Ravi Shekhar, Marko Pranjić, Senja Pollak, Andraž Pelicon, Matthew Purver

Automating News Comment Moderation with Limited Resources: Benchmarking in Croatian and Estonian

Abstract

This article describes initial work into the automatic classification of user-generated content in news media to support human moderators. We work with real-world data — comments posted by readers under online news articles — in two less-resourced European languages, Croatian and Estonian. We describe our dataset, and experiments into automatic classification using a range of models. Performance obtained is reasonable but not as good as might be expected given similar work in offensive language classification in other languages; we then investigate possible reasons in terms of the variability and reliability of the data and its annotation.

1. Introduction

This article describes initial work on the EMBEDDIA project[1] into the automatic classification of user-generated content (UGC) in news media: reader comments posted under news articles. The EMBEDDIA project focuses on the use of cross-lingual techniques to transfer language technology resources to less-resourced languages (as well as English and Russian, the project focuses on Slovene, Croatian, Estonian, Lithuanian, Latvian, Finnish, and Swedish), and the application of these to real-world problems in the news media industry. One such problem is the need for news publishers to allow readers to post comments under articles online, in order to promote engagement with the content, but prevent content being published that would be offensive to other readers, dangerous or in some way compromise the legal position of the publisher. Most publishers currently use manual methods to do this: a team of moderators will monitor comments and block them when required. However, high volumes of comments can often make this impractical. The use of automatic natural language processing methods to detect comments that should be blocked, or referred to human moderators, can speed up the process many times (Pavlopoulos et al., 2017a); and many successful approaches to automated hate and offensive speech detection and categorisation exist (see e.g. MacAvaney et al., 2019; Schmidt & Wiegand, 2017), with datasets and shared tasks made available for several major EU languages (see e.g. Zampieri et al., 2019; V. Basile et al., 2019). However, such resources are generally only available for a few languages (e.g., English, German), leaving a gap for less-resourced languages. For Estonian and Croatian, languages of interest here, the number of studies is very limited (Ljubešić et al., 2018).

In this work, we describe new data collection efforts in two less-resourced European languages (Croatian, Estonian), and our experiments into automated classification. We explain the existing moderation scheme used by humans in news editorial houses, and examine to what extent it

[1] http://embeddia.eu

overlaps with the concept of offensive language as usually defined; describe a range of suitable classifier architectures for automatic detection of problematic comments; and give results showing that although reasonable performance can be achieved on these languages given suitable methods, it does not reach the levels that might be expected given other related work in languages in which more resources are available. We then examine the robustness of both the classifiers and the moderation scheme itself, and find that performance is limited not only by the nature of interactive language and its dependence on context, but by the need to rely on labels gathered under real-world constraints. We conclude that a transfer learning approach is the most promising future direction, providing the opportunity to incorporate information from more, better-curated datasets available in other languages, but that this will require cross-lingual techniques beyond the current state of the art.

2. Data and Task

The task of interest here, broadly defined, is to develop an automatic classifier to automate (or partially automate) the manual process of moderation: deciding which reader comments should be blocked, according to the policy of a particular newspaper.

2.1. Dataset

For this work, we have collected a large new dataset of online reader comments, from a range of news media sources in two less-resourced European languages, as covered by our project partners. Our dataset consists of over 60 million comments from the articles published online by three major news outlets:

- **24sata (`www.24sata.hr`):** The largest-circulation daily newspaper in Croatia, reaching on average 2 million readers daily.[2] Language: Croatian. Size: 21.5M comments.
- **Večernji List (`www.vecernji.hr`):** The third-largest daily newspaper in Croatia. Language: Croatian. Size: 9.6M comments.
- **Eesti Ekspress (`www.ekspress.ee`):** The largest weekly newspaper in Estonia, with a circulation of over 20,000. Languages: Estonian, Russian (articles are written in Estonian, but comments are often also in Russian). Size: 31.5M comments.

2.2. Annotation

In each case, the comments are annotated with metadata including link to the relevant article, ID of the comment author (anonymised) and timestamp; importantly for the purposes of this work, comments are also labelled if they are blocked by human moderators. Details of the moderation policy, and therefore the nature of the labelling, vary with news source, but comments may be blocked for a wide range of reasons. For 24sata, the annotation reflects a moderation policy based on 8 different categories, shown in Table 1; comments should be blocked if they breach any one of

[2] `https://showcase.24sata.hr/2019_hosted_creatives/medijske-navike-hr-2019.pdf`

these categories, although the implications for the comment author vary with the severity of the category. Less serious offences (labelled 'minor' in Table 1) lead to a minor warning: a user may receive up to two minor warnings, but the third one leads to a temporary one-day ban from the site. More serious offences lead to major warnings, of which a user may only receive one – the second one leads to a five-day ban. After a ban, the number of warnings of that type are reset to zero, but breaking the rules multiple times can, at the discretion of the moderators, lead to a permanent ban.

Rule ID	Description	Definition	Severity
1	Disallowed content	Advertising, content unrelated to the topic, spam, copyright infringement, citation of abusive comments or any other comments that are not allowed on the portal	Minor
2	Threats	Direct threats to other users, journalists, admins or subjects of articles, which may also result in criminal prosecution	Major
3	Hate speech	Verbal abuse, derogation and verbal attack based on national, racial, sexual or religious affiliation, hate speech and incitement	Major
4	Obscenity	Collecting and publishing personal information, uploading, distributing or publishing pornographic, obscene, immoral or illegal content and using a vulgar or offensive nickname that contains the name and surname of others	Major
5	Deception & trolling	Publishing false information for the purpose of deception or slander, and "trolling" - deliberately provoking other commentators	Minor
6	Vulgarity	Use of bad language, unless they are used as a stylistic expression, or are not addressed directly to someone	Minor
7	Language	Writing in other language besides Croatian, in other scripts besides Latin, or writing with all caps	Minor
8	Abuse	Verbally abusing of other users and their comments, article authors, and direct or indirect article subjects, calling the admins out or arguing with them in any way	Minor

Table 1: Annotation schema for blocked comments, 24sata.

As Table 1 shows, the categories cover a broad range of grounds for moderation, and many categories potentially overlap. They include a range of categories in the broad area of offensive language, many of which might overlap: threats to others (rule 2); hate speech based on national, racial, sexual or religious affiliation (3); obscene or immoral content (4); bad language (6); and verbal abuse (8). However, they also include a range of other reasons: illegal content (rule 1); comments not allowed by the portal's rules (1); advertising (1); off-topic posts (1); copyright infringement (1); false information (5); use of language other than Croatian (7).

Rule ID	Corresponding 24sata rule ID(s)	Definition	Severity
1	3	Hate speech on a national, religious, sexual or any other basis	Major
2	2	Threats to other users, administrators, journalists or subjects of articles	Major
3	6, part 4, part 8	Insulting other users or use of bad language.	Minor
4	part 4	Publishing personal data	Minor
5	part 1, part 7	Chat, off-topic, writing in all caps, posting links	Minor
6	part 7	Writing in a script other than a Latin script	Minor
7	part 8	Challenging the administrators or arguing with then in any way	Minor
8	part 5	Posting false information	Minor
9	n/a	Using multiple user accounts	Permanent ban

Table 2: Annotation schema for blocked comments, Večernji List, together with corresponding Rule IDs from the 24sata schema (Table 1).

Furthermore, as Table 2 shows, these categories also vary between publishers: the categories for Večernji List (hereafter VL) have many similarities with those for 24sata, but it is not possible to map directly between them. Categories such as hate speech and threats seem to correspond directly (rules 3 and 2 for 24sata, rules 1 and 2 for VL); but others are combined in different ways (e.g. 24sata's rule 5 covers posting false information, which maps to VL's rule 8, but also covers trolling and povocation which does not seem to be explicitly covered in VL's policy; VL's rule 3 covers insults and bad language, aspects of which are covered by parts of 24sata's rules 4, 6 and 8). Ekspress, on the other hand, do not record explicit categories of policy violation, so no such detailed annotation is available.

Three distinct problems therefore arise. First, distinguishing between the categories — rather than just detecting the general category of requiring moderation — is an important task in order to record how the policy was applied when blocking a comment or banning a user, where such a policy exists. Second, the overall category of blocked comments is likely to cover a very heterogeneous sample of language, as it results from a diverse range of phenomena. Third, as the categories are not *a priori* fixed, and can be conceptually divided up in different ways, this heterogeneity is likely to extend even to the individual classes.

Problematic comments are fairly common: for the 24sata subset, articles receive around 45 comments on average, and those that receive problematic comments receive around 5.5 of them. However, the data is highly unbalanced — only around 5-6% of comments require blocking — bringing an added complication to the classification task.

