Not All Comments are Equal: Insights into Comment Moderation from a Topic-Aware Model

Elaine Zosa
University of Helsinki
elaine.zosa@helsinki.fi

Ravi Shekhar
Queen Mary University of London
r.shekhar@qmul.ac.uk

Vanja Mladen Karan
[◇]Queen Mary University of London
m.karan@qmul.ac.uk

Matthew Purver[◇,†]
[†]Jožef Stefan Institute
m.purver@qmul.ac.uk

Abstract

Moderation of reader comments is a significant problem for online news platforms. Here, we experiment with models for automatic moderation, using a dataset of comments from a popular Croatian newspaper. Our analysis shows that while comments that violate the moderation rules mostly share common linguistic and thematic features, their content varies across the different sections of the newspaper. We therefore make our models topic-aware, incorporating semantic features from a topic model into the classification decision. Our results show that topic information improves the performance of the model, increases its confidence in correct outputs, and helps us understand the model's outputs.

1 Introduction

Most newspapers publish their articles online, and allow readers to comment on those articles. This can increase user engagement and page views, and provides readers with an important route to public freedom of expression and opinion, with the ability to interact and discuss with others. Comment sections usually provide some degree of anonymity;[1] while improving accessibility, this can also encourage inappropriate behaviour, and publishers therefore usually employ some moderation policy to regulate content and to ensure legal compliance (in some cases, publishers can be held responsible for user-contributed content on their sites).

One possible approach is a 'moderate then publish' policy, in which comments must be approved by a moderator before they appear; this requires significant manpower and introduces delays and limitations into the user conversation (for example, the New York Times only allows comments for one day after article publication[2]). On the other hand, a 'publish then moderate' strategy, in which comments are published immediately, and later removed if necessary, is less effective at blocking toxic or illegal content. Combined with the increase in comment volumes in recent years there is increasing interest in automatic moderation methods (see e.g. Pavlopoulos et al., 2017a), either as standalone tools or for integration into human moderators' practices (Schabus and Skowron, 2018).

Detecting comments that need moderators' attention is usually approached as a text classification task (see e.g. Pavlopoulos et al., 2017a); but comments can be blocked for a range of reasons (Shekhar et al., 2020). One is the presence of offensive language, a well-studied NLP task (see Section 2 below); however, others include advertising or spam, illegal content, spreading misinformation, trolling and incitement — all distinct categories which might be expected to show distinct features, and perhaps to vary according to the content being commented on. Another aspect that distinguishes the comment moderation task from the usual text classification tasks in NLP is the need for interpretable or explainable models: if classifiers are to be used by human moderators within publishers' working practices, they must be able to understand the outputs (Švec et al., 2018).

Here, we therefore investigate models which can provide both an aspect of interpretability and the ability to take account of the topics being discussed, by incorporating topic information into the comment classifier. Specifically, we incorporate semantic representations learned by the Embedded Topic Model (ETM) (Dieng et al., 2020) into a classifier pipeline based on Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). Our model improves performance

[1]Some newspapers allow completely anonymous posting; some require commenters to create an account with a username, but this does not usually reveal their true identity.

[2]NYT Comment FAQ: https://nyti.ms/2PF02kj

Proceedings of Recent Advances in Natural Language Processing, pages 1652–1662
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_185

by 4.4% over a text-only approach on the same dataset (Shekhar et al., 2020), and is more confident in the correct decisions it makes. Inspection of the topic distributions reveals how different newspaper sections have different language and topic distributions, including differences in the kind of comments that need moderation.[3]

2 Related Work

Automated news comment moderation Most research on this task so far formulates it as a text classification problem: for a given comment, the model must predict whether the comment violates the newspaper's policy. However, approaches to classification vary. Nobata et al. (2016) use a range of linguistic features, e.g. lexicon and n-grams. Pavlopoulos et al. (2017a) and Švec et al. (2018) use neural networks, specifically RNNs with an attention mechanism. Recently, Tan et al. (2020) and Tran et al. (2020) apply a modified BERT model (Devlin et al., 2019) while Schabus et al. (2017) use a bag-of-words approach.

