# Datasets of Slovene and Croatian Moderated News Comments

**Nikola Ljubešić**
Jožef Stefan Institute
Jamova cesta 39,
1000 Ljubljana, Slovenia
nikola.ljubesic@ijs.si

**Tomaž Erjavec**
Jožef Stefan Institute
Jamova cesta 39,
1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

**Darja Fišer**
Faculty of Arts,
University of Ljubljana
Aškerčeva cesta 2, 1000 Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

## Abstract

This paper presents two large newly constructed datasets of moderated news comments from two highly popular online news portals in the respective countries: the Slovene RTV MCC and the Croatian 24sata. The datasets are analyzed by performing manual annotation of the types of the content which have been deleted by moderators and by investigating deletion trends among users and threads. Next, initial experiments on automatically detecting the deleted content in the datasets are presented. Both datasets are published in encrypted form, to enable others to perform experiments on detecting content to be deleted without revealing potentially inappropriate content. Finally, the baseline classification models trained on the non-encrypted datasets are disseminated as well to enable real-world use.

## 1 Introduction

With the rapid rise of user-generated content, there is increased pressure to manage inappropriate online content with (semi)automated methods. The research community is by now well aware of the multiple faces of inappropriateness in on-line communication, which preclude the use of simple vocabulary-based approaches, and are therefore turning to more robust machine learning methods (Pavlopoulos et al., 2017). These, however, require training data.

Currently available datasets of inappropriate on-line communication are primarily datasets of English, such as a Twitter dataset annotated for racist and sexist hate speech (Waseem and Hovy, 2016)[1], the Wikimedia Toxicity Data Set (Wulczyn et al., 2017)[2], the Hate Speech Identification dataset containing tweets annotated as hate speech, offensive language, or neither (Davidson et al., 2017)[3], and the SFU Opinion and Comment Corpus consisting of online opinion articles and their comments annotated for toxicity[4].

Datasets in other languages have recently also started to emerge, with a German Twitter dataset focused on the topic of refugees in Germany (Ross et al., 2017)[5] and a Greek Sport News Comment dataset containing moderation metadata (Pavlopoulos et al., 2017)[6].

In this paper we present two new and large datasets of news comments, one in Slovene, and one in Croatian. Apart from the texts, they also contain various metadata, the primary being whether the comment was removed by the site administrators. Given the sensitivity of the content, we publish the datasets in full-text form, but with user metadata semi-anonymised and the comment content encrypted via a simple character replacement method using a random, undisclosed bijective mapping, similar to the encryption method applied to the Gazzetta Greek Sport News Comments dataset[7] introduced in Pavlopoulos et al. (2017). The two datasets presented in this paper are aimed at further enriching the landscape of datasets on inappropriate online communication overall, but especially the dimension of multilinguality and multiculturality.

[1] https://github.com/ZeerakW/hatespeech

[2] https://figshare.com/projects/Wikipedia_Talk/16731

[3] https://data.world/crowdflower/hate-speech-identification

[4] https://github.com/sfu-discourse-lab/SOCC

[5] https://github.com/UCSM-DUE/IWG_hatespeech_public

[6] https://straintek.wediacloud.net/static/gazzetta-comments-dataset/gazzetta-comments-dataset.tar.gz

[7] https://straintek.wediacloud.net/static/gazzetta-comments-dataset/README.txt

*Proceedings of the Second Workshop on Abusive Language Online (ALW2)*, pages 124–131
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics

The contributions of this paper are the following: (1) we introduce two new datasets annotated for content inappropriateness, (2) we perform a basic analysis of the type of the deleted content, (3) we investigate whether the deleted content is more dependent on specific users, threads or locations in a thread, (4) we build baseline predictive models on these datasets and (5) we publish the full but semi-anonymised and encrypted datasets, as well as the models built on the non-encrypted data ready to be used in real-life scenarios.

