Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1382–1390
Marseille, 11–16 May 2020

WAC: A Corpus of Wikipedia Conversations for Online Abuse Detection

Noé Cécillon, Vincent Labatut, Richard Dufour and Georges Linarès
Laboratoire Informatique d'Avignon – LIA EA 4128, Avignon Université, France
{firstname.lastname}@univ-avignon.fr

Abstract

With the spread of online social networks, it is more and more difficult to monitor all the user-generated content. Automating the moderation process of the inappropriate exchange content on Internet has thus become a priority task. Methods have been proposed for this purpose, but it can be challenging to find a suitable dataset to train and develop them. This issue is especially true for approaches based on information derived from the structure and the dynamic of the conversation. In this work, we propose an original framework, based on the Wikipedia Comment corpus, with comment-level abuse annotations of different types. The major contribution concerns the reconstruction of conversations, by comparison to existing corpora, which focus only on isolated messages (*i.e.* taken out of their conversational context). This large corpus of more than 380k annotated messages opens perspectives for online abuse detection and especially for context-based approaches. We also propose, in addition to this corpus, a complete benchmarking platform to stimulate and fairly compare scientific works around the problem of content abuse detection, trying to avoid the recurring problem of result replication. Finally, we apply two classification methods to our dataset to demonstrate its potential.

Keywords: Wikipedia conversations, Abuse detection, Evaluation framework, Automatic moderation

1. Introduction

The ever growing quantity of content posted online requires more and more moderators to monitor this content. A fast and accurate moderation is highly beneficial to online platforms, but it is increasingly expensive and difficult to maintain. Therefore, the automated detection of abusive online content is an important research topic. Corpora allowing to develop such methods often focus on *single comments* without any conversational context (Pavlopoulos et al., 2017; Razavi et al., 2010). Yet, recent works (Papegnies et al., 2019; Yin et al., 2009) suggest that considering the *entire conversation* thread might improve the automatic detection of abusive content. However, the development of such methods is currently limited by the lack of large scale corpora of conversations. Corpora containing full conversations exists but they have limited number of messages or are not publicly available (Napoles et al., 2017; Cécillon et al., 2019). Karan and Šnajder (2019) offers a solution with *PreTox*, a large corpus of discussion threads from Wikipedia talk pages. However, the quality of their semi-automatically generated annotations might be problematic, as the authors report a Precision of only 51%.

In this paper, we reconstruct a large scale corpus of messages from English Wikipedia talk pages, structured as full conversations and annotated with high quality annotations. This results in a corpus containing roughly 193k conversations and 383k messages annotated as being abusive or not. To encourage further development of context- and thread-based methods in the area of abusive content detection, we publish and make freely available this corpus and the code source used for its extraction. The other objective of this work is to improve replicability and ease the comparison of classification methods. In this context, we introduce an open-source benchmarking platform that we developed.

The contribution of this work is threefold. First, we match two existing corpora of Wikipedia messages and develop a pipeline to create a large publicly available corpus of conversations. Messages are provided together with detailed information such as the message type, author, talk page and high quality annotations. Second, we present a common comparison platform grouping approaches and methods to stimulate communities around automatic detection of abusive content. Third, we illustrate the interest of our corpus and platform by assessing existing abuse detection methods. The rest of this article is organized as follows. First, in Section 2., we describe the existing corpora related to our proposed one, and how they are used in the literature. Then, we describe our corpus in Section 3., as well as the reconstruction pipeline we propose for its constitution. We present a benchmarking platform and some results we obtain on our corpus in Section 4.. Finally, in Section 5., we summarize our results and present some perspectives.

2. Related Work

In this section, we introduce the corpora of Wikipedia messages related to abuse detection. We review how they are used in the literature, and stress the limitations of these corpora as well as the works leveraging them.

2.1. Wikipedia Talk Pages

A *talk page* is a discussion page where users can argue and discuss topics relative to a specific Wikipedia page. Every Wikipedia user and article has a related talk page, identified by a unique `page_id`. But Wikipedia does not propose a standard post system such as those commonly used in online forums. Instead, the talk page is similar to a regular Wikipedia article page, or a wiki page in general: in theory, users have the ability to edit it by adding, modifying or removing text anywhere. However, in practice, a set of writing and formatting conventions[1] allow giving structure to the various conversations taking place on the talk page. For instance, when a user adds his own post, he indents it so as to indicate its hierarchical level in the conversation tree. Figure 1 shows an example of Wikipedia conversation under

[1] `https://en.wikipedia.org/wiki/Help:Talk_pages#Replying_to_an_existing_thread`

the form of the rendered talk page and the corresponding Wikicode (Wikipedia markup language). Note that a talk page generally contains several conversations at once.

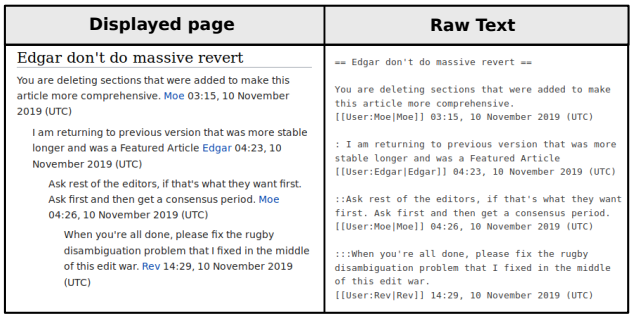

Figure 1: Part of the *Japan* Wikipedia article talk page: rendered page (left) and corresponding Wikicode (right).

