cannot be an option [2], [10], [14], [15]. Most platforms resort to more intense moderation policies in order to keep the comment sections open – while minimizing legal and ethical risks. This includes to enforce mandatory user registration or a change from selective post-moderation (comments moderated after being published, e.g., based on user flagging) towards more rigid pre-moderation (comments are only released after being checked by moderators) [10], [11], [16]. Given the often massive amounts of comments that an outlet receives on a daily basis (e.g., The New York Times reports receiving approx. 9,000 comments per day [17]), moderators often feel overwhelmed. One reason is that comments which contain swear words can be easily filtered [18], but many other comments remain which are not that simple to moderate. However, they still have to be dealt with in a timely manner to keep the profitability of a comment section high [11].

To relieve the moderators and to keep comment sections civil, both practitioners (media and third party community support companies) and academia started to investigate approaches for semi-automated or even completely automated comment moderation [5], [19]. Especially the progress in Natural Language Processing (NLP) and Machine Learning (ML) made these two techniques a promising option. However, the domain of (semi-)automated abusive language detection is still facing substantial challenges. These include the lack of a precise definition of what constitutes abusive language as well as training reliable ML models and ensuring their

acceptance, which we will outline throughout this paper. We further propose a novel Analytics as a Service (AaaS) platform that will address the outlined challenges and especially caters SME media companies, which do not consider dealing with abusive online comments their core competency.

II. Research Design

Our work follows the research paradigm by Österle et al. [20] for design-oriented information systems research. This paradigm consists of four phases: Analysis, Design, Evaluation and Diffusion, as depicted in Fig. 1. For each of the first three phases, we are applying a suitable research method as outlined below. Hence, the research paradigm by Österle et al. leads to a mixed-methods approach.

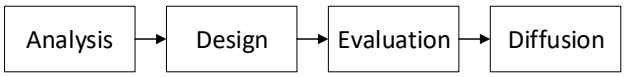

Fig. 1. Research Model

The core part of the **analysis** phase is the identification and description of a business problem [20]. In our case, this is initiated both from practitioner side as well as the scientific side, as discussed in Chapter I. For this, we conducted a structured literature analysis according to [21], in which we have identified four different challenges that are further described in Chapter III. To address the identified challenges, we **design** a business model to develop an Analytics as a Service platform. For the business model generation, we use the Business Model Canvas (BMC) method [22]. These artifacts are presented in Chapter IV. Given that our business model is not yet in the market, it cannot be evaluated in a holistic real-world scenario. Therefore, our **evaluation** follows the principle of logical reasoning, which is an adequate method to use in such a scenario and especially suitable for design specifications [23]. Lastly, the paper at hand aims at the **diffusion** of our work.

III. Challenges

Despite more than one decade of abusive language research that has become increasingly intense, we still face several substantial challenges. These range from the fundamental issue of defining the precise topic we are working on (see Section III-A), the creation of a reliable data basis (see Section III-B) and the actual fine tuning of machine learning models (see Section III-C) to the final acceptance of the end users (see Section III-D).

A. What Constitutes Abusive Language

Looking at the overall challenges that the media industry is facing when it comes to the moderation of abusive comments – and especially automating or at least semi-automating the process – one of the first challenges is a seemingly simple one: to find a definition of what constitutes abusive language. One of the first attempts to define abusive language concepts for the purpose of detecting it by software has been conducted by [24] in 2009; since then a steadily growing stream of research has been taking up the problem. However, so far there is no consistent definition that is used across research. Instead, we are still observing many different definitions and annotation schemes that are used [19], [25]. The only exception are some papers, which pick up existing datasets to optimize ML models which have been previously built for them. Even the few surveys on abusive language concepts, such as "hate speech", explicitly point out that these concepts are typically hard to define [26]–[29]. Taking it a bit further, until now there is not even clarity about the actual concept names that have to be discussed: Is it "harassment", "abusive language", "offensive language", "hate speech" or a combination of multiple concepts? The above stated issue of consistency becomes clearer when one looks closer into these concepts. For example, the concept of "harassment' is defined by [24] as the intentional annoyance of persons, while [30] also subsumed "cyberbullying" and [31], [32] added "hate speech" and "self-harm". Similarly, [33] understand "hate speech" as a sub-concept to "offensive language", while [34], [35] treat it vice versa. The issues associated with such missing or ambiguous definitions are manifold: On the one hand, websites risk to lose users by deleting comments too radically [27], while on the other hand operators risk losing investors or law suits for not filtering highly problematic content [19]. However, repeatedly changing concepts do not only have a severe direct impact on the industry but also inhibit research that focuses on actually solving the problem.

Therefore, it is necessary that the abusive language community further works on a standard or at least a highly standardized process to create such definitions. Here, it will be crucial to go beyond the purely academic perspective, e.g., by deeply integrating a legal perspective into the conceptualization [28], [36]. This aspect will be central in many countries, because typically online platforms are subject to regulation, which often includes limitations on certain kinds of speech. For example Germany restricts "hate speech" [37], "insults" [38] and "threats" [39] while the EU and CoE restrict "hate speech" and "cyber-bullying" [40], [41]. Violations are often linked to fees which easily reach several thousand Euros. However, it is also crucial to consider legal boundaries in the opposite direction, as many democratic countries codify free speech [42]–[45], which in case of violation yet again can create a costly public uproar or even law suits [46], [47].

Beyond these obligatory aspects, future definitions may also have to consider further business and platform requirements. As, e.g., observed by [32], many media companies release platform specific restrictions on language to cater their specific audiences. While some more traditional or educational outlets often implement rather restrictive guidelines, many blogs, independent or smaller websites are rather non-restrictive with the user-generated content. While this may be hard to generalize into one unified definition, the creation process for such a definition should at least account for the option to add a certain degree of customization to filter content unsuitable for a very specific platform.

Given the existence of a massive amount of abusive language research (as indicated above), it will be helpful to consult academic publications as well as to benefit from already existing (legal) interpretation and contextualization. Studies like the ones by [36], [26] and [27] can help researchers and practitioners to get into the topic without having to read through all forms of law books, cases and existing studies. Furthermore, it will help to consider existing academic work for future definitions to keep links to the existing stream of research and to reuse whatever possible in order to avoid lost investments. However, under the premise that definitions should be applicable for real-world moderation scenarios, the focus should be on the legal requirements (and partly the business-specific properties), as these are the ones that practitioners will have to adhere to.

