Controversy and Conformity: from Generalized to Personalized Aggressiveness Detection

Kamil Kanclerz[1], **Alicja Figas**[2], **Marcin Gruza**[1], **Tomasz Kajdanowicz**[1], **Jan Kocoń**[1], **Daria Puchalska**[2], **Przemysław Kazienko**[1]
Wrocław University of Science and Technology
27 Wybrzeże Wyspiańskiego st.
Wrocław, Poland
[1]{kamil.kanclerz,marcin.gruza,tomasz.kajdanowicz,jan.kocon,kazienko}@pwr.edu.pl
[2]{238442,234800}@student.pwr.edu.pl

Abstract

There is content such as hate speech, offensive, toxic or aggressive documents, which are perceived differently by their consumers. They are commonly identified using classifiers solely based on textual content that generalize pre-agreed meanings of difficult problems. Such models provide the same results for each user, which leads to high misclassification rate observable especially for contentious, aggressive documents. Both document controversy and user nonconformity require new solutions. Therefore, we propose novel personalized approaches that respect individual beliefs expressed by either user conformity-based measures or various embeddings of their previous text annotations. We found that only a few annotations of most controversial documents are enough for all our personalization methods to significantly outperform classic, generalized solutions. The more controversial the content, the greater the gain. The personalized solutions may be used to efficiently filter unwanted aggressive content in the way adjusted to a given person.

1 Introduction

Unfortunately, in the pursuit of knowledge on the Internet, one may come across content that they consider inappropriate for various reasons, such as being too aggressive. Many users notoriously come across content that offends them while surfing the Internet. This can cause discomfort and discourage from further expansion of knowledge. To avoid this, it is important to effectively filter out content that a given user may find unwanted. This poses a risk of erroneous assessment of whether a given text is considered inappropriate by a given person. For that purpose, we need to extend commonly applied generalizing solutions and develop personalized methods that take into account beliefs and preferences of the individual user. We expect this information can be obtained from the individual's prior opinions about the offensiveness of some texts. Then, it is crucial to select the relevant texts that allow deriving as much information about users preferences as possible. Our new idea is to use some known, most controversial texts whose offensiveness is very ambiguous and depends more on subjective personal judgment. We examined how many documents has to be annotated by a given user to encapsulate their beliefs sufficiently and to improve personalized reasoning. Independently, we considered personal measures quantifying conformity of each individual. In other words, we measured to what extent a person evaluates documents similarly to others, i.e. "is a part of the mainstream". The conformity measures are used as input features for the classifier. This way, it is possible to find out the user beliefs based on their opinions regarding a relatively small number of texts. In this paper, we present novel methods of personalized aggressive content detection based on the representation of user opinion about aggressive texts. We propose: (1) conformity-based personalization, (2) class-based embeddings, and (3) annotation-based embeddings (Sec. 6). Our experiments were performed on the only relevant dataset *Wikipedia Talk Labels: Aggression* (Sec. 3). Having defined and calculated controversy of documents and conformity of users (Sec. 4), we validated our methods. The results revealed that additional individualized features: simple user conformity measures computed on few texts or embeddings of even four controversial texts significantly boost our personalized classification (Sec. 8). The gain provided by our personalized methods is greater for more controversial documents. This work is based on the results obtained in the article (Kocoń et al., 2021). In addition, in paper (Milkowski et al., 2021), we showed that the personalized approach is also effective for other subjective problems in NLP, such as

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 5915–5926
August 1–6, 2021.

recognizing emotions elicited by text. The source code we used to conduct experiments and evaluation is publicly available in CLARIN-PL GitHub repository[1].

2 Related work

It is observable a steady increase in the number of offensive (Levmore and Nussbaum, 2010), hate (Breckheimer, 2001; Brown, 2018), aggressive, toxic, cyberbullying (Chen et al., 2012), or simply socially unacceptable online messages (Ljubešić et al., 2019). There are many definitions of offensive speech, which can be summarised as speech that targets specific social groups in a way that is harmful to them (Jacobs, 2002). Some countries, such as the USA, protect the rights to use this type of speech as an acceptable form of political expression (Heyman, 2008). In turn, the law prohibits hate speech in many EU countries (Rosenfeld, 2002). Such laws pose a challenge for operators of social networking sites and other online services to identify and moderate unacceptable content. Large companies such as Facebook and Google are often accused of not doing enough to ensure that their platforms are not used to attack other people (Ben-David and Fernández, 2016). On the other hand, attempts to automatically control content often lead to the accidental blocking of content that was not intended to offend anyone.

Ambiguity of the definition of offensiveness is a serious problem. This inconsistency is visible in many reviews related to automatic detection of hate speech (Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017; Alrehili, 2019; Poletto et al., 2020) or more specifically on aggressiveness detection (Sadiq et al., 2021; Modha et al., 2020).