Like for article pages, Wikipedia stores the changes corresponding to each edit, as a *revision* entry containing the text of the page after the edit. Each revision is identified by a unique number called `rev_id`.

2.2. Wikipedia Comment Corpus

As part of the Wikipedia Detox research project, Wulczyn et al. (2017) proposed the *Wikipedia Comment Corpus* (WCC), a corpus of discussion comments from English Wikipedia talk pages. These comments are extracted using the revision history of each considered talk page. The authors consider the textual differences between two consecutive revisions of the talk page, and distinguish two cases depending on the importance of these changes. If the modification is significant, they assume a new comment was posted, which is identified by its own `rev_id`. Otherwise, they suppose an existing comment was modified, and apply these changes without updating its `rev_id`. Therefore, a given `rev_id` is associated to *at most* one comment, and one comment to *exactly* one `rev_id`. However, it is important to note that one large edition can correspond, in practice, to a user writing *several* new posts in distinct conversations of the same page. In this case, all these posts are mistakenly gathered in a single WCC comment.

Wulczyn et al. (2017) used a public dump of the English Wikipedia full history made available in January 2016 to create their corpus, which contains more than 63M comments posted between 2004 and 2015. From this massive corpus, they sampled 3 smaller datasets that they annotated for different types of abuse:

- *personal attack*: abusive content directed at somebody's person rather than providing evidence;
- *aggression*: malicious remark to a person or group on characteristics such as religion, nationality or gender;
- *toxicity*: comment that can make other people want to leave the conversation.

It is important to note that each comment in the three datasets is *explicitly* annotated as abusive or not. By comparison, in the abuse detection literature, datasets are often annotated by considering comments flagged by moderators as abusive, whereas the rest of the comments are deemed non-abusive *by default*, without further check. It is then possible for the non-abusive label to be assigned to *abusive* comments just because they were missed by the human moderators *e.g.* (Papegnies et al., 2017; Delort et al., 2011; Karan and Šnajder, 2019). Having explicitly annotated non-abusive comments makes WCC a more reliable corpus, on this aspect.

Information on the datasets is summarized in Table 1. The *Personal attack* and *Aggression* datasets contain exactly the same 115k comments while the *Toxicity* dataset contains more comments (159k). Among them, 77k appear in all three datasets. The prevalence of abusive comments in the *Personal attack* dataset is 13.4%, 14.7% in the *Aggression* dataset, and 11.5% in the *Toxicity* dataset. This prevalence does not reflect the data though, as Wulczyn et al. (2017) oversampled comments from blocked users to enhance the variety of abusive comments. The original abuse rate in Wikipedia comments is around 1%.

Dataset	Comments	Percentage abusive	Type of annotation
Personal attack	115,864	13.4 %	binary
Aggression	115,864	14.7 %	binary and numerical
Toxicity	159,686	11.5 %	binary and numerical

Table 1: Main properties of the three datasets constituting the Wikipedia Comment Corpus (WCC).

As the *Wikipedia Comment Corpus* (WCC) is one of the largest available human annotated comments corpus, it is used in many works. Wulczyn et al. (2017) themselves tackle the problem of detecting personal attacks in Wikipedia comments. They experiment with logistic regression and multi-layer Perceptron classifiers using word or character n-gram features and report a 96.59 AUC score on the *Personal attack* dataset. Pavlopoulos et al. (2017) apply deep learning methods to the moderation task. They experiment with various methods such as Convolutional Neural Network (CNN) operating on word embeddings, Recurrent Neural Network (RNN) and several variants of RNN using an attention mechanism. Their results on the *Personal attack* and *Toxicity* datasets outperform previously reported results with an AUC score up to 98.42 on the *Toxicity* dataset. Gröndahl et al. (2018) propose a comparative analysis of state-of-the-art hate speech detection models and apply them to the *Personal attack* dataset. They experiment with Logistic Regression (LR) and Multi-Layer Perceptrons (MLP) operating on character n-grams, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) approach. The report results ranging from 85 to 87% in terms of macro-averaged $F1$-score. Mishra et al. (2018) propose various methods to detect abusive content in the *Personal attack* and achieve their best results with a method using context-aware representations for characters. This method is referred to as the *Context hidden-state + char n-grams* method, and obtains a 89.35 and 87.44 macro-averaged $F1$-score on the *Toxicity* and *Personal attack* datasets, respectively.

Furthermore, Dixon et al. (2018) show that classifiers can have unfair biases toward certain people and propose meth-

ods to mitigate unintended bias in text classification models. They illustrate this statement by applying their methods to the *Toxicity* and *Personal attack* datasets.

Table 2 summarizes the performances reported in the previously cited articles. The second column refers to the datasets used to train and test the classifier. Additionally, we can see that the *Aggression* dataset is not used by any of the listed methods. The third column indicates, when available, the percentage of comments in each of the train/development/test split.

Although the classification performances reported are quite good, most of these models can be fooled by basic obfuscation or adversarial methods. Hosseini et al. (2017) demonstrate the efficiency of such an attack against the *Google Perspective* API. Gröndahl et al. (2018) present some basic, but efficient, evasion methods. Most of them induce a significant decrease in the classification performances. The word-based are the most vulnerable ones, which can be completely fooled by introducing or removing manual typos, punctuation and spaces in comments. Because of this possible vulnerability, it can be interesting to rely on more information than only the textual content of each comment.