B. Labeling of Large and Accurate Data Sets

Consequently, the second problem is that it is very difficult to label a sufficiently large and accurate data set for machine learning. This problem is tightly coupled with the general challenges that big data poses: *volume* (large amount of data), *velocity* (high frequency of new data generation), *variety* (many different types of structured and unstructured data) and *veracity* (uncertainty regarding data quality) [48], [49]. User comments typically appear in very large numbers, it is not uncommon for popular Facebook pages or Instagram profiles to receive multiple ten thousands of comments per day. Additionally, as comments are natural language texts, they are unstructured data and need further pre-processing before they can be used as training data for machine learning algorithms. Most important is the labeling of the data, which is the (usually) manual classification of a data set, where labels (like publish/not-publish) are manually assigned to each record in the training data set. This is a very time-consuming and, hence, expensive process, as described by several researchers (e.g., [13], [19]).

One approach towards labeling large data sets is crowd-sourcing [50]. With crowd-sourcing, the labeling of data records is split into mini tasks, where each task consists of assigning one or a few labels. Via an online platform, these tasks are then spread to a set of humans, so called crowd-workers, who solve these tasks. Crowd-workers are usually monetarily rewarded on a per-task basis. Generally, crowd-workers are free to choose their working times and whether they want to solve a task or not. Therefore, the micro tasks might be solved by many different individuals and the customer usually cannot influence the distribution of micro tasks. However, crowd-working platforms like Figure Eight[1], Amazon Mechanical Turk[2], etc. enable the customer to choose crowd-workers according to given skill sets.

In a small pre-study, we have used the crowd-working platform Crowdguru[3] to label user-generated comments. Crowd-workers received an initial introduction regarding the rating schema and tasks. Additionally, attention and concentration tests of crowd-workers were added. All in all, each comment was rated by ten different crowd-workers, which led us to costs of roughly one euro per comment. While the data we gathered with our small pre-study – in line with prior research, e.g., by [51] or [25] – shows that labeling large data sets by means of crowd-working works and helps in labeling large data sets, it is questionable if this is suitable for large scale projects, where costs are a driving factor.

When there are ten different people, there might be – as a proverb says – eleven different opinions. This also applies to crowd-sourcing. In our small pre-study, we followed the simple rule of majority. However, the inter-rater agreement is a problem that is discussed in literature as well (e.g., [34]). Therefore, also other methods should be considered. For example, large media companies usually have full-time community managers, i.e., employees who are concerned with moderating comments. The work of these employees can be collected in a structured way and used as training data for machine learning. While this does not guarantee that two different community managers have the same opinion on a comment, media companies tend to have very precise and transparent rules for moderation and, hence, we expect that labels assigned by professional community managers will differ less than labels assigned by random crowd-workers. A solution to this issue is the development of a ”custom crowd” [52], where workers are carefully selected, groomed and curated. This, however, increases the costs compared to a randomly selected crowd.

Nevertheless, even if applicable methods are found, only few large media companies and platform operators have the required manpower and the financial resources to curate sufficiently big data sets for natural language processing (NLP).

C. Training Machine Learning Models

Going further down the road towards a functioning abusive language classifier, one meets one of the classic machine learning problems: There is no one-size-fits-all algorithm that delivers optimal results for each problem [53]. In consequence, researchers are still struggling to identify a set of optimal classification, respectively detection models, that exhibit sufficient predictive accuracy and stability to reliably identify abusive comments. Skimming through the last 20 years of abusive language detection research, one can easily find dozens of different algorithms that were applied: Many of the early publications primarily focused on standard classification algorithms, such as logistic regression [25], [51], [54]–[56], decision trees [51], [57]–[60], random forests [51], [54], [60], [61], Naïve Bayes [51], [57], [58], [62] and especially support vector machines (SVM) [24], [51], [54], [57]–[59], [61]–[63], which could even well be the most popular classification method used in the domain so far. Even though these algorithms are rather basic algorithms from the early days of machine learning (e.g., logistic regression from 1958 [64] or decision trees like CART [65]) many of them are still popular in recent papers, as indicated by the above citations. While

[1]https://www.figure-eight.com/
[2]https://www.mturk.com/
[3]https://www.crowdguru.de/

these algorithms were apparently state of the art in the early days of abusive language detection, some works still identify them as the top performing algorithms [51], choose them as their primary classifier based on their merits [66] or still indicate their ability to compete with the new state-of-the-art algorithms [25]. Following the general trend in machine learning, neural network (ANN)-based approaches have been gaining traction in the domain of abusive language detection since 2016. Similar to the traditional classifiers, most common ANN-architectures such as Recurrent Neural Networks (RNN) [67]–[69], Convolutional Neural Networks (CNN) [61], [68], [70], [71] and Long Short-Term Memory (LSTM) Networks [3], [56], [61], [70] have been tested in various configurations. Again, different authors report different architectures to be superior — and as we can see in the case of [25] the differences between can be arguably small (less than 1% in accuracy for the best performing models).

Unfortunately, not only the employed algorithms differ across the various publications but also the metrics that are used to assess their predictive performance. The primary four metrics are precision [19], [54], [56], [61], recall [19], [54], [56], [61], [72], the F-score [19], [54], [56], [61] and accuracy [25], [55], [68], which however are not all used in each publication. To make things even worse, each of these metrics are sometimes computed differently (e.g., micro-/macro-averaged, see [73]). This makes it even more difficult to reliably determine algorithms that outperform their competitors, since metrics might be missing in some cases and in others might be conflicting.

Hence, even though a considerable amount of work has been put into the evaluation of suitable algorithms, it remains an open challenge. But even with a powerful state-of-the art classifier at hand, there are further aspects that will require attention: To identify the correct parameter configuration for the algorithm to perform properly. To obtain a reasonable representation of inputs for the classifier to work with – be it lexica, n-grams, word embeddings or another yet unconsidered option [19], [66]. To enable classifier configuration, to fine-tune models to moderate according to user-defined criteria (degree of automation, thresholds for automated deletion vs. warning, kinds of language analyzed), and to make them interpretable so that results can be used to support human moderators [74].