Automatic recognition of offensive speech is the subject of many NLP workshops, such as Semeval 2019 (Zampieri et al., 2019b), GermEval 2018 (Wiegand et al., 2018), FIRE/HASOC 2019 (Mandl et al., 2019) or PolEval 2019 (Ptaszyński et al., 2019). Classic methods do not consider context and word order, e.g. the bag-of-words model (Zhang et al., 2010) or TF-IDF (Sahlgren et al., 2018). The representation may be extended with additional ontologies (Bloehdorn and Hotho, 2004) or WordNets (Scott and Matwin, 1998; Piasecki et al., 2009; Misiaszek et al., 2014; Janz et al., 2017; Kocoń et al., 2019b) and used with SVM (Razavi et al., 2010) or logistic regression models (Waseem and Hovy, 2016; Sahlgren et al., 2018; Kocoń et al., 2018; Kocoń and Maziarz, 2021). New methods often use word embeddings (Wiegand et al., 2018; Bojanowski et al., 2017; Łukasz Augustyniak et al., 2021) (Wiegand et al., 2018; Bojanowski et al., 2017) mixed with character embeddings (Augustyniak et al., 2019), together with deep neural networks, e.g. CNN (Zampieri et al., 2019a) or LSTM (Yenala et al., 2017). The current state-of-the-art are Transformer-based architectures such as BERT (Devlin et al., 2019), ALBERT (Lan et al., 2019), XLNet (Yang et al., 2019) or RoBERTa (Liu et al., 2019). Nevertheless all these methods focus solely on the text itself. Any wider context has been considered very rarely, e.g. as time, thread or author's social network features (Ziems et al., 2020).

In articles focused on detection of aggressiveness (Modha et al., 2018; Risch and Krestel, 2018; Safi Samghabadi et al., 2020), the most often used were datasets shared at the Workshops on Trolling, Aggression and Cyberbullying (TRAC) (Kumar et al., 2018, 2020) at LREC. Few others also used the *Wikipedia Talk Labels: Aggression* (Wulczyn et al., 2017b), where all individual annotations are available, not just the majority vote. Unfortunately, we have not found any other aggression dataset, for which this information would also be given. Moreover the authors focus mainly on the multilingual aspect of the aggression detection (Modha et al., 2018; Risch and Krestel, 2018; Safi Samghabadi et al., 2020). In addition to deep neural models, less complex methods such as logistic regression are also used (Modha et al., 2018; Risch and Krestel, 2018).

To the best of our knowledge, there are no work that dealt with the subjective problem of aggressiveness detection in the personalized way. The disagreement between annotators is usually measured by a single value, e.g. using Cohen's kappa or Krippendorf's alpha, and not investigated further. The researchers prefer a higher agreement level rather than controversy. Therefore, majority annotation is used in modeling, which to some extent leads to the loss of valuable information.

There are several studies focusing on the problem of the disagreement in data annotations. This provides valuable information not only about the annotators, but also about the instances by reflecting their ambiguity (Aroyo and Welty, 2013). There may be no single right label for every text.

[1] https://github.com/CLARIN-PL/controversy-conformity

The disagreement was used to divide annotators into polarized groups (Akhtar et al., 2020) or to filter out the spammers (Raykar and Yu, 2012; Soberón et al., 2013). In (Gao et al., 2019), attention was also drawn to the problem of conformity bias, where the reviewers tend to issue similar opinions. Less frequently, the disagreement is examined at the instance level, to measure its controversy or ambiguity, as in (Aroyo and Welty, 2013). For example, (Chklovski and Mihalcea, 2003) used confusion matrices in word sense tagging task to create and explore coarse sense clusters.

3 Dataset: Wikipedia Talk Labels

We used the *Wikipedia Talk Labels: Aggression* data, gathered in the *Wikipedia Detox* project (Wulczyn et al., 2017b,a). Unlike other collections, it provides information about all annotations given by *Crowdflower* workers (not only the majority vote) for 100k+ comments from English Wikipedia. The assigned aggression score ranged from *very aggressive* (-3), via *neutral* (0), to *very friendly* (3). It was binarized to '1 - aggressive' for negative scores or '0 - nonaggressive' for neutral or friendly annotations. The dataset contained a suggested data split into *train*, *dev* and *test* set.

To enable our experiments, we removed annotations assigned by workers with less than 100 annotations in the *train* set, <20 in the *dev* set or <20 in the *test* set. Otherwise, we would not have data to extract user beliefs from and to perform personalization. We also removed users who did not assign any *aggressive* label in the *dev* set. Information about at least one text, that a specific user considered aggressive was crucial to model his individual perception of such content. Finally, there were 2,450 annotators left (Tab. 1), so we randomly divided them into 10 equal-sized folds.

The *train* set is used to calculate the representations (embeddings) of documents being classified. This is the only data exploited in the classic, generalizing approach (our baseline). The *dev* set provides information about user beliefs, i.e. their previous annotations. Individualized input features are extracted from *dev* data: (1) conformity measures and (2) personal embeddings in class-based and annotation-based personalization. Personalization-related calculations on the *dev* set refer to both training and testing procedure. The documents from the *test* set are embedded and classified by the trained model for the validation purposes.

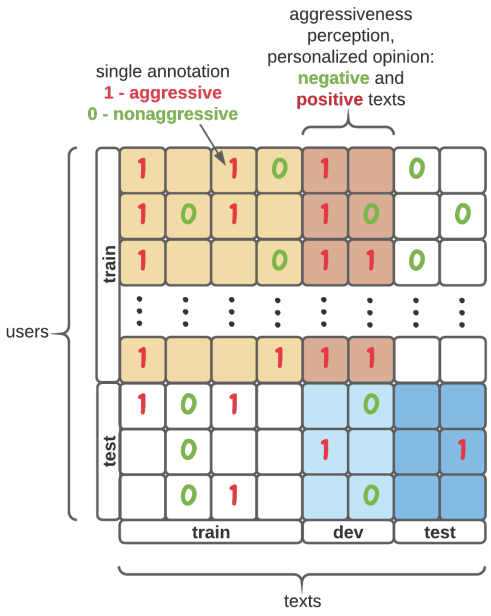

Figure 1: Split of texts and users into train and test set. The *dev* texts are solely used to quantify user beliefs: user conformity and personal embeddings. Each cell is a single text (comment) and its individual annotation.