2.3. WikiConv

WikiConv (Hua et al., 2018) is a large public corpus based on Wikipedia talk pages extracted from a July 2018 Wikipedia dump. This corpus contains *full conversations*, and not only *isolated comments* like WCC. In this corpus, we call "messages" the textual elements constituting a conversation. The structure of a conversation is retrieved by considering the revision history of the talk page containing it. This history is viewed as a sequence of *conversational actions*. A conversational action is an object representing one operation performed by a user on a talk page. It is composed of many attributes about the action, the talk page, the conversation and its structure. Additionally, actions are categorized into 5 types: conversation thread *creation*, new message *addition*, existing message *modification*, message *deletion*, and deleted message *restoration*. All the attributes are listed and described on the WikiConv authors' GitHub repository[2]. As mentioned before, when performing these actions, the Wikipedia users respect a set of formatting conventions. Hua *et al*. define a heuristic leveraging this knowledge to identify actions, retrieve their description, and determine their type. This pipeline only relies on visual markup clues, so it is language-independent and can be applied to any version of Wikipedia archives, as long as the formatting conventions stay the same. The largest component of the corpus is from the English Wikipedia but the pipeline is also applied to Chinese, Russian, Greek and German Wikipedia.

A WikiConv message is the textual content associated to an action, i.e. the text that was added, removed, or edited. A single revision of a talk page can be constituted of several actions, and therefore result in several WikiConv messages. Because of that, the `rev_id`, which is used as a unique comment identifier in WCC, can be shared by multiple actions in WikiConv. Instead, a WikiConv message is uniquely identified by an `action_id`. Another major difference with WCC is that WikiConv contains the full history of the conversation, with each successive version of a message in case it is edited, and not only its final form. Moreover, when several new posts are added in one revision, those are not merged in a single comment as in WCC, but represented by separate WikiConv messages. It is therefore quite common that multiple WikiConv messages correspond to the same WCC comment. Figure 2 illustrates a typical example where the added text is split over two different levels of indentation. In WCC, all the text is concatenated into a single comment while in WikiConv, 2 messages (and therefore actions) are created, each corresponding to a different indentation level. The 2 actions are distinct (different `action_id`) but they have the same `rev_id`. Finally, the most important difference with WCC is that the messages are not annotated for abuses.

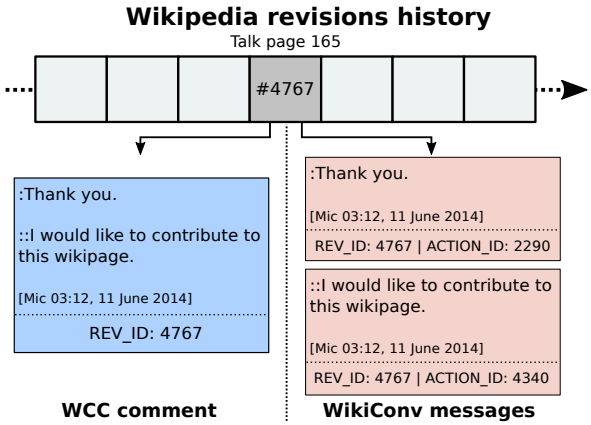

Figure 2: Illustration of the issue regarding the `rev_id` between the WCC and WikiConv. Figure available at 10.6084/m9.figshare.11302385 under CC-BY license.

Moreover, Hua et al. (2018) use the Google Perspective API[3] to score the toxicity of the messages in WikiConv. Based on machine learning models, this API scores messages for several types of abuse, *e.g.*, toxicity, profanity, threat, insult. In WikiConv, all messages are scored on their toxicity and severe toxicity. These 2 are provided as attributes of each message in the corpus.

This corpus is quite recent, so only few researches currently use it. We report one published paper by Chang and Danescu-Niculescu-Mizil (2019), that explores the potential of a small subset of conversations sourced from WikiConv to predict future derailment in online conversations. In the next section, we introduce a corpus extending the work of Hua et al. (2018) on WikiConv.

2.4. PreTox

In a recent work, Karan and Šnajder (2019) proposed *PreTox*, a corpus based on WikiConv (see previous section). *PreTox* is composed of complete discussion threads with semi-automatically generated toxicity annotations. Karan *et al.* rely on a heuristic to flag toxic messages. This heuristic combines two types of information: 1) whether or not the message was deleted by someone else; and 2) the scores generated using the Google Perspective API and provided

[2] https://github.com/conversationai/wikidetox/tree/master/wikiconv

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

Method	Dataset	Split	Metric	Score
Logistic Regression (Wulczyn et al., 2017)	Personal attack	60/20/20 *	AUC	96.24
Multi-layer perceptrons (Wulczyn et al., 2017)	Personal attack	60/20/20 *	AUC	96.59
RNN (Pavlopoulos et al., 2017)	Toxicity	N/A	AUC	98.42
RNN + attention mechanism (Pavlopoulos et al., 2017)	Personal attack	N/A	AUC	97.46
RNN + attention mechanism (Pavlopoulos et al., 2017)	Toxicity/Personal attack	N/A	AUC	98.22
LSTM (Gröndahl et al., 2018)	Personal attack	N/A	F1-score	85.
CNN + GRU (Gröndahl et al., 2018)	Personal attack	N/A	F1-score	87.
Context hidden-state+char n-grams (Mishra et al., 2018)	Personal attack	60/40 +	F1-score	87.44
Context hidden-state+char n-grams (Mishra et al., 2018)	Toxicity	60/40 +	F1-score	89.35

Table 2: Works leveraging the WCC, with the obtained performances. The *Split* column corresponds to the percentage of comments in the train/test or train/development/test sets. Symbols (*,$^+$) denotes a similar split used by several methods.

along with the WikiConv corpus. A binary toxicity annotation is created for each message using this heuristic. Flagged messages are deemed toxic and, unlike WCC, all remaining messages are considered as non-toxic. Karan and Šnajder (2019) report an annotation Precision of 51% for their semi-automated method on a test set of 100 manually annotated messages. Thus, we can suppose that human annotations would surely be more accurate than the semi-automatically generated annotations of *PreTox*.