3. Related Work and Resources

In this section, we investigate what resources might be available which can help; in particular, what datasets might be available to provide training data for suitable classifiers.

3.1. Comment Filtering

Previous work in news comment filtering is limited. Pavlopoulos et al. (2017a) address the problem using data from a Greek newspaper, Gazzetta. They use a dataset of 1.6M comments with labels derived from the newspaper's human moderators and journalists; they test a range of neural network-based classifiers and achieve encouraging performance with AUC scores (area under the ROC curve) of 0.75-0.85 depending on the data subset. However, being in a different language (Greek) their data is not directly usable as a training set for our task. In addition, their moderation labels are binary, representing a "block or not" decision, rather than giving any further information about the reasons behind a decision. They are therefore not suited to investigating the moderation policy labels of interest here; and more fundamentally, it is unclear whether the decisions of Gazzetta's moderators are based on similar aims or policies as the decisions we must try to simulate for 24sata or Ekspress's moderators. Pavlopoulos et al. (2017a) asked additional annotators to classify comments according to a more detailed taxonomy (*"We also asked the annotators to classify each snippet into one of the following categories: calumniation (e.g., false accusations), discrimination (e.g., racism), disrespect (e.g., looking down at a profession), hooliganism (e.g., calling for violence), insult (e.g., making fun of appearance), irony, swearing, threat, other."*) but this was done as a post-hoc exercise and only for a small portion of the test set. It was not used in classification experiments, but only for separate analysis purposes.

Other work with reader comments on news (see Table 3) exists but does not attempt to learn from or reproduce moderation decisions directly in the same way. Kolhatkar et al. (2019) and Napoles et al. (2017) investigate constructivity in comments, and provide datasets which distinguish between constructive and non-constructive comments; these datasets are related to our task, though, as they also include information about toxicity and related categories such as insults and off-topic posting. Barker et al. (2016) investigate quality of comments and their use in summarisation. Wulczyn et al. (2017) investigate a related problem of detection of personal attacks and toxicity in user comments on Wikipedia articles, rather than news; and Zhang et al. (2018) also investigate Wikipedia comments from the point of view of detecting which conversations become toxic. None of these directly solve our problem, although they could in theory provide useful information; however, all are limited to English data.

3.2. Resources for Related Tasks

A variety of related tasks have been studied in data other than user-generated comments on articles. Given the moderation policy details in Section 2 above, the existence of suitable datasets for training classifiers for various categories of offensive language, advertising/spam, and trolling behaviour would be of interest. While none of these categories corresponds directly to the overall category of comments that must be blocked, each one covers a phenomenon that requires blocking.

Corpus	Location	Domain	Language	Size	Type of annotation
Gazzetta	(Pavlopoulos et al., 2017a)	News	gr	1.6M	Moderation
SFU SOCC	(Kolhatkar et al., 2019)	News	en	663k	Constructiveness, toxicity
YNACC	(Napoles et al., 2017)	News	en	522k	Constructiveness, insults, off-topic
SENSEI	(Barker et al., 2016)	News	en	2k	Quality, tone, summaries
DETOX	(Wulczyn et al., 2017)	Wiki	en	115k	Personal attacks, aggression, toxicity
Zhang et al., 2018	(Zhang et al., 2018)	Wiki	en	7k	Personal attacks

Table 3: Existing datasets for filtering user-generated comments on articles. Size is given in number of comments.

Corpus	Location	Domain	Language	Type of annotation
FRENK	(Ljubešić et al., 2019)	Facebook	en,sl	Socially unacceptable language
HASOC	hasoc2019.github.io	Twitter/Facebook	de, en, hi	Hate speech, target
HatEval 2019	(V. Basile et al., 2019)	Twitter	en, es	Hate speech, target, aggression
OLID (OffensEval)	(Zampieri et al., 2019)	Twitter	en	Hate speech, target, threats
GermEval	(Wiegand et al., 2018)	Twitter	de	Abuse, profanity, insults
IBEREVAL	(Anzovino et al., 2018)	Twitter	en,es	Misogynous
MEX-A3T	(Álvarez-Carmona et al., 2018)	Twitter	es-mx	Aggressive
Liu et al 2018	(Liu et al., 2018)	Instagram	en	Hostile
Waseem & Hovy 2016	(Waseem & Hovy, 2016)	Twitter	en	Hate speech, with subcategory
Stormfront	(de Gibert et al., 2018)	Online forum	en	White supremacy

Table 4: Existing datasets: abuse, hate speech and offensive language. "Target" refers to annotation of the group or individual towards which hate speech is directed.

3.2.1. Offensive Language Detection

Recent years have seen a large amount of research on detection of offensive language of various kinds. Many public datasets have been created and distributed, many shared tasks have been run, and many classification systems developed and tested (see Table 4). The exact definition of the categories annotated in these tasks varies, however (see Schmidt & Wiegand, 2017, for a survey), and may include one or all of:

- Threats: hostile speech intended to threaten the addressee with violence or other negative effects;
- Abuse: personal insults directed at others, including 'flaming' or cyberbullying;
- Hate speech: personal attacks on the basis of religion, race, sex, sexuality etc.;
- Offensive content: the use of language which is in itself considered rude, vulgar or profane (including pornographic), even if not targeted at someone in particular.

These terms are often used interchangeably, with some (particularly *hate speech*) often used to cover multiple categories. Exact definitions of the individual categories also vary with task and dataset, so we do not attempt an exhaustive exposition here. As an illustrative example, Waseem & Hovy (2016) define their *hate speech* category for Twitter as a message that:

1. *uses a sexist or racial slur;*
2. *attacks a minority;*
3. *seeks to silence a minority;*
4. *criticizes a minority (without a well founded argument);*
5. *promotes, but does not directly use, hatespeech or violent crime;*
6. *criticizes a minority and uses a straw man argument;*
7. *blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims;*
8. *shows support of problematic hash tags. E.g."#BanIslam", "#whoriental", "#whitegenocide";*
9. *negatively stereotypes a minority;*
10. *defends xenophobia or sexism;*
11. *contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.*

On the other hand, Ljubešić et al. (2019) use a more restrictive set of definitions via a decision tree to separate out different kinds of *socially unacceptable discourse (SUD)* on Facebook into different categories:

Is this SUD aimed at someone's background?
- YES: Are there elements of violence?
  - YES: background, violence
  - NO: background, offensive speech
- NO: Is this SUD aimed towards individuals or other groups?
  - YES: Are there elements of violence?
    - YES: other, threat
    - NO: other, offensive speech
  - NO: Is the speech unacceptable?
    - YES: inappropriate speech
    - NO: acceptable speech

In all these variants, the task is usually defined as a classification task — detecting whether a given text should be classified as hate speech (or abuse, offensive language etc.) or not — although this may be set up as a binary or a multi-class classification problem depending on the definitions used. Many datasets are available for this broad category of tasks, with a number of public shared

tasks having been run over the last few years.[3] The exact categories annotated vary, as do the domain and language of text annotated; we give an indication of each in Table 4.

Most datasets are based on social media (mainly Twitter) posts. Performance varies widely with dataset and domain. OffensEval 2019 reports maximum F1 score 0.829 on the offense classification task; for the white supremacy forum comments (de Gibert et al., 2018) classification accuracy is 0.78.

3.2.2. Spam detection

Another important task for UGC filtering in many domains, corresponding to one of the categories in the 24sata moderation policy in Section 2, is the detection of *spam*: comments which are off-topic, intended not to contribute to an ongoing conversation or relate to a given topic but rather to advertise, and/or to entice readers into clicking on a link either to generate revenue or for more nefarious purposes (e.g. 'phishing', attempting to gain access to personal information). This task is highly relevant for news media companies in order to prevent comments sections being taken over by irrelevant, offputting or dangerous content.

The task is a variant of the familiar spam detection problem for email (see Caruana & Li, 2012, for a survey), but UGC and online comments have their own distinctive characteristics – see for example (Kantchelian et al., 2012) for application to comments in the blog domain, (Aiyar & Shetty, 2018) in the Youtube domain, and (Wu et al., 2018) for a survey of work in the Twitter domain.

Corpus	Location	Size	Language	Domain
NSC Twitter Spam	(Chen et al., 2015)	6 million tweets	en	Twitter
Youtube Spam Collection	(Alberto et al., 2015)	1956 comments	en	Youtube
MPI-SWS	(Ghosh et al., 2012)	41,352 accounts	n/a	Twitter

Table 5: Existing datasets: spam.

Corpus	Location	Size	Language	Domain
FiveThirtyEight	(Linvill & Warren, 2018)	2,973,371 tweets	en	Twitter
Dataturks	(Narayanan, 2018)	20,000 tweets	en	Social media
Mojica 2017	(Mojica de la Vega & Ng, 2018)	5,868 conversations	en	Reddit

Table 6: Existing datasets: trolling and incitement.