Some approaches go beyond the comment text itself: Gao and Huang (2017) add information like user ID and article headline into their RNN to make the model context-aware; Pavlopoulos et al. (2017b) incorporate user embeddings; Schabus and Skowron (2018) incorporate the news category metadata of the article. However, no work so far investigates automatic modelling of topics (rather than relying on categorical metadata), or applies this to the comments rather than just their parent articles.

Some steps towards model intepretability and output explanation have also been taken: both Švec et al. (2018) and Pavlopoulos et al. (2017a) use an attention saliency map to highlight possibly problematic words. However, we are not aware of any work using higher-level topic information as a route to understanding model outputs.

Available datasets Several datasets have been created for the news comment moderation task. Nobata et al. (2016) provide 1.43M comments posted on Yahoo! Finance and News over 1.5 years, in which 7% of the comments are labelled as abusive via a community moderation process. Gao and Huang (2017) contains 1.5k comments from Fox News, annotated with specific hateful/non-hateful labels as a post-hoc task, and having 28% hateful comments. However, both are relatively small, and their labelling methods mean that neither dataset is entirely representative of the moderation process performed by newspapers.

Pavlopoulos et al. (2017a) provides 1.6M comments from Gazzetta, a Greek sports news portal, over c.1.5 years. Here, 34% of comments are labelled as blocked, and the labels are derived from the newspaper's human moderators and journalists. Schabus et al. (2017) and Schabus and Skowron (2018) provide a dataset from a German-language Austrian newspaper with 1M comments posted over 1 year, out of which 11,773 comments are annotated using seven different rules.

More recently, Shekhar et al. (2020) present a dataset from 24sata, Croatia's most widely read newspaper.[4] This dataset is significantly larger (10 years, c.20M comments); and moderator labels include not only a label for blocked comments, but also a record of the reason for the decision according to a 9-class moderation policy. However, their experiments show that classifier performance is limited, and transfers poorly across years. Here, we therefore use this dataset (see Section 3), with a view to improving performance and applying a topic-aware model to improve and better understand the robustness in the face of changing topics.

Related tasks More attention has been given to related tasks, most prominently the detection of offensive language, hate speech, and toxicity (Pelicon et al., 2021). A comprehensive survey of dataset collection is provided by Poletto et al. (2020) and Vidgen and Derczynski (2020).[5]

Topic Modelling Topic models capture the latent themes (also known as *topics*) from a collection of documents through the co-occurence statistics of the words used in a document. Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a popular method for capturing these topics, is a generative document model where a document is a mixture of topics expressed as a probability distribution over the topics and a topic is a distribution over the words in a vocabulary. The Embedded Topic Model (ETM, Dieng et al., 2020) is an LDA-like topic modelling method that exploits the semantic information captured in word embeddings during topic inference. The advantage of ETM over LDA

[3] Source code available at https://github.com/ezosa/topic-aware-moderation

[4] http://24sata.hr/

[5] http://hatespeechdata.com/ provides a comprehensive list of relevant datasets.

Comment Moderation Data			
	Blocked	Non-blocked	Blocking Rate
Train	4984	75016	6.23%
Valid	642	9358	6.42%
Test	37271	438142	7.84%
Topic Modelling Data			
	Blocked	Non-blocked	Blocking Rate
Train	34863	36725	48.70%
Valid	4880	5120	48.80%

Table 1: Details of datasets used in experiments.

Section (– Subsection)	Blocked	Non-blocked	Blocking Rate
Kolumne (*Columns*)	655	6382	9.31%
Lifestyle	2426	30985	7.26%
Show	6827	58896	10.39%
Sport	5882	80820	6.78%
Tech	382	7173	5.06%
Vijesti (*News*)	20094	239835	7.73%
– Crna kronika (*Crime*)	5917	62471	8.65%
– Hrvatska (*Croatia*)	3527	45170	7.70%
– Politika (*Politics*)	6088	80264	7.05%
– Svijet (*World*)	2625	31459	7.24%

Table 2: Details per section, and (for section Vijesti) sub-section, of the comment moderation test set.

is that it combines the advantages of word embeddings with the document-level dependencies captured by topic modelling and has been shown to produce more coherent topics than regular LDA.