## 2 Dataset description

This section gives a description of the two datasets, which we obtained from two different sources from different countries and in different languages. Both datasets are comprehensive in the sense that they contain all the comments from the given time period. Their main value in the context of studying inappropriate content lies in the fact that they also contain all the comments that were deleted by the moderators of the two sites.

The Slovenian **MMC** dataset contains comments on news articles published on the MMC RTV web portal[8], the on-line portal of the Slovenian national radio and television. The dataset comprises all the comments from the beginning of 2010 until the end of 2017, i.e. eight years' worth of content, including deleted comments. The portal is monitored by moderators who delete comments that contain hate speech but also those that are not relevant for the thread or are spam by advertisers.

We obtained the dataset as a CSV file, where we deleted comments with formatting errors, removed illegal UTF-8 characters and remnants of formatting, and then converted the dataset into XML. Apart from the text itself, each comment contains the following metadata: comment ID; ID of the news article that is commented (note, however, that we did not receive the news articles themselves due to copyright limitations); user ID; time stamp; whether the comment was deleted or not; and the number of up- and down-votes.

The Croatian **STY** dataset contains comments on articles from the news portal 24sata[9] which is owned by Styria Media International. They comments in the dataset span from 2007-09-12 to 2017-07-21, i.e. almost ten years of content. Until 2016 the portal was monitored by one moderator, with the last two years of content being moderated by two moderators. Both hate speech and spam are deleted, and the respective users are banned for an amount of time depending on the frequency of their misbehavior.

We received the dataset as a SQL database dump, where we, similarly to MMC, cleaned the file and converted it to a similar XML as the MMC one. Here, the metadata was somewhat different, comprising: comment ID; ID of the news article that is commented (again, we did not receive the news articles themselves); where applicable, ID of comment that is being replied to; the ID of the thread to which the comment belongs; the user ID; the user name; time stamp; whether the comment was deleted or not; and the number of replies to the comment.

### 2.1 The datasets in numbers

Table 1 gives the sizes of the two datasets in users, texts (comments) and words, split into retained and deleted comments, and overall. As can be seen, both datasets are substantial, having together almost 25 million comments, and over 700 million words. The two datasets have a similar size per year, but given that the STY dataset has a longer time span it is also larger, with over 407 million words, against 325 million words in MMC. Interestingly, given their comparable size, the STY dataset has many more comments (17 million, as opposed to only 7.6 million of MMC) as well as many more users (185 thousand as against 42 thousand of MMC). On the other hand, MMC users write significantly longer texts.

As for the deleted comments, we first note that for a user to be classified into either of the Yes/No deleted row, it suffices that one of their comments has been deleted (or not), so the two percentages do not sum to 100%. The percentages reveal that the two portals adopt somewhat different deletion policies: for MMC almost half of the users had at least one comment deleted, while under 10% had a comment deleted in STY. Similarly for texts, the MMC portal deleted over 8% of the texts, while STY deleted under 2%. The proportion of the number of deleted words is lower for MMC and higher for STY, meaning that with MMC the deleted comments are typically shorter than the retained texts, while for STY they are slightly longer.

[8] http://www.rtvslo.si/

[9] https://www.24sata.hr/

| Corpus | Deleted | Users | | Texts | | Words | |
|---|---|---|---|---|---|---|---|
| MMC | No | 41,142 | 96.8% | 6,965,725 | 91.7% | 302,123,513 | 92.9% |
| MMC | Yes | 20,086 | 47.3% | 630,961 | 8.3% | 23,102,063 | 7.1% |
| MMC | Σ | 42,502 | 100.0% | 7,596,686 | 100.0% | 325,225,576 | 100.0% |
| STY | No | 181,626 | 98.0% | 16,732,818 | 98.2% | 399,214,351 | 98.0% |
| STY | Yes | 17,810 | 9.6% | 310,147 | 1.8% | 8,334,776 | 2.0% |
| STY | Σ | 185,266 | 100.0% | 17,042,965 | 100.0% | 407,549,127 | 100.0% |
| Σ | No | 222,768 | 97.8% | 23,698,543 | 96.2% | 701,337,864 | 95.7% |
| Σ | Yes | 37,896 | 16.6% | 941,108 | 3.8% | 31,436,839 | 4.3% |
| Σ | Σ | 227,768 | 100.0% | 24,639,651 | 100.0% | 732,774,703 | 100.0% |