2.5. Discussion

In this section, we reviewed the corpora related to the detection of abusive messages in Wikipedia talk pages. However, they all have some weaknesses. WCC contains high quality annotations for 3 types of abusive content, but does not provide any conversational structure. On the contrary, WikiConv provides full conversations but without annotations. PreTox seeks to extend the latter by semi-automatically annotating the messages, but this process is not accurate enough. In the next section, we address these issues by combining WCC and WikiConv to combine their advantages, and thus compensate for their individual drawbacks.
Moreover, we also reviewed the works leveraging these corpora, and highlighted a major issue, as shown in Table 2: the lack of a standard protocol for evaluating the performances of abuse detection tools. This flaw concerns both the evaluation metric used and the way the data is divided into train/development/test subsets. In addition, none of the listed works give an open source version of their code. Therefore, it is extremely complicated to have a comparative overview of all proposed approaches, which certainly constitutes a major obstacle to progress in this research area.

3. Proposed Corpus

The context of messages (*i.e.* the messages surrounding a targeted message in a conversation) is ignored by many existing abusive content detection methods (Pavlopoulos et al., 2017; Razavi et al., 2010; Djuric et al., 2015), while it seems to have a positive influence on classification performances (Papegnies et al., 2019; Yin et al., 2009). An annotated conversation corpus could allow the use of such information at a large scale, in order to develop context-based methods that takes advantage of conversational structure and dynamics to detect abusive content.
In this work we propose *Wikipedia Abusive Conversations* (WAC), a corpus of messages from Wikipedia integrating conversational information and high quality human annotations. WAC is a combination of the first two corpora described in Section 2. and takes advantage of their complementarity. It is based on the messages and conversations structure from WikiConv (Hua et al., 2018) and the human annotations for 3 different types of abusive content from the WCC (Wulczyn et al., 2017). The textual elements constituting conversations in WAC are called "messages" like WikiConv as they correspond to WikiConv messages matched with a WCC annotation. This reconstruction task is not trivial because of the way that comments and messages are identified in the two corpora. As explained before and shown on Figure 2, there is no guarantee of `rev_id` uniqueness for WikiConv messages, making it difficult to match them with WCC comments. WAC provides a large collection of conversations including at least one human-annotated message per conversation. It is divided into 3 datasets annotated for *Personal attack*, *Aggression* and *Toxicity*. In total, it contains approximately 193k conversations consisting of 4.9 million messages, among which 383k are annotated. It is publicly available online[4].

3.1. Reconstruction Pipeline

We now describe the reconstruction process we developed in order to gather information from existing corpora (Wulczyn et al., 2017; Hua et al., 2018) in a new one and extract useful information. The pipeline is detailed only for the *Personal attack* dataset, but is the same for the other two datasets. Its source code is open source, and publicly available online[5]. The reconstruction process is divided into 5 main steps. It begins with the extraction of the annotation from WCC. The second step is to retrieve messages from WikiConv. The third step consists in filtering these messages in order to keep only the relevant talk pages. The fourth step is the conversation reconstruction. The last step, the most important and difficult one, consists in uniquely identifying all the annotated messages in the conversation. Figure 3 shows the whole pipeline, discussed through this section.

3.1.1. Annotation Extraction

The first step is to extract annotations from the WCC. This corpus provides 10 judgments per annotated comment. Each judgement provides multiple annotations depending on the dataset. The *Personal attack* dataset has

[4] DOI: `10.6084/m9.figshare.11299118`
[5] https://github.com/CompNet/WikiSynch

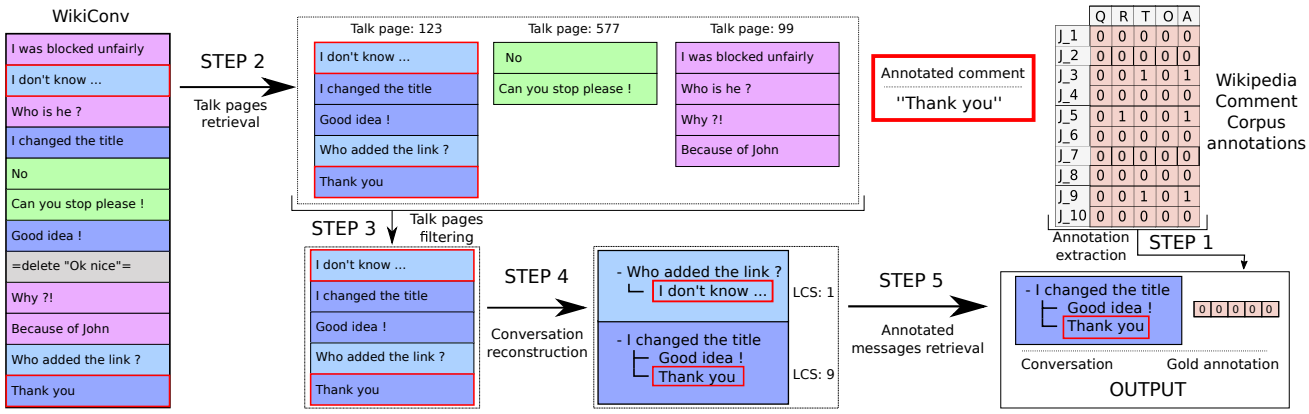

Figure 3: Representation of our pipeline applied to reconstruct the conversation of an annotated comment. Only the textual content of the actions is displayed. The right side shows the annotated comment and its 10 associated human judgments (J_1, ..., J_10). The letters in the WCC table stands for *quoting_attack* (Q), *recipient_attack* (R), *third_party_attack* (T), *other_attack* (O), *attack* (A). Messages with a red frame have the same *rev_id* as the annotated comment. Message colors match both the pages and the conversation containing them. Figure available at 10.6084/m9.figshare.11302385 under CC-BY license.