3 Dataset

We use the 24sata comment dataset (Shekhar et al., 2020; Pollak et al., 2021), introduced in Section 2. This contains c.21M comments on 476K articles from the years 2007-2019[6], written in Croatian. The dataset has details of comments blocked by the 24sata moderators, based on a set of moderation rules–these vary from hate speech to abuse to spam (see Shekhar et al., 2020, for rule description). The dataset also identifies the article under which a comment was posted, together with the section/subsection of the newspaper the article appeared in. These sections/sub-sections relate to the content of the article: for example, the Sport section contains sports-related news while the Kolumne (*Columns*) section contains opinion pieces. The largest section, Vijesti (*News*), is further subdivided as shown in Table 2.

3.1 Data Selection

In this work, we use data from 2018 for training and validation of the topic model and classifiers and data from 2019 for testing. This reflects the realistic scenario where we use data collected from the past to make predictions. For training and validation, we randomly select 50,000 articles out of 65,989 articles from 2018, sampling from the nine most-representative sections/sub-sections (Table 2). Each article comes with c.50 comments on average.

To train the topic model, we sample around 80,000 comments across these articles, with a roughly equal split between blocked and non-blocked comments. This is to encourage a diverse mix of topics from both comment classes. As a preprocessing step we remove comments with less than 10 words from the training data (see Table 1 (lower part)). To train the classifiers, we randomly sample around 80,000 comments such that the sampled set has the same blocking rate as the entire 2018 dataset.

For the test set, we then use all 475,413 comments associated with the 17,953 articles from 2019. Table 1 (upper part) provides the dataset details, with comment moderation blocking rate. For the test set, Table 2 provides details on the section and sub-section of the related articles. These top nine sections account for more than 95% of the comments of the entire test set.

3.2 Content Analysis

To gain some insight into the content of blocked comments, we analyze the linguistic differences between blocked and non-blocked comments and across different sections. First, we compare comment length. As we can see from Table 3, blocked and non-blocked comments have, on average, similar lengths. However, if we further divide blocked comments into two sub-groups — spam and non-spam — we find that on average, spam comments are longer than other comments. We observe a similar pattern across different sections.

Next, we measure lexical diversity using mean-segmental type-token ratio (MSTTR). The MSTTR is computed as the mean of type-token ratio for every 1000 tokens in a dataset to control for dataset size (van Miltenburg et al., 2018). From Table 3, we see that non-blocked comments have higher MSTTR (i.e. higher lexical diversity) than blocked comments (0.62 vs 0.46). However, when we again divide blocked comments into spam and non-spam,

[6]Dataset is available at http://hdl.handle.net/11356/1399

we observe that non-spam blocked comments have a similar MSTTR to non-blocked comments (0.61 vs 0.62), while spam comments have much lower MSTTR (0.35 vs 0.61). This suggests that blocked comments (excluding spam) have as rich a vocabulary as non-blocked. Again, we see a similar pattern across different news sections.

	Mean length	MSTTR
All	23.06	0.61
Non-blocked	23.01	0.62
Blocked	23.65	0.46
Blocked (non-spam)	19.16	0.61
Blocked (Spam only)	28.23	0.35

Table 3: Mean-segmental TTR and average length of comments

Now we look at the top bigrams of each class. We collect all bigrams that occur at least 50 times and rank them according to their pointwise mutual information (PMI) score. In general, we do not see many overlaps between the top bigrams of blocked and non-blocked comments across the different sections. Bigrams in blocked comments indicate spam messages such 'iskustva potrebnog' (*experience required*), 'redoviti student' (*full-time student*) and 'prilika pružila' (*opportunity given*). Removing spam comments, we encounter bigrams used for swearing such as 'pas mater' (*damn it*) and 'jedi govna' (*eat sh*t*). In the non-blocked comments, the top bigrams are more relevant to the section they appear in. For instance, in the Vijesti section, top bigrams include 'new york', 'porezni obveznici' (*taxpayers*) and 'naftna polja' (*oil fields*) while in Sports, top bigrams include 'all star', 'grand slam' and 'man utd'.