Table 1: Sizes of MMC and STY datasets.

| | MMC | STY |
|---|---|---|
| Calumination | 6 | 8 |
| Discrimination | 11 | 10 |
| Disrespect | **37** | 21 |
| Insult | 10 | **42** |
| Irony | 19 | 10 |
| Swearing | 3 | 14 |
| Other | 19 | 18 |
| Σ | 105 | 124 |

Table 2: Categories of deleted comments.

### 2.2 Types of inappropriate content

To gain more insight into the nature of the deleted comments, we manually classified 100 random deleted comments from each datasets into the 9 categories proposed by Pavlopoulos et al. (2017): calumniation, discrimination, disrespect, hooliganism, insult, irony, swearing, threat and other. Both samples were annotated by the same annotator. Where required, multiple labels were assigned to a comment.

As shown in Table 2, there are many differences between the two datasets. In the MMC sample, the most frequently represented category is disrespect (37) while swearing is the least frequent (3). Only 5 of the 100 comments were annotated with double labels. In the STY sample, on the other hand, 17 received a double and 1 a triple label (most frequent combinations being insult and swearing). The most frequent category in the STY sample is insult (42) and the least frequent one threat (1), which does not appear in the MMC sample. In general, the Croatian sample of deleted comments contains worse forms of inappropriate content compared to the Slovene one (e.g., many more cases of insults and swearing compared to more subtle irony which is particularly common in the Slovene sample). Beyond the types of inappropriate comments, we have also observed differences in the persons, groups and institutions towards which the disrespectful comments are targeted. Whereas the targets are the expected "culprits" in the Croatian sample (e.g. marginalized members of the society), in the Slovenian sample, most of them are targeted at the national broadcasting service, its journalists or the administrators of the on-line comments, especially in the category of disrespectful comments.

The reasons for these differences could lie in the different positions of the two media (one national and the other one private) and their subsequently different policies for the treatment of inappropriate content, with MMC deleting more and STY only the most blatant examples of inappropriate comments. Another reason could also be cultural differences, such as more widespread swearing in public discourse in Croatia compared to Slovenia. Interestingly, in both samples a substantial amount (19 vs. 18) of the analysed comments did not belong to any of the categories specified by Pavlopoulos et al. (2017), suggesting that the annotation schema might benefit from further refinements.

## 3 Dataset Analysis

This section presents a basic analysis of deleted vs. retained content in both datasets. We analyze (1) the distribution of deleted vs. retained content through the years, (2) the distribution of deleted content among users, (3) the distribution of deleted content in threads, and (4) the distribution of the relative positions of the deleted comments in a thread. We compare the distributions (2)-(4) with their random counterparts to see

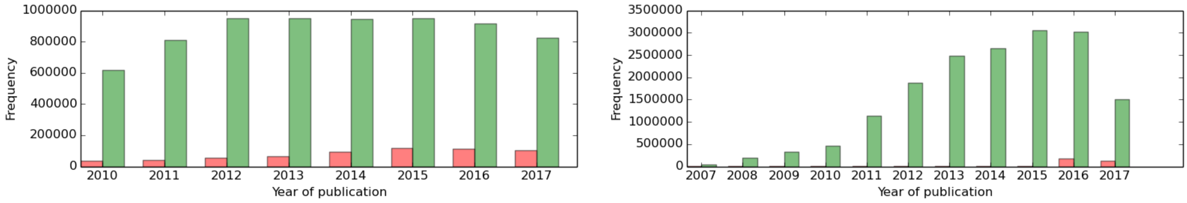

Figure 1: Distribution of deleted (red) and retained (green) comments throughout the publication years for MMC (left) and STY (right).

whether there is a dependence of comment deletion on these three phenomena: user, thread, and location in a thread.