Table 5 shows a sample of the most relevant datasets here. Alberto et al. (2015) provide a dataset of comments on Youtube videos classified as spam or not. Several datasets are available for short text messages in social media, see e.g. (Chen et al., 2015)'s large collection of 6 million spam tweets, and the MPI collection of Twitter accounts detected as spam accounts. Again, this task is usually defined as a binary classification task. Performance varies widely with dataset and

[3] A helpful catalogue of relevant datasets is also available online at `http://hatespeechdata.com/`.

domain. Wu et al. (2018) report accuracies of up to 94.5% on account classification and 88-91% accuracy on individual texts.

3.2.3. Trolling and incitement

Another basis for moderation in the policy of Section 2 is the presence of *trolls* and *bots*: users who may be automated or semi-automated rather than human, and which behave in a disruptive and/or deceptive manner in order to influence discussion, spread propaganda and manipulate opinion or to incite extreme views and disrupt discussion (see e.g. Kim et al., 2019). The effects of such agents in social media and news article comments can be strong, with evidence that they have affected public opinion and outcomes of elections (Badawy et al., 2018). There is a connection with the *fake news* phenomenon, with many trolling accounts being used to spread false rumours and link to fake news.

In this case, although this can be approached in a similar classification manner to the tasks above, labelling texts as coming from trolls, the problem is more often seen as one of classifying user accounts rather than their individual text outputs. Methods used therefore often depend as much on the social network properties of user accounts as on the language they generate. Again, some datasets exist; see Table 5. FiveThirtyEight distribute a dataset of nearly 3 million tweets sent from Twitter accounts *"connected to the Internet Research Agency, a Russian "troll factory" and a defendant in an indictment filed by the Justice Department in February 2018"* between February 2012 and May 2018. Narayanan (2018) then provides a smaller dataset from the same source, but annotated in more detail for level of aggression. Mojica (2017); Mojica de la Vega & Ng (2018) collected a similar dataset of comments on Reddit.

In our domain of UGC comments under news articles, Mihaylov & Nakov (2016) collected a dataset from over 2 years of articles (Jan 2013-April 2015) on the Bulgarian news site Dnevnik (`dnevnik.bg`), totalling 1,930,818 comments by 14,598 users on 34,514 articles. Troll comments were identified by a combination of observing other users' reactions, and checking identities in leaked documents; however, the dataset is not currently available publicly.

Mihaylov & Nakov (2016) achieve around 81% accuracy and F-score on the classification task, on a balanced dataset of news comments, using simple baseline linear classifiers. Mojica (2017) achieves c.90% accuracy on his dataset for the trolling detection task, using a more complex conditional random field classifier.

3.3. The Problem of Monolinguality

As the discussion above shows, datasets are available. However, very few are in the exact domain of automatic moderation: the Gazzetta dataset of (Pavlopoulos et al., 2017b) is the only example from news, with the Wikipedia dataset of (Wulczyn et al., 2017) being quite closely related. More critically, none are available in the languages required here (Croatian, Estonian); the closest are the Facebook dataset of socially unacceptable discourse in Slovenian of Ljubešić et al. (2019), and the Bulgarian news comment trolling data of Mihaylov & Nakov (2016), but neither are publicly available and neither are in the exact domain required.

This problem is a widespread one in NLP: a large majority of research and available datasets is monolingual and in English, and datasets for specific less-resourced languages like Croatian and Estonian are hard to find. Some multi-lingual work exists: Ousidhoum et al. (2019) present a multilingual hate speech study on English, French and Arabic tweets, and A. Basile & Rubagotti (2018) conduct cross-lingual experiments between Italian and English; again, this does not cover our languages or domain.

We also note the existence of Hatebase,[4] a highly multilingual collection of crowdsourced social media posts; however, as its annotation is based only on submission by the public, and it contains no comparable non-abuse language, it is not currently suitable as training or evaluation data for a classifier of the kind needed here.

We therefore conclude that for our present purposes, training on the specific data we have, in the correct language and reflecting the moderation policy of the correct newspaper, is the only practical option. The next section outlines our experiments using this approach.

4. Experiments

Our approach is therefore to treat the task as a classification problem, and use the real-world moderator decisions, recorded in the newspaper databases, as our training and test labels.

4.1. Classification Models

We formulate the problem as a text classification task. The basic task is a binary choice: given a comment, a system has to predict whether it should be *blocked* or *non-blocked*. We can also consider a multi-class task: given a comment, to predict which rule (Table 1 or Table 2) is being violated. We compared four different models, each using a standard method for text classification.

Naïve Bayes As a baseline, we use a Naïve Bayes (NB) classifier. NB is a simple probabilistic generative model which makes the approximation that words are independent of one another: the probability of a text belonging to a particular class can therefore be approximated as the product of the probabilities of the individual constituent words being associated with that class, and those can be calculated directly from frequencies in the training set. While clearly an oversimplification, this approach can provide good results in many text classification tasks, including spam detection (see e.g. Jurafsky & Martin, 2009). It also provides an easily interpretable model: a conditional probability table relating each word to each class.

LSTM In this model, the comment is encoded using a Long Short-Term Memory (LSTM) recurrent neural network (Hochreiter & Schmidhuber, 2015): LSTMs are able to encode not only word sequence but capture dependencies between non-adjacent words. The last hidden state of the LSTM is taken as the representation of the comment, and on top of that, a multi-layer perceptron (MLP) is used to produce the classification decision. Word embedding vectors are randomly initialised, and the whole architecture is trained end-to-end.

[4] http://hatebase.org/

LASER In this model, the comment is represented using Language-Agnostic SEntence Representation (LASER, Artetxe & Schwenk, 2019). LASER produces representations for sentence-length texts, obtained using a five-layer bidirectional LSTM (BiLSTM) encoder with a shared byte-pair encoded (BPE) dictionary for 92 languages. The last states of the LSTM are used to produce a sentence vector by max-pooling, and the model is trained using an encoder-decoder approach, in which the sentence representations are used to generate parallel sentences in another language. This approach gives sentence vectors which capture many aspects of sentence meaning and can be used in many tasks; here, we use a MLP on top of the sentence representations, and train it on our classification task. Only the MLP is trained; the weights of the LASER encoder are kept frozen using the pre-trained models available.[5]

mBERT In our final model, the comments are represented using Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2018). BERT is a deep contextual representation based on a series of layers of Transformer cells (Vaswani et al., 2017), and trained using a variant of a language model objective. As with LASER above, we then pass the comment representation to a MLP for classification. The BERT model weights are initialized using the multilingual pre-trained model (mBERT, trained on 104 languages by sharing embeddings across languages), and fine-tuned end-to-end along with the MLP.[6]

Training Note the difference in the training strategy for our LSTM, LASER, and mBERT models. In the case of LSTM, the whole architecture is initialized randomly and trained end-to-end: we use no pre-trained embeddings, and train only on the data available here. In the case of LASER, only the classification MLP weights are trained, while the LASER model sentence (comment) representation weights are kept fixed at the values in the pre-trained model. For mBERT, the comment representation weights are initialized using the pre-trained model, and the MLP weights initialized randomly, and the whole model is then fine-tuned end-to-end. All the neural models are trained using the Adam optimizer (Kingma & Ba, 2014) with cross-entropy loss.

4.2. Experiment 1: Binary Classification

4.2.1. Data Selection

As Figure 1 shows, the rate of commenting on articles, and the rate at which moderators block comments, vary over time. (Detailed frequency counts are given in Appendix A, Section A.1). For Ekspress, the rate of commenting rises steadily over time; for 24sata, it rises to a peak in 2015/2016 and then reduces slightly. For VL, the commenting rate seems more stable. (Note that the data was collected part-way through the year 2019, so data for that year is not for a complete year period). Particularly of note, though, is that the rate at which moderators block comments rises over time for all newspapers; the effect is particularly marked for VL from 2013 onwards, and for 24sata from 2016 onwards. Note that the rates for VL before 2013, and 24sata before 2016, are not zero, but very low; see Appendix A for details. This effect is not merely one of

[5] Pre-trained model available from `https://github.com/facebookresearch/LASER`.
[6] Pre-trained model available from `https://github.com/google-research/bert`.

comment volume: higher commenting rates do not correspond to higher blocking rates (Figure 1), as might be hypothesized if, say, a rise in commenting rates were caused by a sudden influx of troll accounts or an increase in contentious topics. Instead, the most likely cause is a change in moderation policy: over recent years, more attention has been given by newspapers to moderation, in terms of both overall importance and strictness of adherence to policy. Note also that blocking rates are relatively low in general: even the peak rate for VL is only just over 15% of comments, for Ekspress 12.5%, and for 24sata only 7.8%: this gives an unbalanced dataset which must be accounted for in training and testing.