This suggests that the content of blocked comments tends to share commonalities across sections more than non-blocked comments; but again, these commonalities may be mostly within the spam category, with other blocked categories being more topic-dependent. Our next step therefore is to examine the use of topic modelling to capture these dependencies, with a view to using topic information to improve a moderation classifier.

4 Topic Modelling

We now apply a topic model to gain insight into what characterises a blocked comment and a non-blocked one, and whether this varies between different sections where different subjects are discussed.

4.1 Topic Model

We use the Embedded Topic Model (ETM, Dieng et al., 2020) as our topic model since it has been shown to outperform regular LDA and and other neural topic modelling methods such as NVDM (Miao et al., 2016). We also want to take advantage of ETM's ability to incorporate the information encoded in pretrained word embeddings trained on vast amounts of data to produce more coherent topics. In the ETM, the topic-term distribution for topic k, β_k, is induced by a matrix of word embeddings ρ and its respective topic embedding α_k which is a point in the word embedding space:

$$\beta_k = softmax(\rho^T \alpha_k) \quad (1)$$

The topic embeddings are learned during topic inference while the word embeddings can be pretrained or also learned during topic inference. In this work, we use pretrained embeddings.

The document-topic distribution of a document d, θ_d, is drawn from the logistic normal distribution whose mean and variance come from an inference network:

$$\theta_d \sim LN(\mu_d, \sigma_d) \quad (2)$$

Given a trained ETM, we can infer the **document-topic distribution (DTD)** of an unseen document. In addition, we can also compute a **document-topic embedding (DTE)** as the weighted sum of the embeddings of the topics in a document, where the weight corresponds to the probability of the topic in that document:

$$DTE = \sum_{k=0}^{K} \alpha_k \theta_{d,k} \quad (3)$$

where α_k is the topic embedding of topic k, and $\theta_{d,k}$ is the probability of topic k in doc d.

4.2 Topic Analysis

Now we analyse the usage of topics in our test set. We trained the ETM for 100 topics on the training set and inferred the topic distributions of the comments in the test set. For analysis, we extract the top topics in a set of comments. To do this, we take the mean of the topic distributions over the comments in the set and rank the topics according to their weight in this mean distribution. We then take the top 15 topics for analysis because this is the average number of topics in a comment with a non-zero probability in our test set. Note that in this analysis we only use the document-topic

distributions and not the document-topic embeddings. To more easily discuss the topics here we provide concise labels for each topic as interpreted by a native speaker. Automatic labelling of topics is a non-trivial task and an area of active research (Bhatia et al., 2016; Alokaili et al., 2020; Popa and Rebedea, 2021).

First, we examine the prevalent topics in the blocked and non-blocked comments, separately. The top topics of non-blocked comments cover a diverse range of subjects from politics to football while the top topics in blocked comments are dominated by spam and offensive language (Figure 1). However, we also see many topics shared between blocked and non-blocked comments. [7].

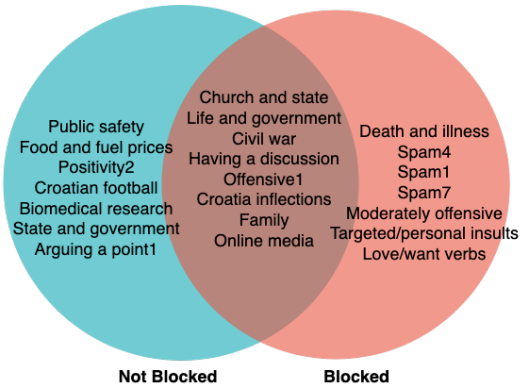

Figure 1: Top topics of the blocked and non-blocked comments for the entire test set.

Next we illustrate how different topics intersect and diverge between blocked and non-blocked comments across sections by looking at the top topics of two thematically-different sections, Lifestyle and Politika (*Politics*).

Figure 2 shows the top topics of these sections and the intersections between them. In Politics, blocked comments tend toward spam and targeted insults. Non-blocked topics include public safety and finances. However, we also see that more than half of the top topics overlap between blocked and non-blocked. This suggests that, thematically, there isn't a very clear distinction between blocked and non-blocked comments in the Politics section.