### 3.1 Distribution through time

Figure 1 shows the distribution of the deleted and retained comments in each dataset throughout the publication years. The trends in the two datasets are quite different. While the STY dataset has an obvious increase in the number of comments throughout the years, the number of comments in the MMC dataset is rather stable. The number of the deleted comments throughout the years is even more different in the two datasets. Most of the deleted content in the STY dataset is from 2016 and 2017, which indicates an obvious change in the policy of content deletion.[10] In the MMC dataset, on the other hand, the percentage of the filtered content is rather stable throughout the years, with only a slight increase through time. The observed difference can be followed back to the type of publishers: STY is a commercial publisher, freely modifying filtering rules, whereas MMC is a national broadcasting company and is as such required to have a much more elaborate and strict, as well as more stable code of conduct.

### 3.2 Distribution per user

Figure 2 depicts the distribution of the percentage of comments deleted from each user, taking into account only users that published 10 or more comments to ensure a proper representation of the percentages on a histogram with 10 bins. The plot shows similar trends in both languages, with most users having 10% or less of their comments deleted. In both datasets we see an increase in the percentage of users having all their comments deleted. This phenomenon can be followed back to the practice of deleting users and all their corresponding content.

We hypothesize that this distribution is significantly different from a random one, i.e., that there are users whose comments are deleted more often than by chance. Put in simpler terms, we assume that inappropriate comments are not a blunder that happens to everyone now and then but that there are consistently "non-conforming" users whose comments are deleted more often than those of other users. We test our hypothesis by applying the Kolmogorov-Smirnov non-parametric two-tailed test (Massey, 1951) of the equality of two distributions, with the null hypothesis that there is no difference between the observed and the expected random distribution. We calculate the expected, random distribution by calculating the probability of a comment to be deleted in a dataset and generating a dataset with the identical distribution of comments among users, calculating whether the comment is deleted via a random function with the deletion probability as estimated on the real dataset. On the MMC test we obtain a statistic of 0.274 with the p-value, i.e., the probability that we might falsely reject the null hypothesis that the two distributions are identical being close to 0.0. For the STY dataset we obtain a statistic of 0.371, with a p-value being close to 0.0. These results show that in each dataset some users' comments are being deleted more often than by chance.

To measure to what amount each of the distributions are different to a random one, we calculate the Wasserstein distance (Ramdas et al., 2017) between the observed and the random distributions. While we obtain a distance of 0.068 for the MMC distribution, this distance for the STY distribution is 0.032.[11] This inspection shows that in the

[10] The increase in the amount of deleted content can be followed back to the internal decision of the newspaper to make the moderation model more strict, resulting in greater identification of inappropriate content, but also an increase in the amount of inappropriate content aimed at the moderators.

[11] We repeat the calculation on the STY dataset only on the data from 2016 and 2017 to control for the different approach to deleting comments before and from 2016, obtaining

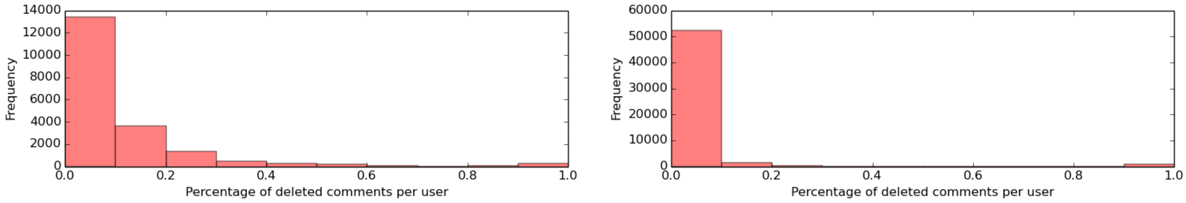

Figure 2: Distribution of the percentage of comments deleted per user publishing 10 or more comments for MMC (left) and STY (right).