Given the sharp change over time, it seems very likely that data from more recent years will be more consistent, and will be more reflective of current moderation policy: earlier years are likely to contain large numbers of false negatives (comments that were not moderated at the time, due to either lack of resources or difference in policy, but would be blocked now). In order to have the cleanest and most relevant data possible, we therefore first selected 2019 data for training, validation, and testing purposes. Since most comments are non-blocked comments, to have a balanced dataset for experiment purposes, we first selected only those articles which have at least one blocked comment. We then divided those articles into training (80%), validation (10%) and test (10%) partitions. Finally, we randomly selected an equal number of blocked and unblocked comments per article in each set. Table 7 shows the resulting data distribution for all three newspapers.

	24sata			Večernji List			Ekspress		
	Train	Val	Test	Train	Val	Test	Train	Val	Test
# Articles	9196	1148	1154	6521	813	821	7490	934	942
# Comments	99246	12364	12472	85916	10490	10855	145154	19310	20312

Table 7: Partitioned dataset distribution, 24sata, Večernji List and Ekspress.

4.2.2. Results

Table 8 shows the results for each classifier model. As our training and test sets have an evenly weighted number of positive (blocked) and negative (non-blocked) examples, we give performance as standard percentage accuracy, and to get an insight into the relative performance we give this not only overall but over the positive and negative portions of the test set individually. 'Blocked' accuracy is therefore equivalent to recall for the positive (blocked) class; 'Non-blocked' accuracy is recall for the negative (non-blocked) class. Standard summary measures such as weighted average F-score are not very helpful in this setting, as they can be so strongly dominated by the majority (non-blocked) class, and accuracy on the two classes has different implications for news publishers; we therefore examine per-class metrics (although see Section 4.4 for results in terms of macro-averaged F-score on the final dataset).

For all three newspapers, the mBERT model gives best performance. Surprisingly, the NB model gives relatively strong performance, with neither the LSTM nor LASER models providing much of an improvement; in fact, for Ekspress they perform worse than NB. Accuracy is higher for 24sata than for Ekspress and VL, but in all cases the absolute level of accuracy is lower than might

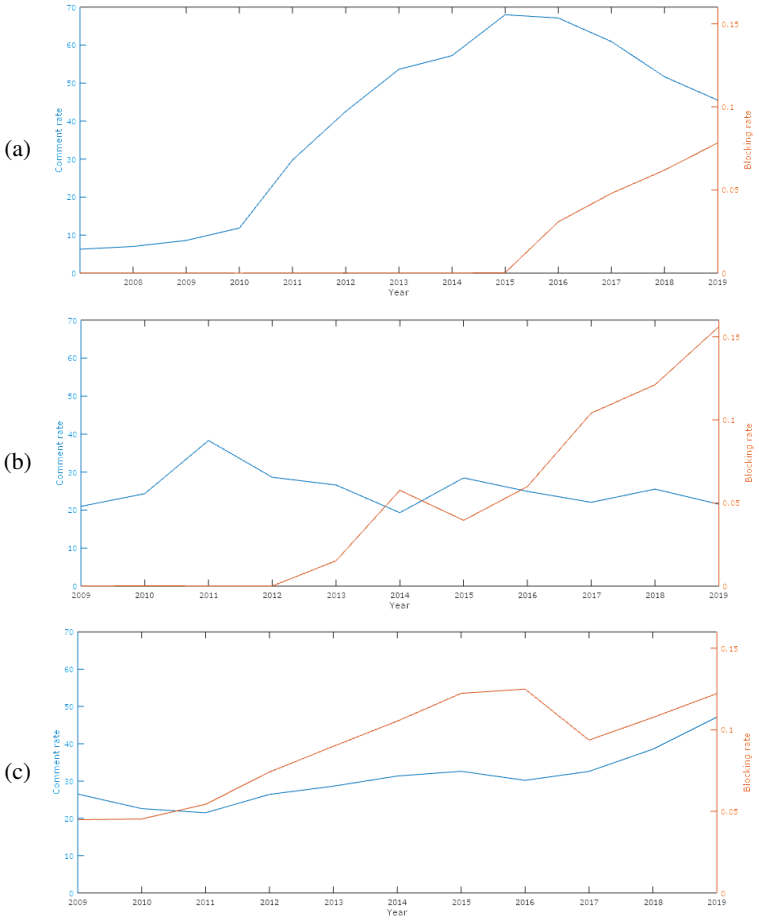


Figure 1: Comment rate $N_{\text{comments}}/N_{\text{articles}}$ in blue, and blocking rate $N_{\text{blocked}}/N_{\text{comments}}$ in red, over time, for (a) 24sata, (b) Večernji List, (c) Ekspress.

be expected given comparable experiments with offensive language detection in other research (Section 3). Accuracy on blocked content is lower than the accuracy of recognition of non-blocked content, particularly for Ekspress.

To calculate the performance that would be expected on real (unbalanced) data, we must take into account the expected real ratio of blocked to non-blocked comments. As Section 2 discusses, blocked comments are rarer than non-blocked, with the most recent estimate of the ratio from 2019 being 0.078 for 24sata. In practice, we would therefore expect for 24sata a recall of 0.67, a

precision of 0.27 and an F-score of 0.38. In other words, the classifier would successfully detect 67% of comments that needed blocking (missing 33%), but 73% of its decisions to block would be false positives; and nearly 15% of innocent comments would be falsely blocked. While this level of performance is potentially useful, it seems it would still require significant manual filtering on the part of moderators. The balance between recall and precision could of course be tuned via the decision boundary, or by weighting the objective function in training, but gains in the recall would correspond to losses in precision, and vice versa (see Pavlopoulos et al., 2017a).

	24sata			Večernji List			Ekspress		
Model	ALL	BLK	NON	ALL	BLK	NON	ALL	BLK	NON
NB	69.43	47.59	**91.26**	66.39	49.75	81.79	64.57	46.48	82.66
LSTM	71.52	61.70	81.33	65.39	54.47	75.50	63.02	41.96	**84.09**
LASER	70.74	**70.11**	71.36	63.31	**59.77**	66.59	61.58	47.07	76.10
mBERT	**76.42**	67.33	85.49	**69.63**	53.18	**84.87**	**68.40**	**58.46**	78.34

Table 8: Classifier performance, as percentage accuracy. Columns are labelled ALL for all comments, BLK for positive instances only (blocked content), NON for negative instances only (non-blocked content).

Inspection of the conditional probability table produced by the NB model allows us to determine the words which are most strongly associated with the blocked and non-blocked classes, on the basis of the ratio of class probabilities. Tables 21 and 22 in Appendix B show full lists of the top 100 words for each class for 24sata. The strongest indicators for the blocked class correspond to vocabulary expected in spam comments: external URLs (*www*, *com*, *google*, *posjetite* (visit)); work and earnings (*poslu/posla* (work), *plaća* (payment), *zaradio/zaraditi* (earn)); amounts of money promised (numbers, *dolara* (dollars), *eura* (euros), *mjesecu* (monthly), *tjedno* (weekly), *dnevno* (daily)). Vocabulary associated with offensive language is also included, but comes further down the list (*jebem/jebo* (fuck), *majmun* (monkey)). Non-blocked indicators include vocabulary associated with discussion of a range of news topics (e.g. football: *inter*, *derbi*) and general evaluative words (*sretno/sritno* (happy/good luck), *predivno* (amazing), *najljepša* (most beautiful), *strašno* (terrible)). However, of a list of 185 blacklisted words used by the moderators at 24sata to flag comments for blocking, only 78 appear in the top 1000 in the NB model; and surprisingly, many words that one might expect to be associated with offensive or highly-charged language (although no blacklisted words) appear in the top 1000 non-blocked indicators in the NB model: *svastiku* (swastika), *terorizam* (terrorism), *trolaš* (you're trolling).

Vocabulary indicators extracted from these annotations are therefore not straightforward, suggesting that the data is fairly heterogeneous: comments may be blocked for many diverse reasons, and therefore display very different textual features. This may be one possible reason for the below-par performance; our next experiment investigates this.

4.3. Experiment 2: Blocking Rule Classification

For 24sata and VL, the publisher's database records the reason behind the moderators' decisions: the specific rule that a comment breaks. Here, we train and test multi-class versions of our classifier models for the problem of rule recognition.

(a)

	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Rule7	Rule8
Train	24329	20	2167	30	2912	992	387	18786
Val	3081	1	216	1	271	114	41	2457
Test	2962	1	248	2	388	134	57	2444

(b)

	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Rule7	Rule8	Rule9
Train	3652	6548	4514	57	3756	9	156	4	24322
Val	572	794	547	7	402	0	13	0	2914
Test	553	864	580	4	456	2	24	0	2951

Table 9: Blocking rule dataset distribution, for (a) 24sata and (b) Večernji List.