In Lifestyle, blocked topics are dominated by spam and while there are topics on offensive insults, they are not as prevalent as the spam-related ones. The non-blocked topics are about family and relationships and commenters arguing with each other. Compared to Politics, we see a clearer distinction between topics in blocked and non-blocked in this section. In terms of topic overlaps between Lifestyle and Politics, blocked comments in both sections are dedicated to spam and insults while non-blocked comments focus on positive sentiments.

The combination of certain topics also provide an indication of the classification of the comment. For instance, we notice the use of topics about football cards in comments that do not do not discuss the sport (for instance, football cards as a topic is prominent in the blocked Lifestyle comments). It turns out that some commenters use the red and yellow cards from football as metaphors for being banned or having their comments blocked by moderators (12% of comments that use these metaphors are blocked by moderators). On the other hand, comments that use the football cards topics *and* any of the sports-related topics are likely to be a genuine discussion of football (only 5% of such comments are blocked by moderators). We show some examples of these comments in Table 5.

So clearly there is a distinction between the usage of topics in the non-blocked and blocked comments. We therefore think it is a good idea to propose a model which incorporates topic information into a comment moderation classifier.

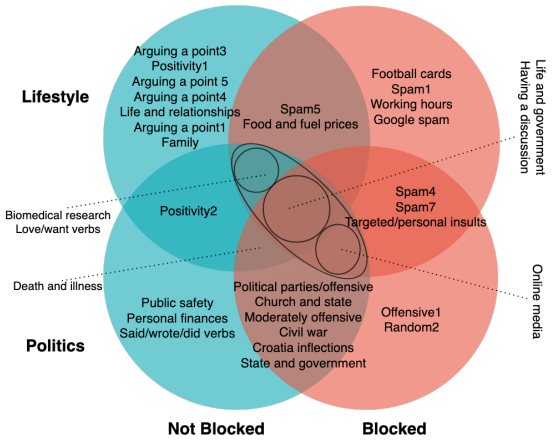

Figure 2: Top topics of the blocked and non-blocked comments in the Lifestyle and Politics sections.

5 Topic-aware Classifier

Our aim is to improve comment moderation predictions by combining textual features with document-level semantic information in the form of topics. To this end, we test several model architectures that combine a language model with topic features.

For the comment text representation, we use a

[7] All 100 topics and labels are available at https://github.com/ezosa/topic-aware-moderation

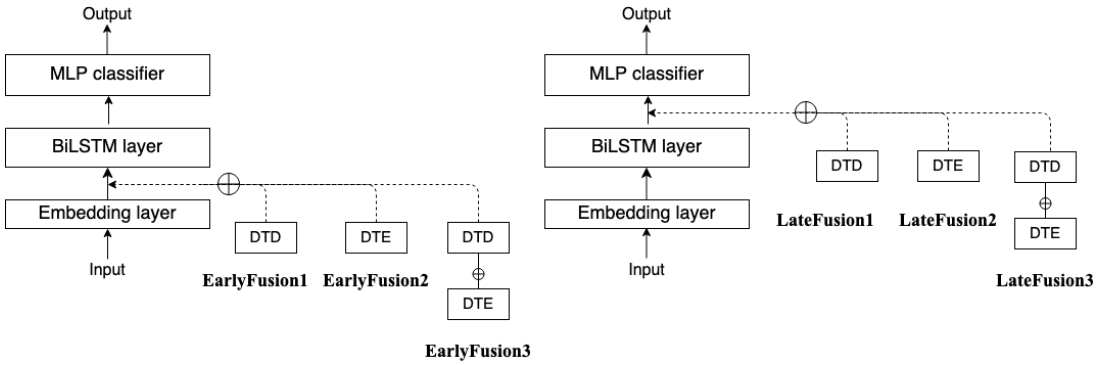

Figure 3: Architectures combining text and topic features. DTD is the topic distribution of a document while DTE is the topic embedding.

bidirectional LSTM (BiLSTM, Schuster and Paliwal, 1997). The comment text is given as input to an embedding layer then a BiLSTM layer where the output of the final hidden state is taken as the encoded representation of the comment. For the topic representations, we use the topic distributions (DTD) and topic embeddings (DTE) discussed in Section 4.1.