Slovene MMC dataset the phenomenon of deleting content of specific users is more prominent than in the Croatian STY dataset.

### 3.3 Distribution per thread

In this subsection we repeat the calculations performed in the previous subsection on user-focused distributions of deleted content, only that this time we utilize the thread structures which are available in both datasets. For this analysis, as well as the following one presented in Section 3.4, we take into account only threads with at least 10 comments, again to ensure a proper representation on a 10-bin histogram.

In Figure 3 we plot the probability distribution of the percentage of comments deleted in each thread, obtaining a similar, long-tailed distribution as with the user-focused distribution. As with the user distributions, the MMC dataset has a larger number of threads with slightly higher percentage of deleted content (higher than 10%), indicating a less random deletion process in the MMC dataset, i.e., that there are threads that have more content deleted than would be expected by chance.

Similar to our previous calculations, we first calculate whether the obtained distributions are different from distributions obtained by randomly deleting comments in threads by applying the Kolmogorov-Smirnov test of the equality of two distributions. We then quantify the distance of the observed distributions to the random distributions via the Wasserstein distance.

When applying the Kolmogorov-Smirnov test on the MMC dataset, we obtain a statistic of 0.292 with a p-value close to 0.0, while on the STY dataset the obtained statistic is 0.333 with a p-value also close to 0.0. Based on this we can reject the null hypothesis that the random and the observed probability distributions are the same on both datasets. In other words, deletion on some threads is more prominent than on others.

We continue by calculating the Wasserstein distance metric between the observed and the random distribution, with an obtained distance of 0.036 on the MMC dataset and a distance of 0.010 on the STY dataset. Similar to the previous measurements on the percentage of users' deleted content, a stronger difference between the two distributions is again observed on the MMC dataset, showing the deletion on that dataset to be less random. Furthermore, the distances obtained on the thread-dependent distributions are almost half the size of the distances calculated on the user-dependent distributions, showing that comments of specific users are deleted more often than comments on specific threads, which is an interesting result. Note that the user-focused and thread-focused distances are directly comparable as both distributions are defined on the same scale between 0 and 1.

### 3.4 Distribution per location in thread

Finally, we inspect the distributions of the deleted comments as per their relative location in a discussion thread. We apply the same calculations as with the user-focused and thread-focused analyses presented in the previous two subsections.

We first analyze the plot of the distribution of the relative location in a thread where a comment is deleted, which is given in Figure 4. Our expectation was that more comments will be deleted in the middle and at the end of the thread because discussions get heated gradually. Both distributions seem very similar, with a close to uniform distribution. However, on both distributions we observe a trend that is opposite to our expectations, namely that comments are deleted more often at the beginning of a thread. The Kolmogorov-Smirnov test of the equality of the random and the observed distribution gives the test statistic of 0.027 and the p-value of $1.448 * 10^{-199}$ on

an identical distance.

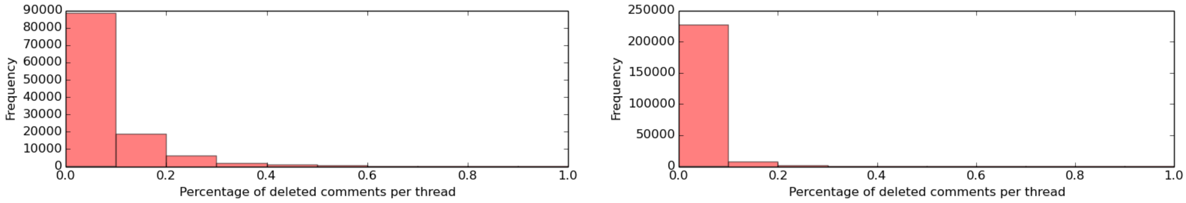

Figure 3: Distribution of the percentage of comments deleted in each thread containing 10 or more comments for MMC (left) and STY (right).