Table 9 shows the distribution of blocked comments by rule within the training, validation and test sets defined above. The distribution is very uneven: for 24sata, rules 1 (unrelated topics, spam, advertising etc.) and 8 (abuse, arguing with administrators) are common, while rules 2 (direct threats) and 4 (obscenity) are extremely rare; others are in between. For VL, rule 9 (using multiple accounts) is most common, with rules 4 (publishing personal data), 6 (using non-Latin script) and 8 (misinformation) very rare. Even for rules which seemingly map directly between the two schemata (e.g. hate speech: 24sata rule 3, VL rule 1; threats: 24sata ule 2, VL rule 2) the distributions seem to vary widely across newspapers: it seems to be very rare for 24sata moderators to class comments as threats, but quite common in VL.

One hypothesis might be that moderators tend to avoid applying rules with more serious consequences if other less serious ones could be used (see Tables 1 and 2); but while this might explain the rarity in 24sata of rules 2 (threats) and 4 (obscenity), it does not explain the distribution in VL, where rules 1 (hate speech), 2 (threats) and 9 (multiple accounts) are all commonly used. It may be that the ambiguity of many rules, together with the cultural practices and habits within particular groups of moderators, have significant effects here.

Results Table 10 shows the results for individual rules, with Table 11 showing the effect this would have on overall blocking accuracy (comments which break any rule should be blocked).

Performance for individual rules varies widely. Less frequent rules are often ignored by all classifiers (rules 2, 4), with better performance for more frequent rules (e.g. rules 1, 8). It is likely that the lower contribution of the less frequent classes to the training objective function means that not enough weight is given to them in the final classifier models. The NB model does much worse than other models, presumably because the pruning of the conditional probability table favours more common words, likely to be significant indicators of the more common classes. The simpler LSTM model seems to have an advantage over the more complex LASER and BERT models, in that accuracy seems more even across classes; this may be because the pre-training of the LASER and BERT models gives them less ability to adjust to the different classes in fine-tuning.

However, the overall performance is not strongly affected. Given the real blocking rate, for 24sata we would expect a recall of 0.48, a precision of 0.32 and an F-score of 0.39. This translates to successfully detecting 48% of comments that needed blocking (missing 52%), while producing 68% false positives; and blocking nearly 8% of innocent comments. Note that the F-score is very

(a)

Model	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Rule7	Rule8
NB	43.01	0	3.23	0	5.67	5.97	8.77	2.74
LSTM	62.42	0	56.05	0	50.52	75.37	43.86	57.53
LASER	51.25	0	9.68	0	1.55	16.42	0	50.12
mBERT	48.68	0	0	0	0	0	0	63.3

(b)

Model	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Rule7	Rule8	Rule9
NB	6.61	5.47	4.56	0	6.4	100	4.55	0	33.73
LSTM	25.73	20.64	33.65	50	35.22	0	13.64	0	40.41
LASER	51.39	45.26	67.49	66.67	61.37	0	63.64	0	57.85
mBERT	0	0	43.54	0	0	0	0	0	42.01

Table 10: Blocking rule classifier performance, measured as percentage accuracy, (a) 24sata (b) Večernji List.

Model	Overall	Blocked	Non-blocked
Chance	11.11	11.11	11.11
NB	60.06	22.19	**97.93**
LSTM	**71.78**	**59.59**	84.16
LASER	67.09	44.82	89.35
mBERT	70.04	47.93	92.19

Table 11: Performance of multi-class rule classifier on binary task, measured as percentage accuracy, 24sata.

similar to the classifier trained on the binary task, although the balance between precision and recall is different; this could be adjusted as discussed above.

To investigate the role of the multi-class objective function in training, we also checked the coverage of the classifiers trained on the binary task in Section 4.2 above. While these classifiers give only binary output and therefore cannot help moderators understand decisions, we can compare their ability to detect the individual rules. Table 12 shows the results. The very rarest classes (rules 2, 4) seem to behave quite randomly (given the very low counts, this is not surprising), but the slightly more common rules (6 and 7, then 3 and 5) get reasonable accuracy for most classifiers. The picture is mixed, however: some classes seem to be inherently hard to detect, with rules 5 (trolling) and 7 (non-Croatian language) getting relatively low scores for all classifiers.

Model	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Rule7	Rule8
NB	52.77	0	45.56	0	27.84	71.64	22.81	43.99
LSTM	63.37	100	61.29	50	52.84	79.85	56.14	60.27
LASER	71.0	100	69.76	100	58.25	84.33	42.11	70.79
mBERT	64.15	0	72.18	100	54.64	88.06	36.84	72.3

Table 12: Performance of binary classifier per blocking rule, measured as percentage accuracy, 24sata.

4.4. Experiment 3: Variation over Time

Another possible reason for variable performance is the reliability and/or variability of the moderation annotation itself. Moderation can be quite a subjective decision, and the large amounts of data to mean that many blockable comments may be missed. One way to test this is to examine how classifier performance changes over time, as the moderation policy and the amount of effort put into moderation changed over the years (see Section 4.2.1); for this experiment we focus on just one dataset, 24sata. The distribution of individual blocking rules also varies over time: Figure 2 shows the proportion of blocking decisions based on each rule for the last four years (the years with most data). (Full details of the rule distributions over time for both 24sata and VL are given in Appendix A, Section A.2). Significant changes can be seen in the proportions. Some changes may reflect changes in behaviour: for example, rule 1 (advertising/spam) is used progressively more over time. However, other changes may be more complex: the commonly used 8 (abuse) becomes less used over time, with related rarer classes such as 2 (threats) and 5 (trolling/provocation) increasing. It therefore seems likely that rules are being applied differently in different cases: with many rules covering a range of phenomena and many phenomena being covered by multiple rules (see details of the rules in Table 1 above), moderators have a choice in which rules to apply, and perhaps more specific rules (often with more stringent penalties) are becoming preferred.

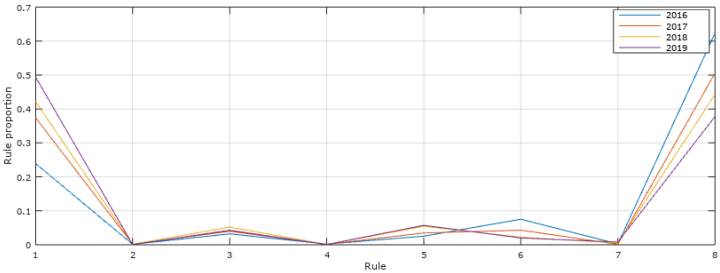


Figure 2: Blocking rule proportion over time, 24sata.

To determine the variability of the models' performance over different years' data, we therefore created a series of test sets, one for each of the last four years. We keep the same training set, taken from 2019 data (see above); the 2019 test set is therefore smaller and based on that used in the previous section. The test sets for 2016-2018 are larger as they can contain all the year's data labelled with rules; as the training set is fixed we can also test on a realistic balance of data, using all the blocked and non-blocked comments available for each year. Table 13 shows the test set distribution over time.

Results Table 14 shows overall accuracy figures per year on the 24sata dataset; we show only performance for the best classifier model, mBERT. Accuracy decreases as we move further away from the year 2019 used in training. Table 15 then shows how the accuracy of the binary blocking

	Articles	Non-blocked	Blocked	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Rule7	Rule8
2016	907	196762	15154	2915	111	992	183	683	1413	227	8630
2017	1045	188639	20579	6351	185	1560	153	1273	1211	137	9709
2018	1678	285620	21838	237	254	2800	125	2616	840	780	14186
2019	1154	68706	6398	3070	3	256	2	396	138	58	2475

Table 13: Yearwise dataset distribution, 24sata.

Year	Overall	Blocked	Non-blocked	F1-macro	Recall (BLK)	Precision (BLK)
2016	72.25	72.20	72.89	54.19	0.73	0.15
2017	75.17	76.16	64.84	58.10	0.65	0.21
2018	76.75	78.36	61.32	59.59	0.61	0.23
2019	80.03	81.19	67.32	62.07	0.67	0.25

Table 14: Binary classification performance over the yearwise testset using mBERT, 24sata. Figures are shown as percentage accuracy overall and for the blocked and non-blocked content separately; as this experiment uses the full data for each year (rather than a balanced subset) we also give F1 score macro-averaged over the two classes, and recall and precision for the blocked class only.

classifier varies with blocking rule class: while figures for many rules decrease in years before 2019, performance for rules 3 (hate speech), 6 (vulgarity) and perhaps 8 (abuse of other users, authors and admins) seems to remain relatively steady. Performance for rule 2 (threats) and rule 7 (non-Croatian language) may even be improving, although these rules have smaller amounts of data. Some of the main categories that relate to offensive language therefore seem to remain relatively consistent, while other categories such as advertising, spam and distribution of obscene content may be changing more. This may be because topics and vocabulary change over time; because authors change their language to avoid detection; because moderators change their criteria and behaviour; or a combination of these factors. What seems clear is that change over time is a significant issue: the ability to re-train classifiers on new data and up-to-date moderation labels will be important in practice.