5 binary annotations: *quoting_attack, recipient_attack, third_party_attack, other_attack* and the more general *attack*. The *Aggression* and *Toxicity* datasets also provides such a general binary score (*aggression* and *toxicity*, respectively). Additionally, they provide an *aggression_score* and a *toxicity_score*) ranging from -2 (very abusive) to 2 (very healthy), 0 being neutral. We aggregate these 10 judgments to determine the gold annotation of all the annotated messages. For the binary annotations, we compute the majority annotation among crowdworkers to determine the gold standard. For the scores, we compute the average value among all crowdworkers. In WAC, we call *annotated messages* the annotated comments extracted from WCC. The right side of Figure 3 shows an example of this step applied to a comment annotated for *Personal attack*. Based on the 10 human judgments from WCC, the gold annotations for each of the 5 types of attack annotated is determined.

3.1.2. Talk Pages Retrieval

The second step is to retrieve the data from WikiConv[6]. Because WCC contains data from the English Wikipedia, we only consider the English part of WikiConv, which is composed of approximately 91M distinct conversations. We group all messages by their respective `page_id`. Note that at this stage, messages are not ordered, and not structured as conversations, as a single talk page can contain multiple conversations. This is for instance the case for the blue talk page in Figure 3. At this step, we also filter the messages based on their type: creations, additions, and modifications are retained while deletions and restorations are filtered out. Indeed, deletions often concern abusive messages which are already outnumbered by non-abusive messages in WCC, so considering that some messages are deleted would unbalance the corpus even more. By doing this filtering, we also retain a maximum of annotated messages in our corpus.
In the example displayed in Figure 3, the colors of the WikiConv messages match the talk page on which they appear. We can distinguish 4 pages: purple, blue, green and grey. However, the grey page contains only 1 message which corresponds to a deletion. Thus, this message is removed and after Step 2, only 3 talk pages remain.

3.1.3. Talk Pages Filtering

Among the WikiConv talk pages retrieved at the previous step, only a fraction contains a message that is annotated in WCC. This filtering step aims at keeping only the pages containing at least one such message, in order to retain only the relevant talk pages. However, it is important to understand that the annotated message can be in any of the conversations taking place on the concerned talk page. In order to perform such a filtering, we rely on the `rev_id`, the id of the revision from which the message was extracted. The retained pages are all the pages whose at least one message has the same `rev_id` as an annotated comment. As previously mentioned, this attribute is available in both WCC and WikiConv but unlike WCC, WikiConv is likely to associate the same `rev_id` to several distinct messages (or rather, these messages correspond to a single comment in the WCC). Therefore, there are more messages in WikiConv having the `rev_id` of an annotated WCC comment than the actual number of annotated comments. This issue is addressed later at Step 5. Messages with a red frame in Figure 3 are messages having the same `rev_id` as the annotated comment. After Step 3, only the blue page is retained, as it is the only one containing messages with the wanted `rev_id`. Messages in this page are still unordered.

3.1.4. Conversation Reconstruction

The fourth step consists in reconstructing the conversation. For each page remaining after the previous step, we reconstruct the distinct conversations taking place on this specific page, using the attributes available in WikiConv. The reconstruction process starts by retrieving all the messages corresponding to the creation of a new conversation. On Figure 3, there are two such messages out of the five messages of this page. To retrieve all the creation of conversations, the `type` attribute is not enough because some messages being the starting point of a new conversation are

[6] DOI: `10.6084/m9.figshare.7376003`

categorized as addition and not as creation. Then, based on the `replyTo_id` allowing to find the message to which this message answers, it is possible to link each message of the conversation and so, to reconstruct its structure. As a result, the structure of each conversation on the page is modeled as a graph of actions, a message being the textual content of an action. Figure 4 is an example of a conversation reconstructed during this step. The left part shows the textual content of all the actions in this conversation. The very first action is the creation of the conversation. Then, all the source-reply relationships are modeled using tabulations. An action is a reply to the nearest previous action with one less tabulation. For instance, Actions 7 and 3 are replies to Action 2 which is itself a reply to Action 1. The right part of Figure 4 shows the corresponding graph of actions.

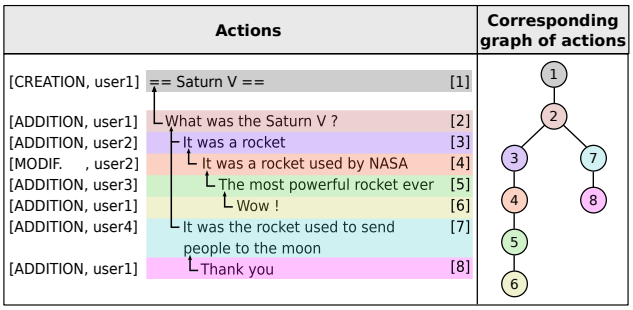

Figure 4: An example of a conversation and its corresponding graph of actions. Arrows illustrate which action is replying to. Figure available at 10.6084/m9.figshare.11302385 under CC-BY license.