As a consequence, future research should put more focus on clearly communicating details of the algorithms used, including their parametrization (either in paper form or via the provision of source code), reporting all relevant evaluation results according to a unified metric schema, so that we can start to reduce the search space by having fully comparable and interpretable results. Another viable alternative could be the concept of automated machine learning (AutoML). Instead of manually selecting and optimizing the ML algorithm, AutoML aims at an automated selection and optimization [75], [76]. However, even though AutoML has proven to be promising in many domains, corresponding results for the abusive language domain are still pending.

D. Acceptance of Moderation Systems

A fourth challenge of (semi-) automated online comment moderation is the acceptance of such systems by the user base. Especially in countries which codify the freedom of speech in their constitutions, people are often fervently arguing against automated or even semi-automated moderation, calling such procedures “censorship”, “oppression”, and the “end of free speech”.

The above-mentioned methods for detecting abusive content have one thing in common: They provide a previously trained decision model, which separates acceptable and publishable from non-acceptable and to be moderated comments. Unfortunately, the degree of comprehensibility and traceability of these models differ largely between approaches. More often than not they present themselves as black boxes to the user rather than an open book. However, once a system influences the user and defines whether his comment is acceptable, he wants to understand how it works. One way that platform providers can try to achieve user understanding and by that acceptance of the deployed systems is through explaining the decision. Previous research lays the ground work for the concepts of explanations for intelligent systems, which comment moderation systems are a subset of. For example, Gregor and Benbasat provide an overview of explanatory constructs that are used in empirical research [77]. In their work, they derived various propositions; the most relevant one for us is that ”explanations will be used when the user experiences an expectation failure, or perceives an anomaly” [77, p. 506]. This is in line with our hypothesis that users of discussion platforms – which apply automatic comment moderation systems – expect explanations of the systems decision, especially if their comment is denied.

Recently, a new European Union (EU) regulation on algorithmic transparency was introduced. It specifies that users of an intelligent system have the right for an explanation of the decisions made [78]. Due to the complex nature of often applied machine learning methods, this similarly calls for a higher transparency of the decision models, which interact with the user-generated content [79], [80].

A higher transparency of the comment moderation system, e.g., through the use of explanations, will enable the user to have increased trust in the system. For example, Wang recently attempted to visualize features with decisive influence in a neural network that identifies hate speech [74]. Trust is a very complex construct to study, but previous information systems (IS) research showed that a user feels more confident in a system, if he understands it [81], [82]. Therefore, higher levels of trust have been found to increase the chance of the user accepting the system [83], [84].

Therefore, it is crucial to develop ML models which deliver interpretable results that can be understood by moderators and users alike; to keep potential deletions transparent at any time. One very promising research stream in this regard is *explainable artificial intelligence*. Here, researchers try to develop methods or concepts that enable just this decision

transparency (e.g. [85]). So far, this research has mostly been applied in other domains or settings and has yet only scratched the surface of the above-mentioned issues. Therefore, further research - especially with a focus on online comment moderation systems - is needed.

IV. How to Save the Media Industry?

As we pointed out before, it will be essential for online discussion platform providers—especially for smaller and medium sized ones—to deploy (semi-) automated comment moderation tools. First successful attempts to deploy such a tool have been reported in the literature (e.g. [70]). However, most of the published solutions only address some of the four challenges outlined above. While some rather seminal papers like [19] handle up to three of the challenges, many others at best deal with two [86], leaving large parts of the overall issue unaddressed. Furthermore, not every platform operator or newspaper organization is capable to develop a system of their own. Therefore, the question arises on how to address all of the above-stated challenges with a viable service-oriented business model that can also provide the required analytic functionality to SME customers.

To start to address this question we make use of the Business Model Canvas (BMC). The BMC is a very common, easy to use and well recognized tool to develop and explain business models. It is a strategic management tool, which uses a visual chart of nine building blocks to describe a business' value propositions, infrastructure, customer and finances [22]. Even though not originally developed for this purpose, the Business Model Canvas is receiving considerable uptake for both design and assessment of data- and analytics-driven business models [87]–[89]. In the following subsections we use the BMC to develop a conceptual AaaS platform business model for online comment moderation (see Fig. 2), which we will then use to further address, match and outline the previously mentioned challenges.

A. Customer Segments

"The customer segments building block defines the different groups of people or organizations an enterprise aims to reach and serve" [22, p. 20]. In this case, the AaaS platform serves a segmented market, because it distinguishes between different types of customers with different needs and problems. There will be small, medium and large enterprise customers. Small to medium sized customers might only need individual comment evaluations from time to time. Medium sized customers need a defined package that enables them to query a predefined amount of comment evaluations, moderation proposals and e.g. a moderation dashboard to handle their requests. Finally, large enterprises might be interested in using a flat rate offer to leverage economies of scale. Additionally, different larger customers might have different moderation styles (more or less strict) that need to be represented in the evaluation models of the platform.

B. Value Proposition

The value proposition building block describes which value the platform offers to the customers through products or services. The AaaS platform creates value for its differently sized customers in three ways. First and foremost the manual moderation effort of the customer is reduced. This is a major cost factor for the customer, as it is done by a human employee and is quite time consuming. Secondly, the user-generated content is moderated in a structured manner and moderation decisions are well documented. This improves the quality and uniformity of the moderation. Finally, platform providers are forced to shut down parts or even their whole discussion platform without the use of (semi-) automated comment moderation systems. Therefore, the deployment of AaaS resurrects and increases their interaction with their user-base, the activity and traffic on their website and thus revenue potential through ads and increased turnover.

C. Channels

The channels building block explains how a firm delivers their value proposition to their customers but also how it reaches new customers and communicates with existing ones. In this work we abstract from communication and marketing channels, as this is an issue of the exact instantiation of this conceptual business model. The delivery of the value proposition, which is the comment evaluation and the ability to moderate comments in an appropriate dashboard, is provided digitally. The data of the customers' own content management system and the AaaS platform is exchanged live through pre-defined application programming interfaces (API), which us push, pull and receive protocols.

D. Customer Relationships

This building block describes which kind of relationships the business maintains with its respective customer segments. In general, all customer segments (small to large customers) are provided with a self-service interface. They can submit the comments that need to be evaluated, receive and display the results and possibly manage them in the provided interface. On top of that, however, specific customers may want the trained evaluation models to be adapted to particular needs. This represents a second type of customer relationship that goes beyond the previously mentioned self-service infrastructure.