4 Controversy and Conformity Measures

For training and testing purposes, both controversy *Contr* for documents and conformity *GConf*, *WConf* for users are calculated within the *dev* set.

4.1 Controversy

Controversy $Contr(d) \in [0, 1]$ of document d is an entropy-based measure expressed in the following

Description		Before filtering	After filtering
Comments	*train*	69,526	69,523
	dev	23,160	23,160
	test	23,178	23,178
Annotations (Ann.)	*train*	762,046	682,517
	dev	253,589	226,996
	test	349,582	304,378
Annotators	*whole set*	4053	2450
Ann. balance	*aggressive*	18.3%	18.1%
	nonaggr.	81.7%	81.9%
Ann. per comment	*mean*	11.78	10.48
	std. dev.	4.88	4.18
Ann. per annotator	*mean*	336.84	495.47
	std. dev	296.59	281.43

Table 1: Wikipedia Talk Labels dataset statistics.

way:

$$Contr(d) = \begin{cases} 0, \text{ if } n_d^0 = n_d \vee n_d^1 = n_d \\ -\sum_{c=0,1} \frac{n_d^c}{n_d} \log_2 \left(\frac{n_d^c}{n_d}\right), \text{otherwise} \end{cases}$$

where n_d^0, n_d^1 is the number of negative and positive annotations assigned to document d, respectively; n_d is the total number of document d's annotations, $n_d = n_d^0 + n_d^1$; $\frac{n_d^c}{n_d}$ approximates the probability that annotation of document d is of class c. $Contr(d) = 0$ means that all users annotated d the same, $Contr(d) = 1$ when 50% of users perceived it aggressive and 50% not.

Controversy $Contr(d)$ is used to rank documents from the *dev* dataset. The most controversial texts (top k) are embedded in class-based or annotation-based personalization. Independently, controversy is computed within the *test* data in order to investigate differences in reasoning quality for more and less controversial documents.

4.2 General conformity

General conformity $GConf(a, C) \in [0, 1]$ of human a quantifies how often a belongs to the majority of annotators evaluating individual texts. It can be of different kind depending on the class C we consider:

$$GConf(a, C) = \frac{\sum_{d \in A_a} \mathbb{1}_{\{l_d \in C \wedge l_d = l_{d,a}\}}}{\sum_{d \in A_a} \mathbb{1}_{\{l_d \in C\}}},$$

where A_a is the set of documents annotated by a; C denotes the conformity type related to the considered classes, i.e. $C = \{0\}$, $\{1\}$ or $\{0, 1\}$; $l_{d,a}$ is the class label assigned by a to document d; l_d is the d's class label obtained by majority voting. In case of equal annotations for both classes document d is considered aggressive. $GConf(a, C) = 1$ when a annotated all documents $d \in A_a$ the same like the others and no one annotated it otherwise.

Note that depending on C, conformity can be calculated in three variants: for nonaggressive ($C = \{0\}$), aggressive ($C = \{1\}$) or any documents ($C = \{0, 1\}$) annotated by a. Such three conformity values are used as input features in conformity-based personalization, Sec. 7.

4.3 Weighted conformity

Weighted conformity $WConf(a, C) \in [0, 1]$ is similar to general conformity $GConf(a, C)$ but it respects the size of the group the annotator belongs to, while evaluating the document. The larger the group with annotator a, the greater annotator a conformity:

$$WConf(a, C) = \frac{\sum_{d \in A} \sum_{c \in C} \frac{n_d^c}{n_d} \mathbb{1}_{\{l_{d,a} = c\}}}{\sum_{d \in A_a} \mathbb{1}_{\{l_{d,a} \in C\}}}.$$

5 Controversy Analysis

To have some insight into our data, we calculated controversy $Contr(d)$ on each dataset (train/dev/test). Fig. 2 presents the distribution of annotations for controversy measure in the *dev* and *test* set. In both, the ratio of aggressive to nonaggressive documents is increasing and reaching 0.5 for the most controversial documents, i.e. $Contr(d) = 1$ resulting from the same number of aggressive and nonaggressive votes. The examples of such texts are following:

"Your behaviour is inappropriate and your reaction is ludicrous. Do they give out admin rights in cornflake packets now?", $n_d^0 = n_d^1 = 5$.

"Far from being ridiculous, it is the recommended approach to follow on wikipedia. We don't simply state what either side claims, rather we report on how they are viewed by neutral 3rd party sources. Take it to WP:NPOVN if you don't believe me, rather than indulging in your continued disruptive habit of always having the WP:LASTWORD.", $n_d^0 = n_d^1 = 14$.

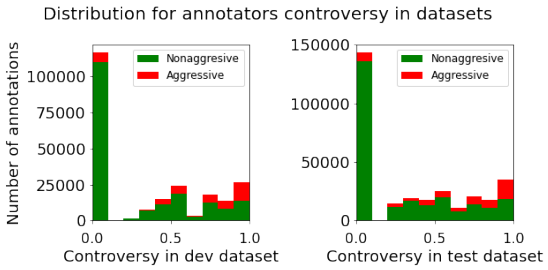

Figure 2: Distribution of controversy in documents calculated on a) the *dev* set, b) the *test* set

We learned that classic methods based solely on content analysis (not personalized) perform worse, the more controversial the documents being tested, Fig. 6. It was the main inspiration for our personalized methods.