We propose two fusion mechanisms to combine the text and topic representations: *early* and *late* fusion. In early fusion, topic features are concatenated with the output of the embedding layer and then passed to the BiLSTM layer. In **EarlyFusion1 (EF1)**, only DTD is concatenated with the word embeddings; **EarlyFusion2 (EF2)** uses DTE instead of DTD; and **EarlyFusion3 (EF3)** uses both DTE and DTD. In late fusion, topic features are concatenated with the output representation of the BiLSTM layer, and passed to the MLP for classification. Again, **LateFusion1 (LF1)** uses DTD; **LateFusion2 (LF2)** uses DTE; and **LateFusion3 (LF3)** uses both. Figure 3 shows the architectures.

Our model is inspired by the Topic Compositional Neural Language Model (TCNLM, Wang et al., 2018) and the Neural Composite Language Model (NCLM, Chaudhary et al., 2020) that incorporate latent document-topic distributions with language models. Both of these models simultaneously learn a topic model and a language model through a joint training approach. The NCLM introduced the use of word embeddings to generate an explanatory topic representation for a document in addition to the document-topic proportions. In our work, instead of using the word embeddings of the top words of the latent topics of a document (where the number of top words is a hyperparameter), we leverage the topic embeddings learned by ETM and combine them with the document-topic proportions to produce the document-topic embeddings (DTE). Also unlike the TCNLM and NCLM, we use pre-trained topics in our model so as to easily de-couple and analyse the influence of topics in the classifier performance. Another related work is TopicRNN (Dieng et al., 2016), a model that uses topic proportions to re-score the words generated by the language model. The topics generated by this model, however, have been shown to have lower coherences compared to NCLM (Chaudhary et al., 2020).

6 Experimental Setup

Dataset As discussed in Section 3.1, we use the 2018 data as the training and validation sets of our topic-aware classifier and the 2019 data as the test set. Details of the train and validation sets are shown in Table 1 and the test set in Table 2.

Baseline models To assess how topic information improves comment classification, we use as baselines the following models trained only on text *or* topics:

- **Text only**: a classifier with BiLSTM & MLP layers, similar to Figure 3 but with comment text alone as input.
- **Document-topic distribution (DTD)**: MLP only, document-topic distributions as input.
- **Document-topic embedding (DTE)**: MLP only, document-topic embeddings as input.
- **DTD+E**: MLP only, concatenated document-topic distributions and embeddings.

Hyperparameters We use 300D word2vec embeddings, pretrained on the Croatian Web Corpus (HrWAC, Ljubešić and Erjavec, 2011; Šnajder, 2014), for training the ETM and to initialize the embedding layer of the BiLSTM. The ETM is trained

for 500 epochs for 100 topics using the default hyperparameters from the original implementation [8]. The BiLSTM is composed of one hidden layer of size 128 with dropout set to 0.5. The MLP classifier is composed of one fully-connected layer, one hidden layer of size 64, a ReLU activation, and a sigmoid for classification with the classification threshold set to 0.5. We use Adam optimizer with $lr = 0.005$. We train all classifiers for 20 epochs with early stopping based on the validation loss.

7 Results

In Table 4, we present the performance of the baselines and proposed models, measured as macro F1-scores. All models that combine text and topic representations perform better than the models that use only text *or* topics. Of the baseline models, the DTD model performs comparatively better than the DTE and DTD+E models, and surprisingly performs almost as well as the Text-only model; however, we show in Section 8 below that DTD is much less confident in its predictions than the Text-only model. Overall, the best performing model is LF1, which improves the Text-only model's performance by $+4.4\%$ (67.37% vs 62.97%); and improves by a similar amount over Shekhar et al.'s results using mBERT (macro-F1 score 62.07 for year 2019).

Interestingly, we see a wide variation in performance across news sections. We observe that comments in Lifestyle and Tech are the easiest to classify (best F1 over 72.00) while Politika (*Politics*) is the most difficult (best F1 around 61.61). The main cause appears to be that Lifestyle and Tech have the highest proportion of spam comments: on average, 49.44% of blocked comments in the test set are spam, but for Lifestyle and Tech this number rises to 77.25% and 69.63%, respectively. As for the Politics section, the most likely reason the comments are difficult to classify is that, excluding spam, there is a high degree of overlap in the subjects discussed in the blocked and non-blocked comments (see the topic analysis in Section 4.2).