During this reconstruction step, modification messages are considered as replies to the original message they are editing in order to keep track of all the content added to the talk page and not only the final form of each message. Moreover, a lot of messages categorized as modifications can actually be considered as additions. Indeed, a typical behavior of Wikipedia users is to reply to a message by adding text straight into the message they want to reply to instead of creating a new one. Thus, some conversations take place in a single message which is modified successively by multiple users. However, even if the conversation takes place in a single message visually, technically, an action is created and saved for each successive edition of the message. So, the graph of action that we produce is exactly the same as if all the messages were posted in successive and distinct messages. This behavior justifies the need to consider modifications as full messages, and not as a simple state of a message at a given time.

In the example of Figure 3, the page retained at the previous step contains 2 conversations modeled by two distinct shades of blue. These conversations are reconstructed at Step 4 and actions are ordered as they appear on the original talk page.

3.1.5. Annotated Messages Retrieval

We now have reconstructed all the conversations appearing in pages known to contain at least one annotated message. However, some of these conversations may not contain any such message, as several conversations generally coexist on the same talk page. The last step is therefore to filter them out. As stated in Step 3, a given `rev_id` can be associated to several WikiConv message, whereas it points out at a unique WCC comment. This is a major issue for us because the `rev_id` is the only attribute available to match WCC comments to WikiConv messages. In order to figure out which of the messages with equivalent `rev_id` is actually the comment annotated in WCC, we compute the *Longest Common Sequence* (LCS) between the original annotated comment and each message in our corpus having the same `rev_id`. We consider that the message with the LCS corresponds to the annotated comment. Approximately 36% of our annotated messages are concerned by this issue, most of them having only 2 or 3 messages with similar `rev_id`. Once every annotated message has been uniquely identified, we filter all the conversations reconstructed at Step 4 to only keep those containing at least one annotated message. In the example of Figure 3, both conversations reconstructed at Step 4 contain a message with the `rev_id` of the annotated comment, the messages with a red frame. The LCS is computed between both conversation messages and the annotated comment. As a result, the message "Thank you" is identified as the actual annotated message and its conversation is retained while the other is discarded. In the end, we get the conversation containing the annotated message and its associated gold annotation computed from WCC.

The described pipeline is applied for all annotations types (*i.e., Personal attack, Aggression and Toxicity*) to create the 3 distinct datasets constituting our WAC corpus. It is composed of conversations containing at least one annotated message. Three files containing the annotations are released along with the corpus, each file corresponding to a dataset.

3.2. Description

A number of annotated comments from the original WCC datasets are discarded during the reconstruction pipeline described in Section 3.1.. This is mostly due to missing data in WikiConv or WCC. Some lost comments also are comments associated to a deletion or restoration in WikiConv which are discarded in the reconstruction pipeline. However, 97.97% of the original annotated WCC comments are retained in WAC. In total, the corpus contains more than 2.2 million unique messages split into 168,827 unique conversations. The number of annotated messages and the division of annotations in all three datasets is summarized in Table 3. As mentioned in Section 2.2., one annotated message can be annotated for different types of abuse. Hence the 382,665 total annotations from the last line of Table 3 are assigned to a total of 193,265 distinct messages.

Wikipedia messages are usually longer than other types of online posts such as tweets or chat messages. Messages in our corpus have an average length of more than 1,000 characters. As shown in Figure 4, the structure of each conversation can be modeled as a graph. On average, there are 13 messages in a conversation. The distribution of the conversations length is shown in Figure 5. Note that the y-axis scale is logarithmic for readability reasons. We can observe that some conversations contain more than 1,000 messages but for a large majority, conversations are only 1- to 20-message long.

Figure 6 shows the distribution of the relative position of

Dataset	Annotated messages	Abuse	Non-abuse
Personal attack	113,174	14,934 (13.20%)	98,240 (86.80%)
Aggression	113,174	16,331 (14.43%)	96,843 (85.57%)
Toxicity	156,317	19,700 (12.60%)	136,617 (87.40%)
Total	382,665	50,965 (13.31%)	331,700 (86.69%)

Table 3: Number of annotated messages and distribution of annotations in the proposed *Wikipedia Abusive Conversations* (WAC) corpus.

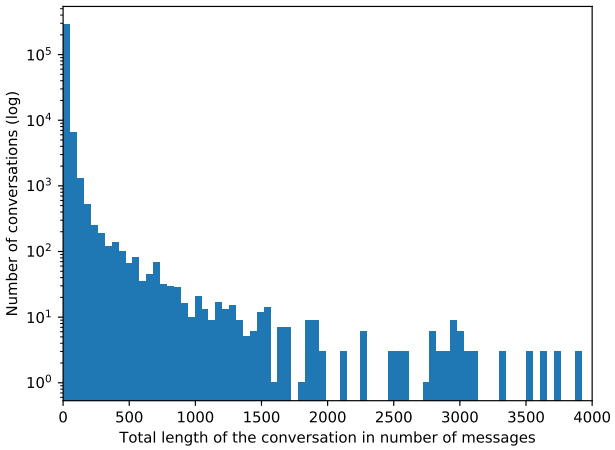

Figure 5: Distribution of the conversation lengths in *Wikipedia abusive Conversations*, expressed in number of messages. The y-axis scale is logarithmic. Figure available at 10.6084/m9.figshare.11302385 under CC-BY license.

annotated messages in conversations. This position is expressed as the percentage of messages posted *before* the annotated message in the conversation. Only conversations with at least 5 messages are considered in this figure. This distribution shows that annotated messages are well distributed over all positions in the conversations, except at the very end of the conversation where a lot more annotated messages appear. This observation holds whether the annotated message is abusive or not. For abusive messages, this position can potentially be explained by the fact that abusive comments are quickly deleted from Wikipedia (Hua et al., 2018), before creating many reactions.