We also checked contribution of aggressive texts for the consecutive most controversial documents included in the personal user embeddings, Fig. 3.

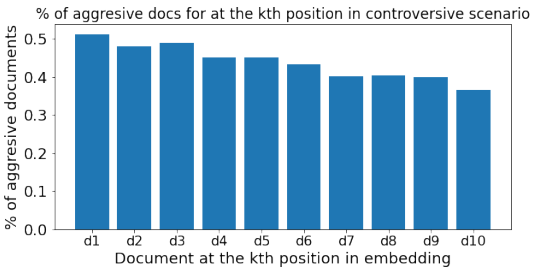

Figure 3: Contribution of aggressive texts in the following positions of the individual ranking of most controversial documents annotated by a given user.

6 Methods for Personalized Aggressiveness Detection

We assume that personal beliefs can be expressed by user activity, i.e. their individual annotations. It means that we can use information about k documents previously annotated by the user in the form of their embeddings or user conformity measures. It leads us to three novel personalization methods: (1) *conformity-based*, (2) *text-based*, and (3) *annotation-based*, Fig. 4. According to our initial studies, the most informative were user annotations provided for most controversial documents.

In *conformity-based personalization*, we exploited simple conformity measures that represent the beliefs of one user in the aggregated way: *GConf* and *WConf*. Each of them can deliver three separate values: for only aggressive, only nonaggressive, and all texts. Finally, we examined input feature sets based on only *GConf*, only *WConf*, and on both, Sec. 7.

We also propose two versions of personal embeddings for previously annotated texts: *class-based* and *annotation-based*.

The *class-based embedding* consists of two fastText embeddings of *k* documents from the *dev* set that the user rated as (1) nonaggressive and (2) separately as aggressive, Fig. 4. Each of the two embeddings can aggregate any and different number of previous user annotations; the embedding size is static for every k. If the user has not annotated any texts of given class (e.g. aggressive), the embedding represents an empty string (zeros). Overall, it is a very rare case in our experiments, mostly happening for $k = 1$.

The *annotation-based embeddings* consider all k user annotations individually. For each such text d, we use the following features: (1) the embedding of the d's content, (2) its controversy *Contr(d)*, (3) the percentage of users who rated d as nonaggressive, (4) the rating of the given user (0/1), and (5) the information on whether this rating is consistent with the the majority rating. Thus, we receive a relatively large number of input features: $300 + k * 304$.

Our general personalized aggressiveness detection procedure is as follows:

1. We ask users to annotate k most controversial documents from the pre-defined set (here *dev*).

2. Information from the first step is used to extract individually-specific features reflecting personal user beliefs, i.e. conformity measures or embeddings of these k texts (class-based and annotation-based methods).

3. A subset of the same users (upper rows in Fig. 1) annotate next documents. The data about their following annotations (embeddings of texts from *train*) together with data from step 2. are used to train the classifier.

4. For some other users (lower rows in Fig. 1), we also collect their annotations (the *test* set). Together with the information about their individual preferences (step 2.) they are used for validation (testing) purposes only.

7 Experimental setup

To validate our three personalized methods, we utilized *Wikipedia Talk Labels: Aggression*, see Sec. 3. We applied 10-fold cross-validation based on users. The first nine sets are used to train the model (upper rows in Fig. 1), while the remaining 10th set for testing (lower rows in Fig. 1). The results presented in plots are averaged over all ten folds.

Since only *dev* texts with annotations are assumed to represent prior knowledge about users, they were used to test personalization scenarios for each of our three methods: class-based, annotation-based, and conformity-based. The last one was in three variants: only three $GConf(a, C)$ measures (for $C = \{0\}, \{1\}, \{0, 1\}$), only three $WConf(a, C)$ measures, all six conformity values. Thus, we analyzed five methods in total. For each of them, we considered: (1) different number k=1,2,..20 of texts d previously annotated by user a: $d \in A_a$ (for conformity-based methods $|A_a| = k$, (2) different selection procedures for texts $d \in A_a$ used to represent a's beliefs (personalization): (2a) k most controversial texts $d \in A_a$,