5. Discussion and Conclusions

In this section we discuss the possible reasons for the overall levels of performance observed, and draw conclusions about what steps can be taken to improve it.

Year	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Rule7	Rule8
2016	52.37	65.00	75.85	46.07	46.77	93.51	63.96	78.62
2017	49.36	76.92	70.27	51.68	46.99	85.71	71.21	73.34
2018	50.67	83.54	71.74	42.74	37.74	90.93	38.20	68.73
2019	64.23	66.67	72.18	100.00	54.36	88.32	35.85	72.17

Table 15: Blocking rule classification performance over the yearwise testset using mBERT, measured as percentage accuracy, 24sata.

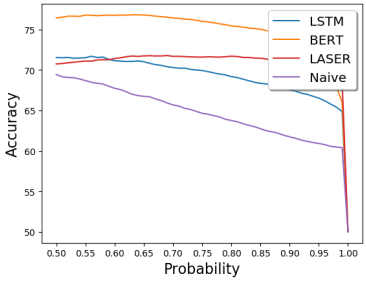
5.1. Analysis of Classifier Outputs

Figure 3 shows the confidence of the different classifier models: the plots are generated by changing the decision threshold of each classifier, increasing from the default 0.5 up to 1.0, and calculating the classification accuracy on the standard 24sata test set of Section 4.2. This is shown for blocked comments in Figure 3a, for non-blocked comments in Figure 3b, and the overall average in Figure 3c. The BERT and LASER models show overall higher confidence: increasing the threshold at which the decision is made has less effect on the accuracy of their output. The NB and to a lesser extent LSTM models' performance drops off more quickly, showing that their outputs give lower confidences for many correct classification decisions. Interestingly, classifier confidences seem significantly higher for blocked comments: the dropoff in performance is much less than that for non-blocked comments as the threshold increases. Although its performance was generally lower, the LASER model may provide some advantages here: its confidence curve is flatter with less dropoff for non-blocked comments.

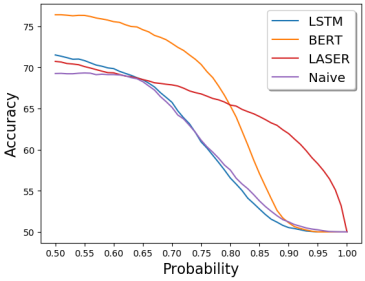
This general tendency suggests that non-blocked comments are harder to classify in many cases. This may be due to variability or lack of reliability in moderation, with many comments that should be blocked labelled as non-blocked. Classifiers would therefore be learning decision boundaries that fit these examples where possible, but having to leave them close to the boundary given their similarity to other blocked comments.

Manual inspection of classifier errors was carried out over a set of approximately 350 comments on which the best (mBERT) classifier output disagreed with the moderator's decisions. These comments were passed back to 24sata's moderators, who were asked to moderate them again and produce a new set of labels. Of 101 comments which were originally not blocked, the majority (82) were still not blocked, but with a significant proportion (19) now marked as blocked. The problem of moderators missing comments which should be blocked is therefore a real one, as suspected. However, a bigger effect may be the variability of moderation decisions. Of 244 comments which were originally blocked (but given a non-blocked label by our classifier), approximately half (124) were still judged to be blocked, but half (120) were now marked non-blocked. Of the 124 which remained blocked, over half (81) were given a different rule as justification for blocking.

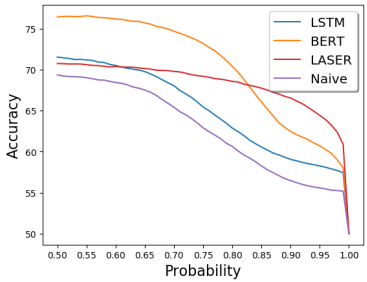
Examination of the errors also helps shed some light on the phenomena which cause difficulties for automatic classification. Some examples show classic language processing problems: non-standard spelling and vocabulary, and complex references and indirect statements can all be hard for classifiers to recognise without extremely large training sets. Two particular phenomena emerge as covering a large proportion of examples, however. One is that reader comments occur in the context of the article and the preceding comments, and many references need that context to be understood (see example (1), in which the phrase "that symbol" refers to an important concept from the previous discussion, probably the swastika. Treating comments as independent texts (as we do here) misses this – without the reference, it is hard to understand the comment as problematic. The second is that many comments use culture- and country-specific references which must also be resolved before the stance of the comment is clear. Example (2) appears on the face of it as a political trolling attempt; but if one knows that the HDZ and SDP are not only opposing political parties, but the only two large parties in Croatia, it can be understood as even-handed. In example (3), one must know that Pavelić headed a fascist government, and that



(a) Blocked Comments.



(b) Non-Blocked comments.



(c) Both comments.

Figure 3: Confidence of the Classifier.

Tuđman founded the currently governing, right-of-centre HDZ, in order to see its provocative nature.

(1) U čemu je problem? Dotični je pod tim simbolom živio i djelovao.
What's the problem? The person in question lived and worked under that symbol.
Moderator decision: blocked, rule 8

(2) HDZ je proslost a i Sdp !
HDZ is the past, and so is the SDP!
Moderator decision: not blocked

(3) Naime, preko natpisa "Franjo Tuđman, prvi hrvatski predsjednik"... Profesor Milan Kangrga je u emisiji NU2 rekao da je prvi hr pred bio Ante Pavelić.
Namely, via the inscription "Franjo Tuđman, the first Croatian president" ... Professor Milan Kangrga said on the NU2 show that the first Croatian president was Ante Pavelić.
Moderator decision: blocked, rule 8/rule 5 (moderators disagree)

5.2. Conclusions and Further Work

The high levels of variability in moderation decisions, and in the justifications given for them according to the moderation policy, indicate that an iterative approach may be of benefit in this task. Working with moderators to jointly define a more reliable policy, based partly on observation and use of high-confidence classifier outputs as in the error analysis above, would allow us to work towards less noisy data together with more reliable and useful classifiers. This could be framed within a general active learning approach, and we hope to explore this in future work. However, working within a real-world setting constrains the time and resources that can be dedicated to such work; great care must be taken to find an approach which does not further burden moderators and news publishers.

Second, the use of moderation flags as training labels, as pursued here and in other related work (Pavlopoulos et al., 2017a), may not be the most practical way to proceed in order to produce an accurate classification tool. A more effective and reliable way may be to use other, better-understood and curated datasets which represent the categories of language and author behaviour which should be blocked. By training classifiers on these cleaner datasets, a more reliable set of classifier outputs may be obtained which can feed into an active learning approach as outlined above. However, as Section 3 explains, such datasets are simply not available in the languages of interest here (Croatian and Estonian), or in many other language other than the majority well-resourced languages such as English, German and Spanish. One helpful step might be to pre-train word embeddings and/or models on data in the target language, even if annotated data is not available, to help smooth the noise from the training set; but note that the LASER and BERT models used here already benefit from large amounts of multi-lingual data, and in any case this is unlikely to go far towards solving the problem. Cross-lingual approaches (Ruder et al., 2017) would therefore be of great benefit if they can permit transfer learning from well-understood datasets in better-resourced languages to tasks in less-resourced languages.

However, while some work in hate speech and offensive language detection has been multi-lingual, studying datasets in more than one language, cross-lingual work is rare. A. Basile & Rubagotti (2018) use a *bleaching* approach (van der Goot et al., 2018) to conduct cross-lingual experiments between Italian and English in the EVALITA 2018 misogyny identification task, and Pamungkas & Patti (2019) propose a cross-lingual approach using a LSTM joint-learning model with multilingual MUSE embeddings. However, as far as we are aware, no work has yet tried to apply this to the problem of comment filtering, or focused on the languages needed here. As our error analysis shows, the task here poses significant challenges for cross-lingual techniques: many phenomena of interest are dependent on region- or culture-specific references and understanding of the related context, as in the need to understand country-specific relations between political parties and individuals discussed in the previous section. Current cross-lingual techniques depend on parallel corpus training, or on mapping of embedding spaces based on known synonymous anchor points (e.g. digits); these are unlikely to capture such phenomena well. Our next steps will therefore be to adapt techniques for cross-lingual learning to try to better map the entities, events and similar references found in news text between languages.