7.1 Analysis of Classifier Outputs

In general, we observe that blocked comments tend to use similar topics across different sections while non-blocked comments have more diverse topics. Of the nine sections that we analyzed, there are five topics that are prominent in blocked comments in all sections ('Targeted/personal insults', 'Spam4', 'Spam7', 'Online media', and, 'Having a discussion') and only three topics prominent in non-blocked comments ('Having a discussion', 'Online media', and, 'Life and government'). This suggests that blocked comments are more semantically-coherent across sections than non-blocked ones. In contrast, topics in non-blocked comments tend to be more relevant to their respective sections: for instance, family and relationships are not discussed a lot in the Politics section, while Lifestyle commenters do not tend to talk about political issues.

The higher topical coherence then of blocked comments explains why a text classification approach can achieve reasonable performance; but the variation in blocked comment content between some sections explains why adding topic information improves our classification results.

Next, we analyze the confidence of classifiers and examine some of the outputs of the models. To analyze confidence, we gradually increase the classification threshold from 0.5 to 1.0 in increments of 0.05. For every new threshold, we plot the macro-F1 for the different models (Figure 4). We compare the confidence of four models: DTD, Text-only, EF2 (the strongest early fusion model), and LF1 (the overall best-performing model). We find that the most confident model is LF1 and the least confident is DTD. The two fusion classifiers display similar levels of confidence. The Text-only classifier is not as confident as the fusion classifiers but still more confident than DTD. This suggests that adding topic features to text not only improves performance, it also increases classifier confidence.

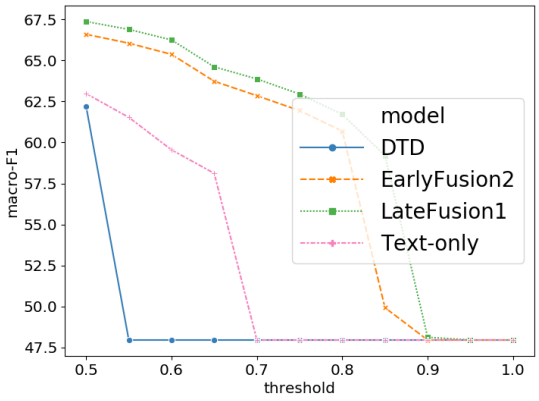

Figure 4: Confidence of the top performing models.

In Table 5 we give some examples of comments and the classifier decisions of the Text-only classifier and LF1 (our best-performing fusion model) and their top topics (topics with $prob > 0.10$). The

[8] https://github.com/adjidieng/ETM

Section – Subsection	Text only	Topics only			Text+Topic Combinations					
		DTD	DTE	DTD+E	EF1	EF2	EF3	LF1	LF2	LF3
All	62.97	62.20	59.3	58.33	66.33	66.58	65.61	**67.37**	66.22	66.95
Kolumne	59.86	59.65	56.25	55.33	62.40	62.90	63.13	63.25	62.38	**63.6**
Lifestyle	69.21	70.07	65.93	64.47	72.73	70.9	69.36	72.00	72.39	**72.92**
Show	61.97	61.30	58.62	57.60	65.24	65.63	64.26	**66.50**	65.00	65.86
Sport	63.22	61.42	58.61	57.90	67.11	67.86	66.74	**68.26**	67.14	67.82
Tech	64.87	66.37	63.17	62.55	67.72	68.74	67.65	68.76	67.68	**69.15**
Vijesti (*News*)	62.38	61.49	58.79	57.77	65.58	65.99	65.24	**66.77**	65.53	66.24
– Crna kronika	64.67	63.98	61.03	59.84	68.10	68.88	68.11	**69.60**	67.89	68.88
– Hrvatska	63.61	63.50	60.10	58.93	67.24	66.86	65.95	67.90	67.12	**67.95**
– Politika	57.93	56.49	54.95	54.20	60.51	61.52	60.84	**61.61**	60.63	61.30
– Svijet	63.58	62.55	59.62	58.35	66.83	66.95	66.33	68.44	67.21	67.57

Table 4: Classifier performance measured as macro-F1.