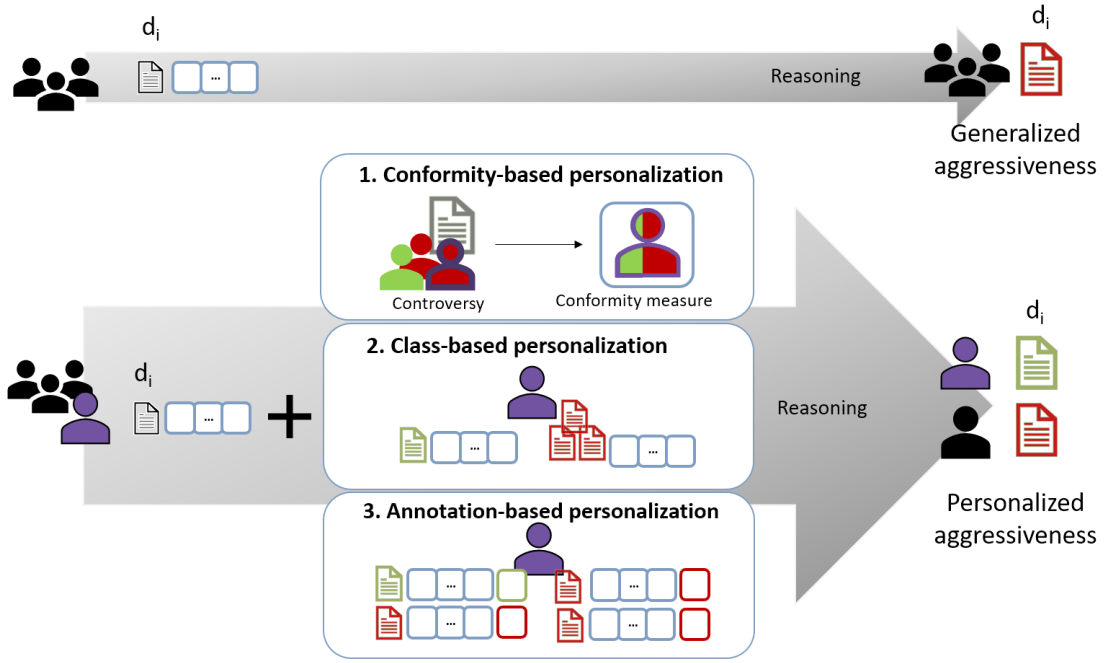

Figure 4: A classic approach generalizing output based solely on textual content (the same decision for all users) – an upper flow (our baseline). Three personalized methods proposed in the paper: (1) *Conformity-based* – additional input features – personal conformity measures (*GConf*, *WConf* or both, each for aggressive, nonaggressive or any texts); (2) *Class-based* – two embeddings of $k = 4$ texts previously annotated by a given user, one embedding for one *aggressive* text and the second for three *nonaggressive* ones; (3) *Annotation-based* – embeddings, classes and additional features for each of $k = 4$ most controversial texts previously annotated by a given user.

(2b) k class-balanced most controversial (like 2a but with class balancing), (2c) most aggressive $d \in A_a$ (rank according to % of aggressive annotations among all for d), (2d) random selection of k texts $d \in A_a$. In total, we tested: 10 folds x (5 methods x 20 distinct k no. of texts x 4 selection + 1 baseline) = 4,010 models.

The logistic regression models were optimized during the training process by using the L2 regularization and the early stopping mechanism. Both of them aim to prevent overfitting and the early stopping mechanism additionally ensures that the model instance that achieved the best loss function score is preserved. The models were run on Intel Xeon Processor E5-2650 v4.

We also compared our personalized methods with the baseline, i.e. the commonly investigated approach generalizing user perception. It exploited only the evaluated text embeddings as the input.

We considered classification performance not only for the whole test set but also in its breakdown of 10 percentage buckets according to three independent rankings of test docs: (1) most controversial ($Contr(d)$), (2) with least conformity $GConf(a, \{0, 1\})$, averaged over all $a \in Test$ annotating d, (3) least $WConf(a, \{0, 1\})$. Here, the measures were computed for the *test* set only, not for *dev*. It was used to investigate where our models more outperform the baseline. In order to generate text embeddings in each personalization method, we used the fastText library (Bojanowski et al., 2017; Joulin et al., 2017). It offers pre-trained word vectors for 157 languages, based on the continuous bag of words (CBOW) model in a 300-dimensional space, with character n-grams of length 5.

8 Validation of personalization methods

Both class-based and annotation-based methods were tested using various rankings while selecting texts for personal embeddings: most controversial, class-balanced most controversial, most aggressive, and random. The conformity-based methods were evaluated in terms of the measure variant used: general conformity, weighted conformity, and both, all with random selection of texts.

8.1 Conformity-based Personalization

The results for three conformity-based personalization methods, i.e. three different sets of input conformity features (Sec. 7) and various number k