E. Revenue Streams

The revenue streams represent all incoming turnover that the business generates from its customers. Similar to the previous building block the revenue stream is twofold. The first includes consulting, adapting and also implementing on a customer basis; e.g. the previously mentioned customization of the evaluation models or an API-provision for individual content management systems. The second and major revenue stream represents the subscription model that the different customer segments utilize to receive evaluations for their comments.

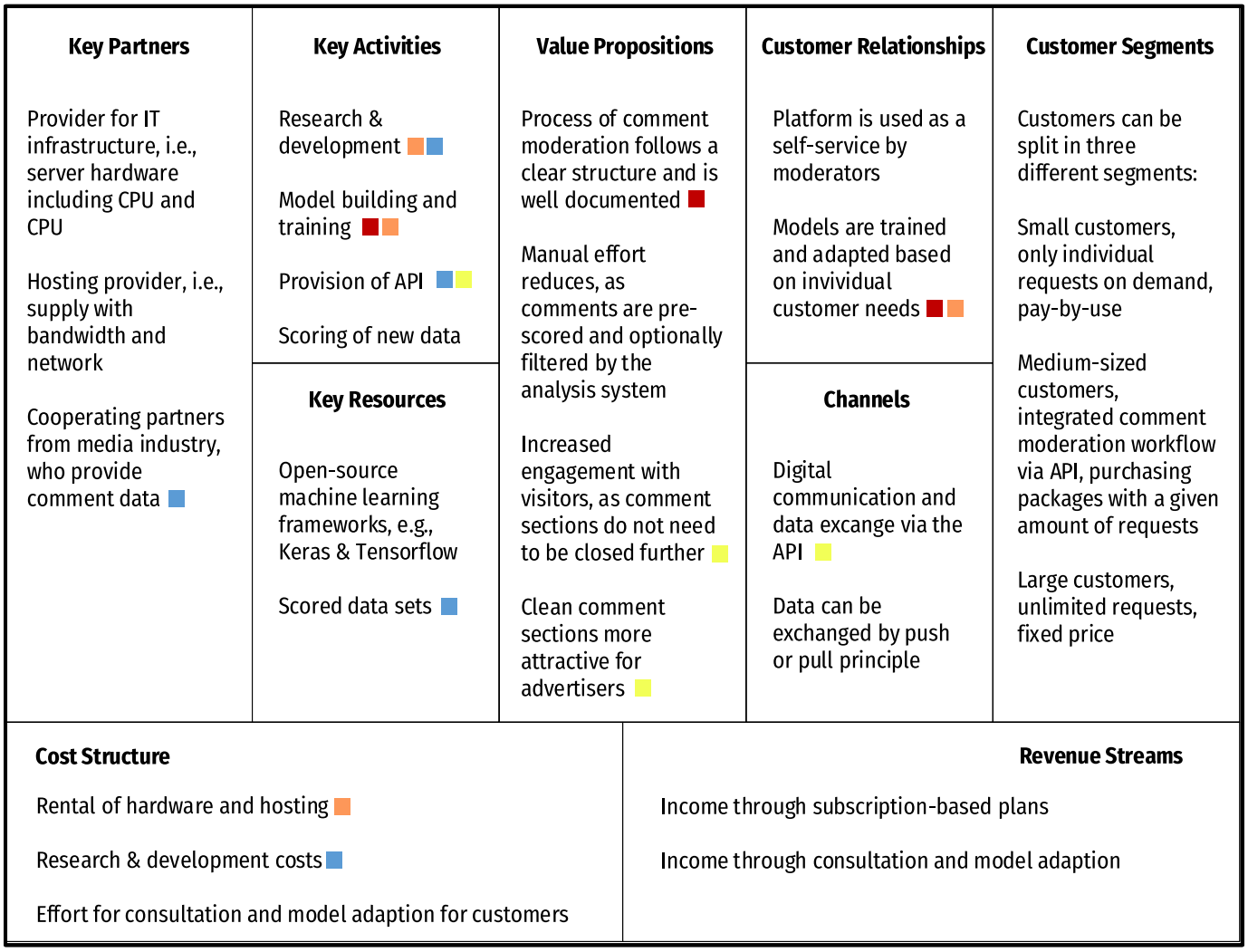

Fig. 2. Business Model According to the BMC

F. Key Resources

This building block represents the key assets that the business model is based upon and that are essential to its success. The AaaS platform makes use of a variety of machine learning methods, tools, frameworks and libraries to calculate the evaluation models. Many of the applied assets are open source or available under different accessible licenses. At the point of writing, these would, e.g., include tools, such as TensorFlow or Keras. Another key resource that is essential for the calculation of the evaluation models are data sets. These data sets include real world comments annotated with labels and additional information that the machine learning algorithms leverage to extract their decision making patterns. These data sets can stem from partners, customers or might be even self developed.

G. Key Activities

The key activities building block describes the most important tasks that the business needs to perform in order to deliver its business value to the customers. For our platform, the most important activity is the evaluation of comments submitted by the customers. To be able to do this, two secondary activities need to be performed. On the one hand, the APIs need to be implemented and provided beforehand. And on the other hand the evaluation model needs to be trained, which is a continuous challenge, as language is not a static but evolving construct.

H. Key Partnerships

Key Partnerships include all the relations and strategic partnerships that enable the business model to function. Considering the previous two building blocks, a strategic partnership with the developers of the machine learning tools and frameworks is of importance. Through this, the platform can ensure a timely and continuous delivery of necessary updates. Similarly, data set providers can be important partners. For example, some customers might also be data deliverers at the same time and therefore the relationship should receive special attention. Another more basic, but as important, partnership is the IT infrastructure and hosting of the platform. The

calculation and provision of the machine learning evaluation models requires great amounts of calculation power and the availability of the service must always be granted.

I. Cost Structure

Last but not least, the cost structure includes all the cost that accumulate by executing the business model. Of course, the previously mentioned infrastructure, hosting, and computing power incurs major costs on the business. Furthermore research and continuous development on how to improve, adapt, and optimize the evaluation models is very important. And ultimately, the staff that executes the consulting and adaption actions with (future) customers must be noted here as well.