Conversations can also be divided into multiple sub-conversations, a sub-conversation being a sequence of messages in which each message is an answer to the previous. For instance, the conversation represented in Figure 4 is composed of 8 messages and 2 sub-conversations. On average, there are 3 sub-conversations per conversation in the corpus. However, they only contain approximately 4 messages in average, which is quite limited for a conversation. This value highlights the fact that Wikipedia talk pages are not used in the same way as forums or social media. Indeed, many messages are informative messages explaining what changes have been made to the article or suggestions on how to edit the article associated to the talk page. Most

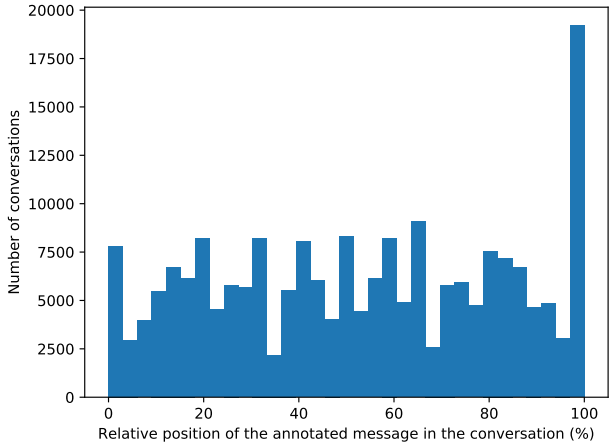

Figure 6: Position of the annotated messages in the conversation, expressed in percentage of the messages appearing *before* the annotated message. Only conversations with 5 and more messages are considered. Figure available at 10.6084/m9.figshare.11302385 under CC-BY license.

of the time, these messages do not imply answers, which can explain the relatively small number of messages per sub-conversation that we observe.

4. Proposed Benchmarking Platform

In Section 2.5., we highlighted some issues with the evaluation and comparison of current abuse detection methods. To overcome this problem, we propose a common benchmark platform, described in Section 4.1.. Then, we assess, in Section 4.2., existing detection methods to illustrate the interest of our corpus and platform.

4.1. Platform Description

An important issue is that the reported performance in different works are often almost impossible to compare since systems may be evaluated on different datasets, and many different metrics are used to perform these evaluations. Even if the used corpus is identical, which is the case with works on WCC for example, the way it is split into train/development/test differs, and is often not precisely described by the authors.

All these issues hinder replicability. In this section, we present a benchmarking platform that we developed in order to address them. It is an open-source tool available online[7], and aimed at grouping classification methods in the area of automatic abusive content detection to ease and stimulate the replicability of the reported performances. Moreover, we take advantage of our new corpus to address the difficulties in the comparison of the results. All the methods of the platform are assessed using the corpus we developed (WAC). We propose a split into train (60%), development (20%) and test (20%) sets for each of the 3 datasets of WAC. This split was randomly generated, but is publicly available online. Using this split for all the methods ensure that all the results are obtained with the same data and so, are truly comparable. Additionally, we leave open the possibility to

[7] https://github.com/CompNet/Alert

Ground Truth Dataset	Perspective - Toxicity			Perspective - Severe toxicity			Hybrid method		
	Prec.	Recall	F-measure	Prec.	Recall	F-measure	Prec.	Recall	F-measure
Personal attack	84.96	86.96	85.96	54.05	93.06	68.38	81.24	70.47	75.47
Aggression	82.89	87.49	85.13	53.71	92.52	67.96	81.75	70.23	75.55
Toxicity	84.68	90.82	87.64	53.49	94.29	68.26	74.17	74.77	74.47

Table 4: Macro Precision, Recall and F-measure obtained by the 3 tested methods.

implement and add further metrics to the methods if needed, the tool being designed to ease the addition of new metrics. Different variants of the F-Measure as well as the Area Under the ROC Curve are currently implemented, since they are the metrics mainly used by the methods listed in Table 2. As mentioned before, the source code of all the works presented in Section 2. is not publicly available, so we could not include them in our platform. Instead, we focused on our previously published abuse detection method (Cécillon et al., 2019). It is a hybrid approach combining two distinct methods previously proposed by our team (Papegnies et al., 2019). The first one is content-based and relies on a set of features describing exclusively the textual content of the messages to perform the classification. Though this approach does not require a corpus of conversations to be tested, it is still interesting to assess it on the WAC corpus because of its size. Indeed, the reported performances for this method were obtained on a small dataset of less than 3,000 comments. The second approach is graph-based, and requires full conversations in order to be applied. It consists in extracting conversational networks modeling the interactions between users, before computing topological measures to describe these graphs, and using them as features during the classification process. Hence, this method completely ignores the content of the messages and only relies on the structure and dynamics of the conversation. It is typically the kind of methods for which the WAC was created. Our hybrid tool combines both text- and graph-based features.
Our benchmarking platform is currently available online, but still needs some work to implement more existing approaches. Indeed, while it currently contains only two methods, the objective is to include more methods using *Wikipedia conversations* to make it a comparative platform of many approaches in the area of automatic abuse detection.