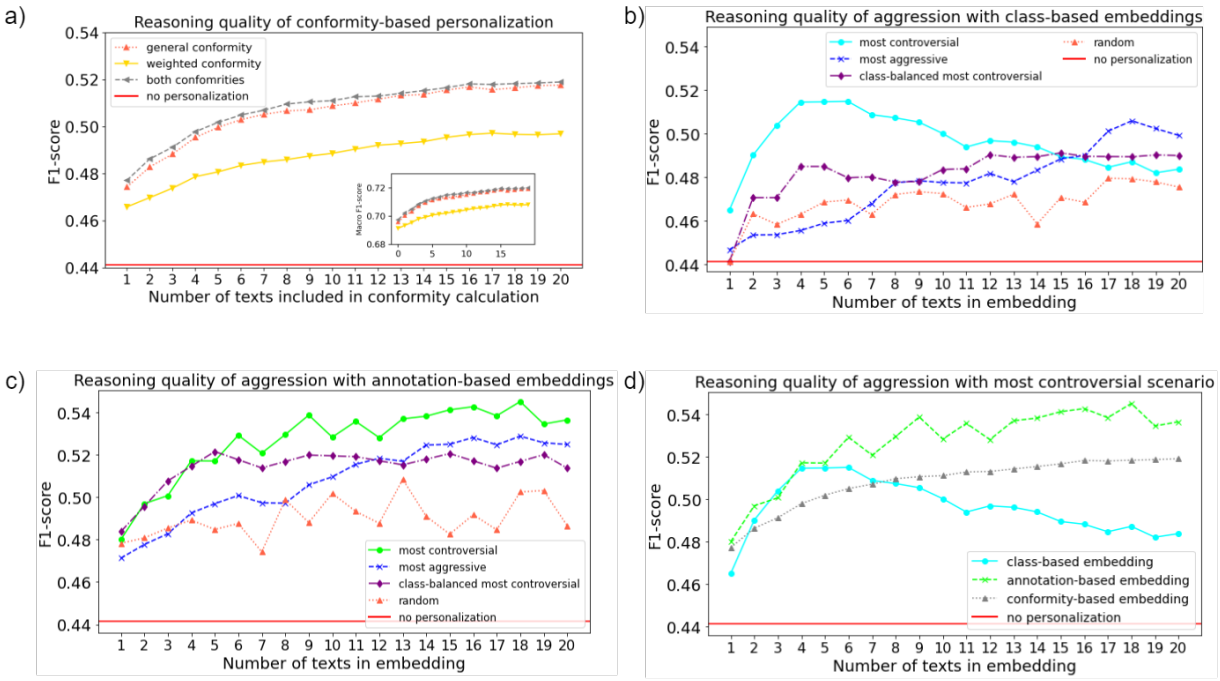

Figure 5: Performance of three personalized methods proposed in the paper, only for the *aggression* class: (a) conformity-based, in inset we inserted evaluation results for both classes; (b) class-based; (c) annotation-based; (d) comparison of the best method of each type. Both (b) and (c) were evaluated using various rankings while selecting texts for personal embeddings: most controversial, class-balanced most controversial, most aggressive, and random. Macro F1 score for both classes have the same shapes by with different range for Y: 0.68–0.73.

of texts used to calculate user conformity are shown in Fig. 5a. The greater k results in more precise evaluation of user conformity. It also directly and positively impacts on model performance, although gains for $k > 15$ are very small.

Additionally, we considered the performance for more and less controversial documents in the *test* set, Fig. 6a. It is clearly visible that the non-personalized method is completely lost for the most controversial documents. However, our conformity-based models lose relatively less. It appears that their gain (smaller loss) is greater for 30% most controversial texts. In other words, the greater controversy, the greater gain from personalization.

8.2 Class-based Embeddings

Fig. 5b describes evaluation of class-based embeddings for various text selection approaches and different number of previously annotated texts. The performance was shown only for texts from the *aggression* class (the same plot shapes were for macro F1 and both classes). The models using the most controversial texts for selection reached the best results in 14 out of 20 cases (70%). The highest F1 score was achieved for only 4 texts representing user beliefs. It was greater than the model without any personalization by over 7pp.

8.3 Annotation-based Embeddings

Annotation-based embeddings were tested for the same rankings as in Sec. 8.2, Fig. 5c. The most controversial texts used to generate user representations and feed the model provided the best results in 17 out of 20 cases (85%). The best performance was achieved while using 18 texts to represent user personal beliefs – then, the input consisted of 5,772 features. The F1 score of this model was greater than the baseline by over 10pp.

The greater gain compared to the not personalized method is exposed for 50% of the most controversial texts in the *test* set; the greatest for 10% of the most controversial – even 22.7 percentage points (twice better: 44.0% vs. 21.3%), Fig. 6b.

8.4 Comparison of personalization methods

The best models from each personalization method, which were achieved for annotations of most controversial texts, are compared in Fig. 5d. Models based on annotation-based embeddings provided significantly better results than the others in 10 out of 20 cases of k values (50%). The conformity-based models performed better than other models in

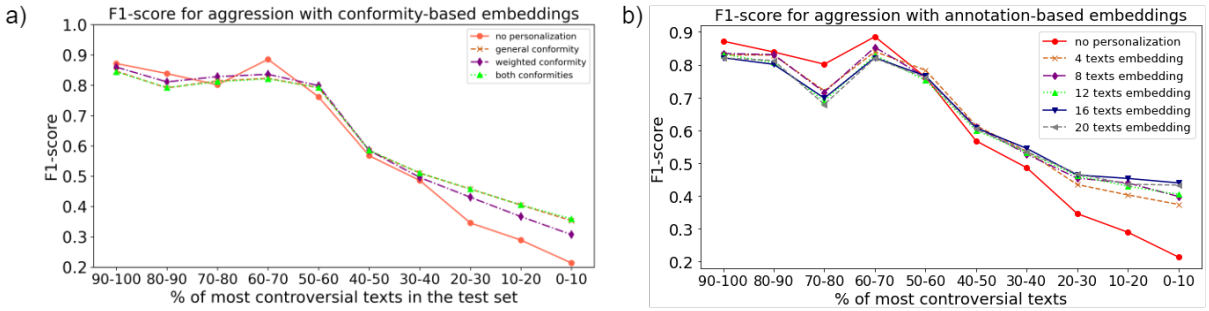

Figure 6: Performance of two personalized methods proposed in the paper, only for the *aggression* class: (a) conformity-based; (b) annotation-based. Both were evaluated on documents d in the *test* set, sorted in ascending order by $Contr(d)$ measure, 0-10 denotes 10% of the most controversial texts.

3 out of 20 cases (15%); it referred to the smallest number of texts considered ($k = 1 \div 3$). The highest value of F1 score was achieved by the model using 18 texts to represent user personal beliefs. However, this solution used 5,772 input features, whereas the much simpler conformity-based model with 306 input features was only 2.7 percentage points worse. Simultaneously, conformity-based model training time was 38.6 times faster than the annotation-based one, Fig. 7.