V. Discussion

So far we pointed out the necessity of an AaaS platform for online comment moderation and described how a viable business model could look like. In the following, we match the identified challenges of the domain to the developed business model canvas and its building blocks. By doing so, we evaluate our design and identify key areas for research and development. This advancement will ultimately enable the implementation of our envisioned AaaS platform. For this purpose, each challenge and the relevant building blocks of the BMC are marked with a colour to represent their association (see Fig. 2).

A. What Constitutes Abusive Language

The first challenge deals with the complicated, uniform and inconsistent definition of what abusive language and the to-be filtered content actually is. As pointed out before, this is an essential piece to the puzzle of creating good evaluation models. Therefore, this challenge strongly weights on the key activity *model building and training*. This is especially true, if you consider that part of the value proposition and customer relationship support is, that *models are trained and adapted based on individual customer needs*. This means that different discussion platform operators will have different thresholds on the scale of unacceptable to acceptable content and might even need to include language features, which others do not. But in the end, a common understanding of what abusive language is, is needed to initiate this discussion. With its absence, the value proposition to create a *clear structure and process of comment moderation* can not be reached.

B. Labeling of Large and Accurate Data Sets

The second challenge describes the process of creating a large and accurate data set, which is then used to learn the evaluation models. An important part of this challenge are the *cooperating partners from the media industry, who provide comment data*. If these customers/partners follow the jointly developed moderation process and the included decision rules, then they automatically generate new data and thus data sets. On the providers site, the key activity *provision of API* is important, since the previously described work flow is only possible, if the APIs support it.

As language is constantly evolving, *research and development* as well as its *costs* are influenced by this challenge. The value of the existing data continuously looses value as time progresses. Therefore new data sets need to be developed and the models and definitions need to be adapted. Additionally, there are two more perspectives to this segment. One is to curate data sets yourself and the second is to use third party *scored data sets*. These represent one of the key resources.

C. Training Machine Learning Models

The third challenge also touches on some building blocks that the labeling of data sets is related to. Similarly, the training of machine learning models needs intensive *research and development* in order to perform the *model building and training*. The models, which are *adapted based on individual customer needs* display an even higher degree of difficulty. For being able to perform the computation intensive training of machine learning models, the *rental of hardware and hosting* is needed to provide the necessary computational power.

D. Acceptance of Moderation Systems

The fourth and final identified challenge concerns itself more with the end user acceptance rather than the providers side of the BMC. Nevertheless, there are key building blocks that are affected. The working theory at the moment is, that, e.g., through explanations of the intelligent and (semi-) automated systems, user acceptance can be increased. Only if user acceptance exists we can deliver the value proposition of *clean comment sections, which are more attractive for advertisers* and via that an *increased engagement with visitors, as comment sections do not need to be closed further*. To provide reasonable explanations, the key activity of the *provision of API* and the channel of *digital communication and data exchange via the API* need to support the exchange of these secondary attributes.

VI. Conclusion and Outlook

Looking back at slightly more than one decade of abusive language detection and moderation research, a lot of important work has been done in almost all related areas: Defining the actual issue at hand, collecting data sets and engineering first machine learning models to assist newspaper moderators. However, we can still observe many points being left disregarded or at least not being addressed in a conclusive manner: Till the current day we have no undebated definition of what actually constitutes abusive language (or hate speech, harassment, ...). Datasets are still often proprietary and not open for public usage, respectively very difficult to compile for independent actors given the complexity and cost of such undertakings. This especially applies to SME newspapers, who also experience problems with training suitable ML algorithms given the sparsity of capable data scientists and the cost associated. Even for large companies and academics many points such as optimal algorithms or metrics are still open for debate. Last but not least, interpretability and algorithmic

transparency are becoming increasingly important, both as a support for moderators as well as as an explanation for users to convey moderation decisions.

With our work, we have analyzed recent literature on online comment moderation systems and have identified four challenges towards the implementation of analytics as a service platforms. We have proposed and designed a business model with the BMC method, which supports both practitioners and researchers on the specification and implementation of such platforms. Currently, there are still some aspects, which require further research: Data security and privacy have not been considered so far. For implementing an appropriate platform, the participating SME newspapers need to agree on a suitable model for data ownership. In addition, legal requirements may (partly) require anonymization of exchanged data. Therefore, legal aspects need to be considered carefully. Lastly, we have not yet distinguished between pre- and post-moderation of comments. While pre-moderation would happen by an ML algorithm as discussed in this paper, community managers still need to be able to change the algorithm's decision in a post-moderation process. While from a process and data perspective this is a trivial process, it still needs to be considered when workflow models are created during an implementation phase.

Acknowledgements

The research leading to these results received funding from the federal state of North Rhine-Westphalia and the European Regional Development Fund (EFRE.NRW 2014-2020), Project: MODERAT! (No. CM-2-2-036a).

References

[1] A. Bastidas, E. Dixon, C. Loo, and J. Ryan, "Harassment detection: a benchmark on the #HackHarassment dataset," in *Proc. Collab. Eur. Res. Conf.*, ser. CERC 2016, U. Bleimann, B. Humm, R. Loew, I. Stengel, and P. Walsh, Eds., Cork, Ireland, sep 2016, pp. 76–79. [Online]. Available: http://arxiv.org/abs/1609.02809

[2] B. Gardiner, M. Mansfield, I. Anderson, J. Holder, D. Louter, and M. Ulmanu, "The dark side of Guardian comments," 2016. [Online]. Available: https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments

[3] V. Kolhatkar and M. Taboada, "Constructive Language in News Comments," in *Proc. First Work. Abus. Lang. Online*, ser. ALW1, Z. Waseem, W. H. K. Chung, D. Hovy, and J. Tetreault, Eds. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 11–17.

[4] S. C. Lewis, A. E. Holton, and M. Coddington, "Reciprocal Journalism: A concept of mutual exchange between journalists and audiences," *Journal. Pract.*, vol. 8, no. 2, pp. 229–241, 2014.

[5] R. Bilton, "Why some publishers are killing their comment sections," 2014. [Online]. Available: https://digiday.com/media/comments-sections/

[6] M. Mansfield, "How we analysed 70m comments on the Guardian website," 2016. [Online]. Available: https://www.theguardian.com/technology/2016/apr/12/how-we-analysed-70m-comments-guardian-website

[7] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn, "Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web," in *Proc. 11th Int. Conf. Web Soc. Media*, ser. ICWSM-2017. Montral, Canada: Association for the Advancement of Artificial Intelligence, 2017, pp. 92–101.