4.2. Usage Example

We now illustrate the interest of our corpus and platform to assess the performance of the Google Perspective API as well as our own hybrid method. The way the data is split between train, development, and test sets is described in a file provided with the data. We assess the performance in terms of Precision, Recall and F-measure, separately for the 3 datasets constituting WAC (*Personal attack*, *Aggression* and *Toxicity*). Our results are presented in Table 4.
Each message in WikiConv is provided with two scores computed through the Google Perspective API: a `toxicity` and a `severe_toxicity`. To evaluate their quality, we first convert them into binary classes (*abusive* vs. *non-abusive*), by using the equal error thresholds calculated by Hua et al. (2018) following the methodology of Wulczyn et al. (2017). A message is considered toxic if its `toxicity` score is above 0.64 and severely toxic if its `severe_toxicity` score is above 0.92. Unsurprisingly, the best F-measure for the `toxicity` score is obtained with the *Toxicity* dataset. However, the performances obtained for the other 2 datasets are not much lower. Based on this observation, we can hypothesize that the method used to generate the `toxicity` and `severe_toxicity` scores may not really distinguish between *Personal attack*, *Aggression* and *Toxicity*, and relies on a more general definition of abuse. The `severe_toxicity` score yields a higher Recall than the `toxicity` one for all 3 abuse types, but the Precision is only around 54%. This poor precision is due to a lot of toxic messages being mistaken for severely toxic messages. This confirms our assumption from Section 3., *i.e. PreTox* annotations (largely based on the Google Perspective API) are less accurate than human annotations. For the hybrid method implemented in our platform, performances are similar for the 3 datasets, with a F-measure around 75%. This puts our method between both variants of the Google Perspective API. There is a clear drop in performance compared to the results obtained in our previous work, on a different corpus (Cécillon et al., 2019). This can be explained by several factors. First, the text-based part of our method relies on very standard features and could be improved by using more sophisticated ones. Second, the graph-based part was designed to operate on chat messages, and therefore to handle very large and linear conversations. In WAC, conversations have a limited size and are not linear, which decreases a lot the efficiency of this method. Therefore, there is room for improvement, and we plan to adjust both parts to better handle the characteristics of Wikipedia talk pages. In any way, our goal in this paper was only to illustrate the usefulness of our platform and corpus, and we leave the improvement of our classifier to future work.

5. Conclusion and Future Work

In this paper, we introduced a large corpus of 383k annotated user messages along with the conversations they appear in. We presented the pipeline that we developed to link two existing corpora of Wikipedia comments and extract high quality labels and thread-level information. So far, the development of context-based methods in the area of abusive comment detection was limited by the lack of large annotated corpora of conversations. This new publicly available corpus opens perspectives for new work and for extending existing work. For example, content-based methods could incorporate information about the conversation and its structure. Furthermore, the large number of messages in the corpus allows us to use it with any machine learning approach. In a second part, we presented a tool that we developed to assess some detection methods on this new corpus. A future work is to further develop this platform by integrating more methods in it. The objective is to make it a comparison platform for classification methods using the conversational corpus we proposed.

6. Bibliographical References

Chang, J. and Danescu-Niculescu-Mizil, C. (2019). Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In Conference on Empirical Methods in Natural Language Processing & International Joint Conference on Natural Language Processing, pages 4745–4756.

Cécillon, N., Labatut, V., Dufour, R., and Linarès, G. (2019). Abusive language detection in online conversations by combining content- and graph-based features. In International Workshop on Modeling and Mining Socia-Media Driven Complex Networks, volume 2 of *Frontiers in Big Data*, page 8.

Delort, J.-Y., Arunasalam, B., and Paris, C. (2011). Automatic moderation of online discussion sites. *International Journal of Electronic Commerce*, 15(3):9–30.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In AAAI/ACM Conference on AI, Ethics, and Society, pages 67–73.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In 24th international conference on world wide web, pages 29–30.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., and Asokan, N. (2018). All you need is "love": Evading hate speech detection. In 11th ACM Workshop on Artificial Intelligence and Security, pages 2–12.

Hosseini, H., Kannan, S., Zhang, B., and Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. *arXiv*, cs.LG:1702.08138.

Hua, Y., Danescu-Niculescu-Mizil, C., Taraborelli, D., Thain, N., Sorensen, J., and Dixon, L. (2018). WikiConv: A corpus of the complete conversational history of a large online collaborative community. In Conference on Empirical Methods in Natural Language Processing, pages 2818–2823.

Karan, M. and Šnajder, J. (2019). Preemptive toxic language detection in Wikipedia comments using thread-level context. In 3rd Workshop on Abusive Language Online, pages 129–134.

Mishra, P., Yannakoudakis, H., and Shutova, E. (2018). Neural character-based composition models for abuse detection. In 2nd Workshop on Abusive Language Online, pages 1–10.

Napoles, C., Tetreault, J., Pappu, A., Rosato, E., and Provenzale, B. (2017). Finding good conversations online: The yahoo news annotated comments corpus. In 11th Linguistic Annotation Workshop, pages 13–23.

Papegnies, E., Labatut, V., Dufour, R., and Linarès, G. (2017). Detection of abusive messages in an online community. In 14ème Conférence en Recherche d'Informations et Applications, pages 153–168.

Papegnies, E., Labatut, V., Dufour, R., and Linarès, G. (2019). Conversational networks for automatic online moderation. *IEEE Transactions Computational Social Systems*, 6(1):38–55.

Pavlopoulos, J., Malakasiotis, P., and Androutsopoulos, I. (2017). Deep learning for user comment moderation. In 1st Workshop on Abusive Language Online, pages 25–35.

Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive language detection using multi-level classification. In Canadian Conference on Artificial Intelligence, pages 16–27.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In 26th International Conference on World Wide Web, pages 1391–1399.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., and Edwards, L. (2009). Detection of harassment on web 2.0. In WWW Workshop: Content Analysis in the Web 2.0, pages 1–7.