Practically, we would like to avoid bothering the user with too many previous annotations, i.e. we may want to limit k to just a few, for example $k = 4$. Then, we should select k most controversial texts and use either class-based or conformity-based personalization. They learn just as fast but keep the same performance: 7.3 percentage points, 5.7 percentage points greater F1 for class *aggressive*, respectively, and 3.9 percentage points, 3.2 percentage points greater macro F1 (for both classes), respectively.

The worst performance was observed for models using class-based embeddings. The results of evaluation on all texts are presented in Fig. 5d.

Random selection of k texts for personalization is almost always worse than dedicated rankings, Fig. 6b,c. Most controversial texts turned out to be the best option that usually outperformed the most aggressive and class-balanced most controversial.

9 Discussion

A valuable observation from our experiments is that already one document used to valuate user beliefs is enough to significantly improve reasoning, Fig. 5d. Anyway, more texts in personalization keep boosting the performance, but about 4-5 previously annotated most controversial documents seem to be a reasonable trade-off between reasoning quality and user annoyance.

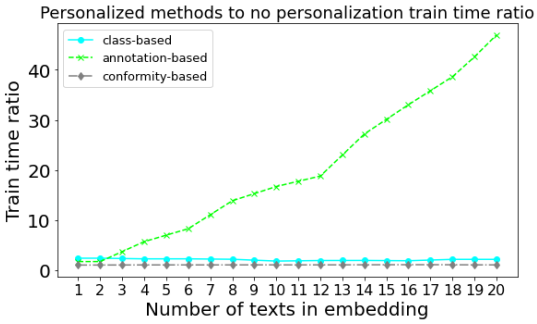

Figure 7: Training computing time for personalization methods in reference to no personalization method.

Annotation-based embeddings most precisely express user opinions, but it comes at the cost of linearly longer learning and demand for more samples. They also cannot easily adapt to different number k of personalization documents.

We decided to utilize very fast logistic regression model with fastText embeddings, since we wanted to examine thousands of models related to multiple scenarios, not all are presented here.

We belief our personalization methods establish a new research direction: how to effectively and efficiently embed user beliefs? We expect new methods will be developed for that purpose.

One of the most important postulate derived from our research is the demand for new datasets collections. We need annotations of individual humans rather than aggregated and agreed general beliefs received by majority voting, by annotator training, or by removal of controversial texts.

Besides, our personalization methods may be applied to any NLP problem with inconsistencies between people. It especially refers to diverse emotions evoked by textual content, hate speech, detection of cyberbullying or offensive, toxic, abusive,

harmful, or socially unaccepted content.

The common problem of imbalanced classes in aggressiveness detection (Tab. 1, Fig. 12) will be addressed in future work.

10 Conclusions

The main conclusion from our research is that the natural controversies associated with individual perceptions of contents should not be overlooked or reduced but rather directly exploited in personalized solutions. Ultimately, this reflects the diversity in our societies.

Our three new personalization methods make use of texts previously annotated by a given user by means of conformity measures, class-based or annotation-based embeddings. Just a few documents are able to capture individual user beliefs, the more so, the more controversial documents they relate to. As a result, all our methods outperform classic solutions that generalize offensiveness understanding. The gain is greater for more controversial documents.

The personalization solutions can also be applied to other NLP problems, where the content tends to be subjectively perceived as hate speech, cyberbullying, abusive or offensive, as well as in prediction of emotions elicited by text (Kocoń et al., 2019a; Milkowski et al., 2021) and even in sentiment analysis (Kocoń et al., 2019; Kanclerz et al., 2020).

We keep working on testing of our methods on more resource-demanding but also more SOTA language representations: XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), and XLM-RoBERTa (Conneau et al., 2020).

Acknowledgments

This work was financed by (1) the National Science Centre, Poland, project no. 2020/37/B/ST6/03806; (2) the Polish Ministry of Education and Science, CLARIN-PL Project; (3) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN - Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing*, pages 151–154.

A. Alrehili. 2019. Automatic hate speech detection on social media: A brief survey. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6.

Lora Aroyo and Chris Welty. 2013. Harnessing disagreement in crowdsourcing a relation extraction gold standard. Technical report, Technical Report.

Łukasz Augustyniak, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Comprehensive analysis of aspect term extraction methods using various text embeddings. *Computer Speech & Language*, 69:101217.

Łukasz Augustyniak, Tomasz Kajdanowicz, and Przemysław Kazienko. 2019. Aspect detection using word and char embeddings with (bi) lstm and crf. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 43–50. IEEE.

Anat Ben-David and Ariadna Matamoros Fernández. 2016. Hate speech and covert discrimination on social media: Monitoring the facebook pages of extreme-right political parties in spain. *International Journal of Communication*, 10:27.

Stephan Bloehdorn and Andreas Hotho. 2004. Text classification by boosting weak learners based on terms and concepts. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 331–334. IEEE.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Peter J Breckheimer. 2001. A haven for hate: the foreign and domestic implications of protecting internet hate speech under the first amendment. *S. Cal. L. Rev.*, 75:1493.

Alexander Brown. 2018. What is so special about online (as compared to offline) hate speech? *Ethnicities*, 18(3):297–326.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80.

Timothy Chklovski and Rada Mihalcea. 2003. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *In Proceedings of RANLP 2003*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

P. Fortuna and S. Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51:1 – 30.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. *CoRR*, abs/1903.11367.

Steven J Heyman. 2008. Hate speech, public discourse, and the first amendment. *Oxford University Press, Forthcoming*.