[8] Z. Papacharissi, "Democracy online: civility, politeness, and the democratic potential of online political discussion groups," *New Media Soc.*, vol. 6, no. 2, pp. 259–283, 2004.

[9] J. Cheng, "Report: 80 percent of blogs contain "offensive" content," 2007. [Online]. Available: https://arstechnica.com/information-technology/2007/04/report-80-percent-of-blogs-contain-offensive-content/

[10] S. Boberg, T. Schatto-Eckrodt, L. Frischlich, and T. Quandt, "The Moral Gatekeeper? Moderation and Deletion of User-Generated Content in a Leading News Forum," *Media Commun.*, vol. 6, no. 4, pp. 58–69, 2018.

[11] R. Pöyhtäri, "Limits of Hate Speech and Freedom of Speech on Moderated News Websites in Finland , Sweden , the Netherlands and the UK," *Ann. · Ser. hist. sociol.*, vol. 24, no. 3, pp. 513–524, 2014.

[12] The Coral Project Community, 2016. [Online]. Available: https://community.coralproject.net/t/shutting-down-onsite-comments-a-comprehensive-list-of-all-news-organisations/347

[13] S. Köffer, D. M. Riehle, S. Höhenberger, and J. Becker, "Discussing the Value of Automatic Hate Speech Detection in Online Debates," in *Tagungsband Multikonferenz Wirtschaftsinformatik 2018*, ser. MKWI 2018, P. Drews, B. Funk, P. Niemeyer, and L. Xie, Eds. Lüneburg, Germany: Leuphana Universität, 2018.

[14] N. Diakopoulos, "Picking the NYT Picks : Editorial Criteria and Automation in the Curation of Online News Comments," *#ISOJ, Off. Res. J. ISOJ*, vol. 5, no. 1, pp. 147–166, 2015.

[15] S. Plöchinger, "Über den Hass," 2016. [Online]. Available: http://ploechinger.tumblr.com/post/140370770262/{\"{u}}ber-den-hass

[16] T. B. Ksiazek, "Civil Interactivity: How News Organizations' Commenting Policies Explain Civility and Hostility in User Comments," *J. Broadcast. Electron. Media*, vol. 59, no. 4, pp. 556–573, 2015.

[17] B. Etim, "The Most Popular Reader Comments on The Times," 2015. [Online]. Available: https://www.nytimes.com/2015/11/23/insider/the-most-popular-reader-comments-on-the-times.html

[18] A. Muddiman and N. J. Stroud, "News Values, Cognitive Biases, and Partisan Incivility in Comment Sections," *J. Commun.*, vol. 67, no. 4, pp. 586–609, 2017.

[19] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive Language Detection in Online User Content," in *Proc. 25th Int. Conf. World Wide Web*, ser. WWW '16. Montreal, Canada: ACM Press, 2016, pp. 145–153.

[20] H. Österle, J. Becker, U. Frank, T. Hess, D. Karagiannis, H. Krcmar, P. Loos, P. Mertens, A. Oberweis, and E. J. Sinz, "Memorandum on design-oriented information systems research," *European Journal of Information Systems*, vol. 20, no. 1, pp. 7–10, 2011.

[21] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS quarterly*, pp. xiii–xxiii, 2002.

[22] A. Osterwalder and Y. Pigneur, *Business model generation: a handbook for visionaries, game changers, and challengers*. John Wiley & Sons, 2010.

[23] C. Sonnenberg and J. vom Brocke, "Evaluations in the science of the artificial — reconsidering the build-evaluate pattern in design science research," in *Proceedings of the 7th International Conference on Design Science Research in Information Systems: Advances in Theory and Practice*, ser. DESRIST'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 381–397. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-29863-9_28

[24] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0," in *Proc. Content Anal. WEB*, ser. CAW2.0, Madrid, Spain, 2009, pp. 1–7.

[25] E. Wulczyn, N. Thain, and L. Dixon, "Ex Machina," in *Proc. 26th Int. Conf. World Wide Web*, ser. WWW '17. Perth, Australia: ACM Press, 2017, pp. 1391–1399.

[26] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–30, 2018.

[27] K. Gelber, "Differentiating hate speech: a systemic discrimination approach," *Crit. Rev. Int. Soc. Polit. Philos.*, pp. 1–22, 2019.

[28] M. Niemann, D. M. Riehle, J. Brunk, and J. Becker, "What is Abusive Language? Integrating Different Views on Abusive Language for Machine Learning," in *Multidiscip. Int. Symp. Disinformation Open Online Media*, ser. MISDOOM 2019, Hamburg, Germany, 2019.

[29] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," in *Proc. Fifth Int. Work. Nat. Lang. Process. Soc. Media*, ser. SocialNLP 2017, L.-W. Ku and C.-T. Li, Eds. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 1–10.

[30] U. Bretschneider, T. Wöhner, and R. Peters, "Detecting Online Harassment in Social Networks," in *Proc. Int. Conf. Inf. Syst. - Build. a Better World through Inf. Syst.*, ser. ICIS 2014, M. D. Myers and D. W. Straub,

Eds. Auckland, New Zealand: Association for Information Systems, 2014, pp. 1–14.

[31] J. Guberman and L. Hemphill, "Challenges in Modifying Existing Scales for Detecting Harassment in Individual Tweets," in *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, ser. HICSS 2017. Waikoloa Village, Hawaii, USA: Association for Information Systems, 2017, pp. 2203–2212.

[32] J. A. Pater, M. K. Kim, E. D. Mynatt, and C. Fiesler, "Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms," in *Proc. 19th Int. Conf. Support. Gr. Work*, ser. GROUP '16. Sanibel Island, FL, USA: ACM Press, 2016, pp. 369–374.

[33] W. Warner and J. Hirschberg, "Detecting Hate Speech on the World Wide Web," in *Proc. Second Work. Lang. Soc. Media*, ser. LSM '12, S. O. Sood, M. Nagarajan, and M. Gamon, Eds. Montreal, Canada: Association for Computational Linguistics, 2012, pp. 19–26.