6. Acknowledgements

This research is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The work of AP was funded also by the European Union's Rights, Equality and Citizenship Programme (2014-2020) project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, grant no. 875263). We acknowledge also the funding by the Slovenian Research Agency (ARRS) core research programme Knowledge Technologies (P2- 0103). The results of this publication reflect only the authors' views and the Commission is not responsible for any use that may be made of the information it contains.

References

Aiyar, S., & Shetty, N. P. (2018). N-gram assisted Youtube spam comment detection. *Procedia Computer Science*, *132*, 174 - 182. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1877050918309153` (International Conference on Computational Intelligence and Data Science) doi: https://doi.org/10.1016/j.procs.2018.05.181

Alberto, T., Lochter, J., & Almeida, T. (2015, December). TubeSpam: Comment spam filtering on YouTube. In *Proceedings of the 14th ieee international conference on machine learning and applications (icmla'15)* (p. 1-6).

Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H. J., Villasenor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. In *Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain* (Vol. 6).

Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on Twitter. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, & F. Meziane (Eds.), *Natural language processing and information systems (NLDB)* (Vol. 10859, p. 57-64). Springer.

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610.

Badawy, A., Ferrara, E., & Lerman, K. (2018, Aug). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *Ieee/acm international conference on advances in social networks analysis and mining (asonam)* (p. 258-265). doi: 10.1109/ASONAM.2018.8508646

Barker, E., Paramita, M. L., Aker, A., Kurtic, E., Hepple, M., & Gaizauskas, R. (2016, September). The SENSEI annotated corpus: Human summaries of reader comment conversations in on-line

news. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue* (pp. 42–52). Los Angeles: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W16-3605 doi: 10.18653/v1/W16-3605

Basile, A., & Rubagotti, C. (2018). Crotonemilano for ami at evalita2018. a performant, cross-lingual misogyny detection system. In *Evalita@ clic-it.*

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., ... Sanguinetti, M. (2019, June). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proc. semeval* (pp. 54–63). Retrieved from https://www.aclweb.org/anthology/S19-2007 doi: 10.18653/v1/S19-2007

Caruana, G., & Li, M. (2012, March). A survey of emerging approaches to spam filtering. *ACM Computing Surveys*, *44*(2), 9:1–9:27. Retrieved from http://doi.acm.org/10.1145/2089125.2089129 doi: 10.1145/2089125.2089129

Chen, C., Zhang, J., Chen, X., Xiang, Y., & Zhou, W. (2015, June). 6 million spam tweets: A large ground truth for timely Twitter spam detection. In *2015 ieee international conference on communications (icc)* (p. 7065-7070). doi: 10.1109/ICC.2015.7249453

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018, October). Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd workshop on abusive language online (ALW2)* (pp. 11–20). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-5102 doi: 10.18653/v1/W18-5102

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Gautam, K., Benevenuto, F., ... Gummadi, K. (2012, April). Understanding and Combating Link Farming in the Twitter Social Network. In *Proceedings of the 21st International World Wide Web Conference (WWW'12).* Lyon, France.

Hochreiter, S., & Schmidhuber, J. (2015). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.

Jurafsky, D., & Martin, J. (2009). *Speech and language processing* (2nd ed.). Pearson Prentice Hall.

Kantchelian, A., Ma, J., Huang, L., Afroz, S., Joseph, A. D., & Tygar, J. D. (2012, October). Robust detection of comment spam using entropy rate. In *Proceedings of the 5th acm workshop on artificial intelligence and security* (p. 59-70).

Kim, D., Graham, T., Wan, Z., & Rizoiu, M. (2019). Tracking the digital traces of Russian trolls: Distinguishing the roles and strategy of trolls on Twitter. *CoRR*, *abs/1901.05228*. Retrieved from http://arxiv.org/abs/1901.05228

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd international conference on learning representations (iclr).*

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2019, Nov 02). The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*. Retrieved from https://doi.org/10.1007/s41701-019-00065-w doi: 10.1007/s41701-019-00065-w

Linvill, D. L., & Warren, P. L. (2018). *Troll factories: The internet research agency and state-sponsored agenda building.* Retrieved from http://pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf

Liu, P., Guberman, J., Hemphill, L., & Culotta, A. (2018). Forecasting the presence and intensity of hostility on instagram using linguistic and social features. In *Twelfth international aaai conference on web and social media.*

Ljubešić, N., Erjavec, T., & Fišer, D. (2018, October). Datasets of Slovene and Croatian moderated news comments. In *Proc. 2nd workshop on abusive language online* (pp. 124–131). Retrieved from https://www.aclweb.org/anthology/W18-5116 doi: 10.18653/v1/W18-5116

Ljubešić, N., Fišer, D., & Erjavec, T. (2019). The FRENK datasets of socially unacceptable discourse in Slovene and English. *CoRR*, *abs/1906.02045*. Retrieved from http://arxiv.org/abs/1906.02045

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019, 08). Hate speech detection: Challenges and solutions. *PLOS ONE*, *14*(8), 1-16. Retrieved from https://doi.org/10.1371/journal.pone.0221152 doi: 10.1371/journal.pone.0221152

Mihaylov, T., & Nakov, P. (2016, August). Hunting for troll comments in news community forums. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 399–405). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P16-2065 doi: 10.18653/v1/P16-2065

Mojica, L. G. (2017). A trolling hierarchy in social media and A conditional random field for trolling detection. *CoRR*, *abs/1704.02385*. Retrieved from http://arxiv.org/abs/1704.02385

Mojica de la Vega, L. G., & Ng, V. (2018, May). Modeling trolling in social media conversations. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC-2018).* Miyazaki, Japan: European Languages Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L18-1585

Napoles, C., Tetreault, J., Rosata, E., Provenzale, B., & Pappu, A. (2017, April). Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th linguistic annotation workshop* (pp. 13–23). Valencia, Spain: Association for Computational Linguistics.

Narayanan, A. (2018). *Tweets dataset for detection of cyber-trolls.* Retrieved from https://www.kaggle.com/dataturks/dataset-for-detection-of-cybertrolls

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Pamungkas, E. W., & Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop* (pp. 363–370).

Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2017a, September). Deeper attention to abusive user content moderation. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1125–1135). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D17-1117 doi: 10.18653/v1/D17-1117

Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2017b, August). Deep learning for user comment moderation. In *Proceedings of the first workshop on abusive language online* (pp. 25–35). Vancouver, BC, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W17-3004 doi: 10.18653/v1/W17-3004

Ruder, S., Vulić, I., & Søgaard, A. (2017). A survey of cross-lingual word embedding models. *CoRR*, *abs/1706.04902*. Retrieved from http://arxiv.org/abs/1706.04902

Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In *Proceedings of the 5th international workshop on natural language processing for social media* (pp. 1–10). Retrieved from https://www.aclweb.org/anthology/W17-1101 doi: 10.18653/v1/W17-1101

van der Goot, R., Ljubešić, N., Matroos, I., Nissim, M., & Plank, B. (2018). Bleaching text: Abstract features for cross-lingual gender prediction. *arXiv preprint arXiv:1805.03122*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (p. 5998-6008).

Waseem, Z., & Hovy, D. (2016, 01). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the naacl student research workshop* (p. 88-93).

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018, September). 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).* Vienna, Austria.

Wu, T., Wen, S., Xiang, Y., & Zhou, W. (2018). Twitter spam detection: Survey of new approaches and comparative study. *Computers and Security*, *76*, 265-284. Retrieved from http://www.sciencedirect.com/science/article/pii/S016740481730250X doi: https://doi.org/10.1016/j.cose.2017.11.013

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web* (pp. 1391–1399).

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). *CoRR*, *abs/1903.08983*. Retrieved from http://arxiv.org/abs/1903.08983

Zhang, J., Chang, J., Danescu-Niculescu-Mizil, C., Dixon, L., Hua, Y., Taraborelli, D., & Thain, N. (2018, July). Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1350–1361). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P18-1125 doi: 10.18653/v1/P18-1125

A. Yearwise Data Distribution

This section gives the full details of the dataset distributions over time, in terms of overall numbers of articles, comments and moderator's blocking behaviour for all three newspapers (Section 4.2.1), and the frequency of application of individual blocking rules for 24sata and VL (Section 4.4).