James B Jacobs. 2002. Hate crime: Criminal law and identity politics: Author's summary. *Theoretical Criminology*, 6(4):481–484.

Arkadiusz Janz, Jan Kocoń, Maciej Piasecki, and Zaśko-Zielińska Monika. 2017. plWordNet as a Basis for Large Emotive Lexicons of Polish. In *LTC'17 8th Language and Technology Conference*, Poznań, Poland. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Kamil Kanclerz, Piotr Miłkowski, and Jan Kocoń. 2020. Cross-lingual deep neural transfer learning in sentiment analysis. *Procedia Computer Science*, 176:128–137.

Jan Kocoń, Arkadiusz Janz, and Maciej Piasecki. 2018. Classifier-based polarity propagation in a wordnet. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jan Kocoń and Marek Maziarz. 2021. Mapping wordnet onto human brain connectome in emotion processing and semantic similarity recognition. *Information Processing & Management*, 58(3):102530.

Jan Kocoń, Piotr Miłkowski, and Monika Zaśko-Zielińska. 2019. Multi-level sentiment analysis of polemo 2.0: Extended corpus of multi-domain consumer reviews. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 980–991.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: from data-centric to human-centred approach. *Information Processing & Management*.

Jan Kocoń, Arkadiusz Janz, Piotr Miłkowski, Monika Riegel, Małgorzata Wierzba, Artur Marchewka, Agnieszka Czoska, Damian Grimling, Barbara Konat, Konrad Juszczyk, Katarzyna Klessa, and Maciej Piasecki. 2019a. Recognition of emotions, valence and arousal in large-scale multi-domain text reviews. In Zygmunt Vetulani and Patrick Paroubek, editors, *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 274–280. Wydawnictwo Nauka i Innowacje, Poznań, Poland.

Jan Kocoń, Arkadiusz Janz, Monika Riegel, Małgorzata Wierzba, Artur Marchewka, Agnieszka Czoska, Damian Grimling, Barbara Konat, Konrad Juszczyk, Katarzyna Klessa, and Maciej Piasecki. 2019b. Propagation of emotions, arousal and polarity in WordNet using Heterogeneous Structured Synset Embeddings. In *Proceedings of the 10th International Global Wordnet Conference (GWC'19)*.

Ritesh Kumar, Atul Kr Ojha, Bornini Lahiri, Marcos Zampieri, Shervin Malmasi, Vanessa Murdock, and Daniel Kadar, editors. 2020. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA).

Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, editors. 2018. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Saul Levmore and Martha Craven Nussbaum. 2010. *The offensive Internet: Speech, privacy, and reputation*. Harvard University Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The frenk datasets of socially unacceptable discourse in slovene and english. In *International Conference on Text, Speech, and Dialogue*, pages 103–114. Springer.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC Track at FIRE 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Piotr Milkowski, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling, and Jan Kocoń. 2021. Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*. Association for Computational Linguistics.

Andrzej Misiaszek, Przemysław Kazienko, Marcin Kulisiewicz, Łukasz Augustyniak, Włodzimierz Tuligłowicz, Adrian Popiel, and Tomasz Kajdanowicz. 2014. Belief propagation method for word sentiment in wordnet 3.0. In *Asian Conference on Intelligent Information and Database Systems*, pages 263–272. Springer.

Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2018. Filtering aggression from the multilingual social media feed. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 199–207, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sandip Modha, Prasenjit Majumder, Thomas Mandl, and Chintak Mandalia. 2020. Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. *Expert Systems with Applications*, 161:113725.

Maciej Piasecki, Bernd Broda, and Stanislaw Szpakowicz. 2009. *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław.

Fabio Poletto, Valerio Basile, M. Sanguinetti, Cristina Bosco, and V. Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. In *LREC 2020*.

Michal Ptaszyński, Agata Pieciukiewicz, and Paweł Dybala. 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. In *Proceedings of the PolEval 2019 Workshop*, pages 89–110. Institute of Computer Science, Polish Academy of Sciences.

Vikas C. Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J. Mach. Learn. Res.*, 13(1):491–518.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence*, pages 16–27, Berlin, Heidelberg. Springer Berlin Heidelberg.

Julian Risch and Ralf Krestel. 2018. Aggression identification using deep learning and data augmentation. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 150–158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Michel Rosenfeld. 2002. Hate speech in constitutional jurisprudence: a comparative analysis. *Cardozo L. Rev.*, 24:1523.

Saima Sadiq, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, Gyu Sang Choi, and Byung-Won On. 2021. Aggression detection through deep neural model on twitter. *Future Generation Computer Systems*, 114:120 – 129.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Magnus Sahlgren, Tim Isbister, and Fredrik Olsson. 2018. Learning representations for detecting abusive language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 115–123, Brussels, Belgium. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Sam Scott and Stan Matwin. 1998. Text classification using wordnet hypernyms. In *Usage of WordNet in Natural Language Processing Systems*.

Guillermo Soberón, Lora Aroyo, Chris Welty, Oana Inel, Hui Lin, and Manfred Overmeen. 2013. Measuring crowd truth: Disagreement metrics combined with worker behavior filters. In *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web - Volume 1030*, CrowdSem'13, page 45–58. CEUR-WS.org.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In

Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), pages 1–10.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017a. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017b. Wikipedia talk labels: Aggression.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Harish Yenala, Ashish Jhanwar, Manoj Chinnakotla, and Jay Goyal. 2017. Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52.

Caleb Ziems, Ymir Vigfusson, and Fred Morstatter. 2020. Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):808–819.