[34] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis," in *Proc. 3rd Work. Nat. Lang. Process. Comput. Commun.*, ser. NLP4CMC III, M. Beißwenger, M. Wojatzki, and T. Zesch, Eds. Bochum, Germany: Stefanie Dipper, Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, 2016, pp. 6–9.

[35] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in *Proc. NAACL Student Res. Work.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 88–93.

[36] D. Fišer, T. Erjavec, and N. Ljubešić, "Legal Framework, Dataset and Annotation Schema for Socially Unacceptable Online Discourse Practices in Slovene," in *Proc. First Work. Abus. Lang. Online*, ser. ALW1, Z. Waseem, W. H. K. Chung, D. Hovy, and J. Tetreault, Eds. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 46–51.

[37] § 130 StGB, "Volksverhetzung," in *Strafgesetzb. der Fassung der Bekanntmachung vom 13. Novemb. 1998 (BGBl. I S. 3322), das zuletzt durch Artik. 22 Absatz 5 des Gesetzes vom 23. Juni 2017 (BGBl. I S. 1693) geändert worden ist*. Deutscher Bundestag, 2017.

[38] § 185 StGB, "Beleidigung," in *Strafgesetzb. der Fassung der Bekanntmachung vom 13. Novemb. 1998 (BGBl. I S. 3322), das zuletzt durch Artik. 22 Absatz 5 des Gesetzes vom 23. Juni 2017 (BGBl. I S. 1693) geändert worden ist*. Deutscher Bundestag, 2017.

[39] § 241 StGB, "Bedrohung," in *Strafgesetzb. der Fassung der Bekanntmachung vom 13. Novemb. 1998 (BGBl. I S. 3322), das zuletzt durch Artik. 22 Absatz 5 des Gesetzes vom 23. Juni 2017 (BGBl. I S. 1693) geändert worden ist*. Deutscher Bundestag, 2017.

[40] Council of Europe, "Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "Hate Speech"," 1997.

[41] European Commission against Racism and Intolerance, "ECRI General Policy Recommendation No. 6 on Combating the Dissemination of Racist, Xenophobic and Antisemitic Material via the Internet," 2000.

[42] Amendment 1, "Amendment 1 - Freedom of Religion, Press, Expression," in *Const. United States*, 1791.

[43] Art 5 GG, "Grundgesetz für die Bundesrepublik Deutschland," in *Grundgesetz für die Bundesrepublik Deutschl. der im Bundesgesetzblatt Tl. III, Gliederungsnummer 100-1, veröffentlichten bereinigten Fassung, das zuletzt durch Artik. 1 des Gesetzes vom 13. Juli 2017 (BGBl. I S. 2347) geändert worden ist*. Deutscher Bundestag, 2014.

[44] Council of Europe, "European Convention on Human Rights," 2010.

[45] European Union, "The Charter of Fundamental Rights of the European Union," *Off. J. Eur. Communities*, vol. C 364, pp. 1–22, 2000.

[46] P. Evans, "Will Germany's new law kill free speech online?" 2017. [Online]. Available: https://www.bbc.com/news/blogs-trending-41042266

[47] C. George, "Challenging Hate Speech - A Dilemma for Journalists," 2016. [Online]. Available: https://ethicaljournalismnetwork.org/resources/publications/ethics-in-the-news/hate-speech

[48] A. McAfee and E. Brynjolfsson, "Big data: the management revolution." *Harvard business review*, vol. 90, no. 10, p. 60, 2012.

[49] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics–challenges in topic discovery, data collection, and data preparation," *International journal of information management*, vol. 39, pp. 156–168, 2018.

[50] J. Howe, "The rise of crowdsourcing," *Wired magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[51] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in *Proc. Elev. Int. Conf. Web Soc. Media*, ser. ICWSM-2017. Montreal, Canada: AAAI Press, 2017, pp. 512–515.

[52] R. Lukyanenko, J. Parsons, Y. Wiersma, G. Wachinger, B. Huber, and R. Meldt, "Representing crowd knowledge: Guidelines for conceptual modeling of user-generated content," *Journal of the AIS*, vol. 18, no. 4, pp. 297–339, 2017.

[53] D. H. Wolpert and W. G. Macready, "No Free Lunch Theorems for Optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, 1997.

[54] P. Burnap and M. L. Williams, "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015.

[55] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate Speech Detection with Comment Embeddings," in *Proc. 24th Int. Conf. World Wide Web*, ser. WWW '15 Companion. Florence, Italy: ACM Press, 2015, pp. 29–30.

[56] B. van Aken, J. Risch, R. Krestel, and A. Löser, "Challenges for Toxic Comment Classification: An In-Depth Error Analysis," in *Proc. Second Work. Abus. Lang. Online*, ser. ALW2, D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont, Eds. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 33–42.

[57] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in *Soc. Mob. Web, Pap. from 2011 ICWSM Work.*, ser. ICWSM 2011. Barcelona, Spain: Association for the Advancement of Artificial Intelligence, 2011, pp. 11–17.

[58] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," in *Proc. 2012 ASE/IEEE Int. Conf. Soc. Comput. 2012 ASE/IEEE Int. Conf. Privacy, Secur. Risk Trust*, ser. SOCIALCOM-PASSAT '12. Amsterdam, Netherlands: IEEE, 2012, pp. 71–80.

[59] K. Reynolds, A. Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyberbullying," in *Proc. 10th Int. Conf. Mach. Learn. Appl. Work.*, ser. ICMLA'11, X.-w. Chen, T. Dillon, H. Ishbuchi, J. Pei, H. Wang, and M. A. Wani, Eds. Honolulu, Hawaii, USA: IEEE, 2011, pp. 241–244.

[60] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean Birds: Detecting Aggression and Bullying on Twitter," in *Proc. 2017 ACM Web Sci. Conf.*, ser. WebSci '17. Troy, New York, USA: ACM Press, 2017, pp. 13–22.

[61] P. Mathur, R. Sawhney, M. Ayyar, and R. R. Shah, "Did you offend me? Classification of Offensive Tweets in Hinglish Language," in *Proc. Second Work. Abus. Lang. Online*, ser. ALW2, D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont, Eds. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 138–148.

[62] Y. Lee, S. Yoon, and K. Jung, "Comparative Studies of Detecting Abusive Language on Twitter," in *Proc. Second Work. Abus. Lang. Online*, ser. ALW2, D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont, Eds. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 101–106.