A.1. Summary data, commenting and blocking rates

Year	# Articles	# Comments	# Blocked	Comment rate	Blocking rate
2007	6054	38005	3	6.3	7.9×10^{-5}
2008	26523	185578	12	7.0	6.5×10^{-5}
2009	38024	326609	31	8.6	9.5×10^{-5}
2010	38777	459227	2	11.8	4.4×10^{-6}
2011	38330	1140555	111	29.8	9.7×10^{-5}
2012	43978	1870449	251	42.5	1.3×10^{-4}
2013	46457	2490285	130	53.6	5.2×10^{-5}
2014	46429	2656841	171	57.2	6.4×10^{-5}
2015	44919	3054087	724	68.0	2.4×10^{-4}
2016	47595	3194761	98487	67.1	3.1×10^{-2}
2017	45891	2795824	134080	60.9	4.8×10^{-2}
2018	48777	2519279	156083	51.7	6.2×10^{-2}
2019	17953	816692	63972	45.5	7.8×10^{-2}
Total	489707	21548192	454057		

Table 16: Yearwise data distribution, 24sata; comment rate $= N_{\mathrm{comments}}/N_{\mathrm{articles}}$, blocking rate $= N_{\mathrm{blocked}}/N_{\mathrm{comments}}$.

Year	# Articles	# Comments	# Blocked	Comment rate	Blocking rate
2009	7724	162017	4	20.98	2.47×10^{-5}
2010	31423	764134	175	24.32	2.29×10^{-4}
2011	32521	1245946	91	38.31	7.30×10^{-5}
2012	35693	1022186	29	28.64	2.84×10^{-5}
2013	41408	1101234	16747	26.59	1.52×10^{-2}
2014	43251	835152	48099	19.31	5.76×10^{-2}
2015	43469	1237714	48930	28.47	3.95×10^{-2}
2016	40485	1009070	60390	24.92	5.98×10^{-2}
2017	38136	840677	87476	22.04	1.04×10^{-1}
2018	42092	1073953	130054	25.51	1.21×10^{-1}
2019	16453	354551	55295	21.55	1.56×10^{-1}
Total	372655	9646634	447290		

Table 17: Yearwise data distribution, Večernji List; comment rate $= N_{\mathrm{comments}}/N_{\mathrm{articles}}$, blocking rate $= N_{\mathrm{blocked}}/N_{\mathrm{comments}}$.

Year	# Articles	# Comments	# Blocked	Comment rate	Blocking rate
2009	109352	2898438	130040	26.51	4.49×10^{-2}
2010	105173	2377591	107735	22.61	4.53×10^{-2}
2011	127037	2729389	148302	21.49	5.43×10^{-2}
2012	127663	3372776	249880	26.42	7.41×10^{-2}
2013	114914	3289393	295608	28.63	8.99×10^{-2}
2014	101936	3195502	336450	31.35	10.53×10^{-2}
2015	98198	3202592	391758	32.61	12.23×10^{-2}
2016	94353	2848624	355868	30.19	12.49×10^{-2}
2017	87098	2838075	265810	32.58	9.37×10^{-2}
2018	82887	3194597	343538	38.54	10.75×10^{-2}
2019	32691	1540382	188197	47.12	12.21×10^{-2}
Total	1081302	31487359	2813186		

Table 18: Yearwise data distribution, Ekspress; comment rate $= N_{\mathrm{comments}}/N_{\mathrm{articles}}$, blocking rate $= N_{\mathrm{blocked}}/N_{\mathrm{comments}}$.

A.2. Blocking rule distribution

	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Rule7	Rule8
2007						1		2
2008	12							
2009	29		1					1
2010	2							
2011	107							4
2012	144				2	9	13	83
2013	112				5		1	12
2014	108	1	1		45	2		14
2015	659	2	7		18	1		37
2016	23551	111	3152	183	2479	7400	227	61384
2017	50178	185	5310	153	4631	5752	137	67734
2018	65775	254	8099	125	8483	3453	780	69114
2019	31592	26	2734	37	3658	1270	498	24157

Table 19: Yearwise blocking rule data distribution, 24sata.

	Rule1	Rule2	Rule3	Rule4	Rule5	Rule6	Rule7	Rule8	Rule9
2009		4							
2010	91		1		1				82
2011	44		1		4				42
2012	8		4	2	4				11
2013	1748	52	6575	16	7192	15	618	275	256
2014	4913	83	20911	19	16462	114	813	142	4642
2015	5438	82	16729	24	21858	109	187	4	4499
2016	4859	118	14007	10	38076	147	889	2	2282
2017	28888	169	15251	30	35957	195	608		6378
2018	33660	8076	17311	4	37572	45	256	8	33122
2019	4860	8477	5748	72	4712	11	199	4	31212

Table 20: Yearwise blocking rule data distribution, Večernji List.

B. Word Lists

This section gives the full top 100 words lists for blocked and non-blocked comments as inferred by the Naïve Bayes classifier trained on the binary classification task (Section 4.2.2).

Word	Ratio	Word	Ratio	Word	Ratio
20544	2.16	ponio	1.50	jebo	1.40
22000	2.14	ovom	1.48	odnio	1.40
17000	2.14	ovog	1.48	ženu	1.40
pridružio	2.08	želiš	1.47	sada	1.40
mreži	2.08	internetu	1.47	dobivanje	1.40
www	2.03	radno	1.46	nepunim	1.39
com	2.02	jebem	1.46	redoviti	1.39
mjesečno	1.94	promijenilo	1.45	pogledam	1.39
google	1.94	slijedite	1.45	radeci	1.39
mjesecu	1.85	dnevno	1.45	sponzoru	1.39
kuće	1.81	paycheck	1.44	šokiran	1.38
dolara	1.80	eura	1.44	redovne	1.38
mjeseca	1.79	odlučio	1.44	počeo	1.38
prvom	1.78	dnevne	1.44	stanicom	1.38
poslu	1.77	nabijem	1.43	odabirete	1.38
zaradio	1.76	litte	1.43	primio	1.38
rad	1.74	24857	1.43	vremenom	1.37
radeći	1.70	čula	1.43	zarađivati	1.37
promijenjen	1.69	web	1.42	želite	1.36
plaća	1.69	top	1.42	blogu	1.36
dobrodošli	1.69	započela	1.42	prije	1.36
7645	1.67	premise	1.42	dodatni	1.36
9264	1.67	rasponu	1.42	86	1.36
27936	1.67	prošlog	1.42	prethodni	1.36
tjedno	1.57	počinjem	1.41	zaradite	1.35
online	1.57	četiri	1.41	rate	1.35
pronaći	1.55	jednostavan	1.41	39	1.35
mom	1.54	29584	1.41	stranicu	1.35
posla	1.53	22738	1.41	posjetite	1.35
zaraditi	1.53	sam	1.41	majmune	1.35
noć	1.52	debil	1.40	mijenjam	1.34
skraćeno	1.52	računalo	1.40	govno	1.34
satu	1.51	jo	1.40	nepuno	1.34
				mjesec	1.34

Table 21: Top 100 word features for blocked comments, in order of class probability ratio

Word	Ratio	Word	Ratio	Word	Ratio
sritno	1.26	vrtić	1.18	gripa	1.16
nii	1.25	noja	1.18	kapetan	1.16
sretno	1.24	liniju	1.18	ličnost	1.16
strašno	1.24	tekma	1.17	težak	1.16
inter	1.23	ponovilo	1.17	niš	1.16
derbi	1.21	šanse	1.17	sudar	1.16
napišite	1.21	osijek	1.17	petak	1.16
naklon	1.21	strah	1.17	bok	1.16
malena	1.20	ajmoo	1.17	vrhova	1.16
var	1.20	vozac	1.17	cirkusanti	1.16
štima	1.20	miša	1.17	šubi	1.16
zavisi	1.20	nima	1.17	terorizam	1.16
humbla	1.20	glumac	1.17	probaju	1.16
điri	1.20	kiša	1.17	jela	1.16
prekrasna	1.20	miru	1.17	sjeveru	1.16
svašta	1.20	išlo	1.17	cudimo	1.16
pocelo	1.20	vakula	1.17	potpisujem	1.16
počivaj	1.19	svizac	1.17	nadje	1.16
gledanost	1.19	dvojno	1.17	cares	1.16
drž	1.19	pila	1.17	žiri	1.16
oja	1.19	zasluženo	1.17	hrabro	1.16
horor	1.19	ligama	1.17	kip	1.16
predivno	1.19	najte	1.17	blagi	1.16
obožavam	1.19	tragedija	1.17	dizel	1.16
mokra	1.19	baš	1.17	tuzno	1.16
odlično	1.18	teško	1.17	nasmijao	1.16
sumljam	1.18	skupit	1.17	informaciji	1.16
pocivao	1.18	troše	1.17	srećom	1.16
pravna	1.18	anđeli	1.17	trolaš	1.16
sućut	1.18	svastiku	1.17	prolazak	1.16
bisera	1.18	hep	1.17	lepi	1.16
ludost	1.18	najljepša	1.17	pretjerao	1.16
filmova	1.18	izvoli	1.16	čekala	1.16
				snijeg	1.16

Table 22: Top 100 word features for non-blocked comments, in order of class probability ratio