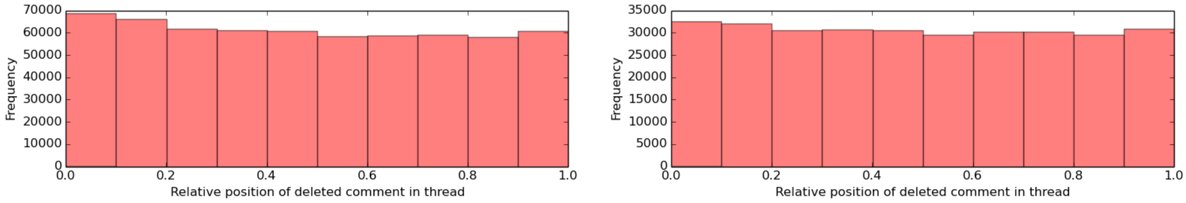

Figure 4: Distribution of the relative position of comments deleted in a thread containing 10 or more comments for MMC (left) and STY (right).

the MMC dataset, while the test statistic on the STY dataset is 0.017 and the p-value $3.26 * 10^{-39}$. The differences between the observed and the random distribution are still highly significant, but less than with the other analyses. Again, the STY dataset appears more random-like than the MMC dataset. We double-check these observations by calculating the Wasserstein distance between observed and random distributions. On the MMC dataset the distance is 0.018, while on the STY dataset it is 0.011. On the MMC dataset this is the smallest distance measured, while on the STY distance this distance is similar to the one calculated on the thread-dependent distribution. The distance to a random distribution is, again, smaller on the STY dataset than on the MMC dataset.

### 3.5 Discussion

Regarding the three analyses performed, namely the analysis of content deletion among users, threads and relative locations in threads, we can conclude that on the MMC dataset all the distributions are further away from random distributions than on the STY dataset, hinting at a more careful supervision of comments on that dataset. However, we must be careful about drawing such a conclusion because a less random behavior might also point towards targeting specific users (e.g., previously misbehaving users), threads (e.g., threads on specific topics) or locations of the comment in the thread (e.g., the beginning of the thread).

With these three levels of analysis we have shown that the user-dependent distribution of deleted comments is the least random, followed by the thread-dependent distribution, with the location of the comment in the thread being closest to random. In other words, specific users seem to be the most filtered, followed by specific threads, with specific locations in the thread being least prone to filtering. The observation that comments of specific users are more prone to deletion than comments in specific threads is interesting and should be compared to the distributions in other datasets. Given that we observe the same phenomenon in two datasets of different origin, we assume that such regularity would hold in other datasets as well.

## 4 Availability of data and baseline model

### 4.1 Data availability

Both datasets are published on the CLARIN.SI repository with all the metadata pseudo-anonymised, and the text encrypted via a simple character replacement method using a random, undisclosed bijective mapping to comply with the terms-of-use of our data providers as well as to mitigate propagation of inappropriate content. The Slovene dataset, published together with the baseline model described in the following subsection, is available from http://hdl.handle.net/11356/1201, while the Croatian counterpart is available from http://hdl.handle.net/11356/1202.

### 4.2 Model availability

Given that the distributed datasets are encrypted, because of which only in-vitro experiments on inappropriate content identification can be run, but the final systems cannot be applied on real data, we have also published baseline models trained on the non-encrypted data. While these models are capable of identifying potentially inappropriate content, by sharing them we do not propagate inappropriate content above the token level.

For building the baseline models, we split each dataset into training, development and testing portions in a 8:1:1 distribution, and trained, tuned and evaluated fastText (Joulin et al., 2016) classification models on that data split. The published fastText models were trained on the full datasets.

When splitting the data into train, dev and test, we randomly shuffled on the thread level, ensuring that there is no spillover between threads, i.e., that there are no portions of the same thread to be found in train and dev or test data.