[63] S. O. Sood, J. Antin, and E. F. Churchill, "Using Crowdsourcing to Improve Profanity Detection," in *AAAI Spring Symp. Ser.*, Palo Alto, CA, USA, 2012, pp. 69–74.

[64] D. R. Cox, "The Regression Analysis of Binary Sequences," *J. R. Stat. Soc. Ser. B*, vol. 20, no. 2, pp. 215–242, 1958.

[65] L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone, *Classification And Regression Trees*. Monterey, CA, USA: Chapman & Hall/CRC, 1984.

[66] M. Sahlgren, T. Isbister, and F. Olsson, "Learning Representations for Detecting Abusive Language," in *Proc. Second Work. Abus. Lang. Online*, ser. ALW2, D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont, Eds. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 115–123.

[67] Y. Mehdad and J. Tetreault, "Do Characters Abuse More Than Words?" in *Proc. 17th Annu. Meet. Spec. Interes. Gr. Discourse Dialogue*, ser. SIGDIAL 2016. Los Angeles, CA, USA: Association for Computational Linguistics, 2016, pp. 299–303.

[68] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "Deep Learning for User Comment Moderation," in *Proc. First Work. Abus. Lang. Online*, ser. ALW1, Z. Waseem, W. H. K. Chung, D. Hovy, and J. Tetreault, Eds. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 25–35.

[69] J. Serrà, I. Leontiadis, D. Spathis, G. Stringhini, and J. Blackburn, "Class-based Prediction Errors to Categorize Text with Out-of-

vocabulary Words," in *Proc. First Work. Abus. Lang. Online*, ser. ALW1, Z. Waseem, W. H. K. Chung, D. Hovy, and J. Tetreault, Eds. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 36–40.

[70] A. Švec, M. Pikuliak, M. Šimko, and M. Bieliková, "Improving Moderation of Online Discussions via Interpretable Neural Models," in *Proc. Second Work. Abus. Lang. Online*, ser. ALW2, D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont, Eds. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 60–65.

[71] J. H. Park and P. Fung, "One-step and Two-step Classification for Abusive Language Detection on Twitter," in *Proc. First Work. Abus. Lang. Online*, ser. ALW1, Z. Waseem, W. H. K. Chung, D. Hovy, and J. Tetreault, Eds. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 41–45.

[72] H. Chen, S. Mckeever, and S. J. Delany, "Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media," in *Adv. Comput. Intell. Syst. Contrib. Present. 16th UK Work. Comput. Intell. Sept. 7-9, 2016, Lancaster, UK*, P. Angelov, A. Gegov, C. Jayne, and Q. Shen, Eds. Cham: Springer, 2017, pp. 187–205.

[73] O. Luaces, J. Díez, J. Barranquero, J. J. del Coz, and A. Bahamonde, "Binary relevance efficacy for multilabel classification," *Prog. Artif. Intell.*, vol. 1, no. 4, pp. 303–313, 2012.

[74] C. Wang, "Interpreting Neural Network Hate Speech Classifiers," in *Proc. Second Work. Abus. Lang. Online*, ser. ALW2, D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont, Eds. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 86–92.

[75] M. Feurer, A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum, and F. Hutter, "Efficient and Robust Automated Machine Learning," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, ser. NIPS'15, Montreal, Canada, 2015, pp. 2755–2763.

[76] ——, "Auto-sklearn: Efficient and Robust Automated Machine Learning," in *Autom. Mach. Learn. Methods, Syst. Challenges*, ser. The Springer Series on Challenges in Machine Learning. Springer, Cham, 2019, pp. 113–134.

[77] S. Gregor and I. Benbasat, "Explanations from intelligent systems: Theoretical foundations and implications for practice," *MIS Quarterly*, vol. 23, no. 4, pp. 497–530, 1999.

[78] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"," *AI Mag.*, vol. 38, no. 3, p. 50, oct 2017.

[79] K. R. Fleischmann and W. A. Wallace, "A covenant with transparency: Opening the black box of models," *Communications of the ACM*, vol. 48, no. 5, pp. 93–97, 2005.

[80] P. Owotoki and F. Mayer-Lindenberg, "Transparency of Computational Intelligence Models," in *Proc. AI-2006, Twenty-sixth SGAI Int. Conf. Innov. Tech. Appl. Artif. Intell.*, ser. AI-2006, M. Bramer, F. Coenen, and A. Tuson, Eds. Cambridge, UK: Springer London, 2007, pp. 387–392.

[81] H. Cramer, B. Wielinga, S. Ramlal, V. Evers, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga, "The effects of transparency on perceived and actual competence of a content-based recommender," in *Semantic Web User Interaction workshop at CHI, Florence, Italy*, 2008.

[82] R. Sinha and K. Swearingen, "The role of transparency in recommender systems," in *CHI'02 extended abstracts on Human factors in computing systems*. ACM, 2002, pp. 830–831.

[83] D. H. McKnight, V. Choudhury, and C. Kacmar, "The impact of initial consumer trust on intentions to transact with a web site: a trust building model," *The Journal of Strategic Information Systems*, vol. 11, no. 3, pp. 297–323, 2002. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0963868702000203

[84] Gefen, Karahanna, and Straub, "Trust and TAM in online shopping: An integrated model," *MIS Quarterly*, vol. 27, no. 1, pp. 51–90, 2003. [Online]. Available: http://www.jstor.org/stable/10.2307/30036519

[85] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, vol. 1, pp. 1–10, 10 2017.

[86] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweet," in *Proc. 26th Int. Conf. World Wide Web Companion*, ser. WWW '17 Companion. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 759–760.

[87] A. Immonen, M. Palviainen, and E. Ovaska, "Requirements of an Open Data Based Business Ecosystem," *IEEE Access*, vol. 2, pp. 88–103, 2014.

[88] D. Naous, J. Schwarz, and C. Legner, "Analytics As A Service: Cloud Computing and the Transformation of Business Analytics Business Models and Ecosystems," in *Proc. 25th Eur. Conf. Inf. Syst.*, ser. ECIS 2017. Guimarães, Portugal: AIS, 2017.

[89] O. Ylijoki, J. Sirkiä, J. Porras, and V. Harmaakorpi, "Innovation capabilities as a mediator between big data and business model," *J. Enterp. Transform.*, pp. 1–18, 2019.