During tuning, we optimized the following fastText hyperparameters: word n-gram length (default is 1), minimum and maximum character n-gram length (no character n-grams are used by default) and number of epochs (default is 5). We optimize them by training on the train portion and evaluating on the dev portion. In Table 3 we give the results of the best-performing non-default values on each of the hyperparameters. We evaluate via the ROC AUC score, i.e., the area under the ROC curve, which is 0.5 in case of random results and 1 for perfect results where all positive instances are ranked higher than all negative instances. The presented results show that most impact can be obtained with adding character n-grams to the word, i.e., text representation procedure. By adding character n-grams of length 3 to 7 we lower the error rate on both datasets by 18%. Adding word n-grams longer than 1 does not have a positive impact on performance of fastText. Optimizing the number of epochs has a slight positive impact as long as other hyperparameters are kept default. If we combine optimal character n-gram and epoch hyperparameter values (last row in Table 3, there is no significant difference to using optimal character n-grams only. This is why we decided to use for our final setting all default hyperparameters, except for the character n-gram lengths that we set between 3 and 7.

We evaluate our final system setting on the test set and obtain a ROC AUC result of 0.794 for the MMC dataset and 0.793 for the STY dataset. These results are more than half way from random to perfect, which is still far from satisfying. We leave the task of building stronger prediction models to future work.

The baseline fastText models trained on the full non-encrypted datasets with optimal hyperparameter values can be obtained from the CLARIN.SI repository together with the encrypted datasets, as mentioned above.

| | MMC | STY |
|---|---|---|
| default (ngram=1;epoch=5) | 0.755 | 0.746 |
| ngram=2 | 0.717 | 0.711 |
| charngram=3,7 | **0.798** | **0.791** |
| epoch=3 | 0.762 | 0.753 |
| charngram=3,7;epoch=3 | **0.796** | **0.792** |

Table 3: Results of tuning hyperparameters on both datasets. Results are ROC AUC scores.

## 5 Conclusions

In this paper we have introduced two new large on-line news comment datasets annotated for inappropriate content by the content providers for Slovene and Croatian, languages typically rarely represented in similar datasets, making them all the more valuable for the research community.

We have performed a small manual analysis of the kinds of comments that get deleted in each of the datasets. The Croatian deleted comments contain more severe types of inappropriate content, such as insults, as well as more swearing. The Slovene ones, on the other hand, are more covert, formulated as irony, and are frequently aimed at the broadcaster or are off-topic, which indicates differences in the policy of handling user comments by the two media outlets.

The initial statistical analysis of the distribution of filtering among users, threads and locations in threads, has shown that all the distributions are less random on the MMC dataset than on the STY dataset, which is probably caused by more constant and vigilant moderation on the MMC RTV portal. Regarding the three levels of analysis, we showed that the distribution of deleted content among users is the least random, followed by the distribution of that content among threads, with the location of the deleted content being closest to random. This shows that specific users seem to be the most deleted, followed by specific threads, with specific locations in the thread being least prone to deletion.

Finally, by building baseline models on the new datasets, we have shown that fastText classifiers can be improved most easily by adding character n-gram information to the text representation. The obtained results are promising, but still far from production-ready, at least for most usages. However, we have published not only the datasets with metadata pseudo-anonymised and texts encrypted, but also the baseline models trained on the non-encrypted full text of each dataset. We expect that our baseline classification models will not serve just as a point of comparison, but will also be used in real world scenarios, either for feature extraction and text representation for similar tasks, or for the task of ranking or classifying inappropriate content directly.

## Acknowledgments

We would like to thank MMC RTV and Styria Media International for providing their databases of comments. The work presented in this paper was funded by the Slovenian Research Agency within the national basic research project Resources, methods and tools for the understanding, identification and classification of various forms of socially unacceptable discourse in the information society (J7-8280, 2017-2020).

## References

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *CoRR*, abs/1703.04009.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*.

F. J. Massey. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper Attention to Abusive User Content Moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1125–1135.

Aaditya Ramdas, Nicols Garca Trillos, and Marco Cuturi. 2017. On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19(2).

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. *CoRR*, abs/1701.08118.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1391–1399, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.