Comment	Label	Text-only	LF1	Top topics
1. konačno. gamad lopovska crno bijela prevarantska (*finally. the black and white cheating thieving bastards*)	1	1 (0.501)	1 (0.687)	Arguing a point, Political parties (offensive)
2. ...dobro jutro,moze crveni karton za novinara koji je osmislio naslov ;-) (... *good morning, how about a red card for the journalist who came up with this title ;-)*)	1	0 (0.315)	0 (0.456)	Football cards
3. Ne bum komentiral, dosta mi je kazni od žutih i crvenih kartona. Strah me je cenzure i bradate cure. (*No comment, I'm tired of getting yellow and red cards. I'm afraid of censorship and bearded ladies.*)	0	0 (0.054)	0 (0.335)	Football cards, Random
4. Koji kurac Rumunjski sudac ne da koji karton više Čehima. Pa svake tri minute sa leđa sruše Olma !!!! *(Why the fuck does the Romanian referee not give a few cards more to the Czechs, They tackle Olm from behind every three minutes.)*	0	0 (0.303)	1 (0.587)	Targeted/personal insults
5. Baš ste jadnici kao i ovi sa 24sata koji u ovome uživaju ! (*All of you are lame as well as those from 24sata who enjoy this.*)	1	0 (0.171)	0 (0.229)	Online media, Moderately offensive
6. Google sada plaća između 15.000 i 30.000 dolara mjesečno za rad na mreži od kuće. Pridružio sam se ovom poslu prije 3 mjeseca i zaradio 24857 dolara u prvom mjesecu ovog posla. >>> URL (*Google now pays between 15.000 and 30.000 dollars per month for working remotely from home. I started this job 3 months ago and made 24857 dollars in the first month of this job. >>> URL*)	0	1 (0.67)	1 (0.90)	Spam4

Table 5: Sample comments and classifier decisions.

first example contains swearing which both models pick up on and classify as blocked although LF1 is more confident in its decision then Text-only. In the second example, both models predict the wrong label but LF1 treats this as a borderline case because it is targeted at the moderators. However since this is only a mild provocation of the moderators, this might be a case where the gold label is incorrect. The topics also pick up on the fact that this comment talks about football cards but only has a tenuous connection to the sport ("getting a red card" is an expression used for "being banned"). In contrast, the third comment also uses the banning sense of "card" but is not directed at anyone, and is thus labeled as 0 (non-blocked), which both models get right. Again the topics indicate that the comment is not really about the sport. The fourth example shows a case where "cards" are mentioned in their standard football sense but also contains a swear word, making the gold label of 0 (non-blocked) questionable. The better performance of LF1 on such examples, compared to Text-only, implies that

LF1 is better aware of the different semantics of "card" (sports-related vs. metaphorical), likely due to added topic information.

The fifth example contains a moderately offensive insult that is not directed at any single group except the 24sata readership in general. One reason why both classifiers do not get this right is that the word *jadnici* is not strong enough to be considered offensive. Finally the last example is clearly a spam comment that both classifiers correctly classify but for which the gold label is incorrect.

Overall, compared to the Text-only model, we find that LF1 more often than not improves the confidences (and sometimes the classification), especially in cases in which the gold label is clear. This is valuable in practice, as better confidences might lead to better prioritisation of comments for manual moderation, reducing the time required to remove the most problematic ones.

8 Conclusion

In this work, we propose a model to combine document-level semantics in the form of topics with text for comment moderation. Our analysis shows that blocked and non-blocked comments have different linguistic and thematic features, and that topics and language use vary considerably across news sections, including some variation in the comments that should be blocked. We also found that blocked comments tend to be more semantically coherent across sections than non-blocked ones. We therefore see that the use of topics in our model improves performance, and gives more confident outputs, over a model that only uses the comment text. The model also provides topic distributions, interpretable as keywords, as a form of an explanation of its prediction. As future work, we plan to incorporate comment, article, and user metadata into the model.

Acknowledgements

This work has been supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA), and by the UK EPSRC under grant EP/S033564/1.

Ethics and Impact Statement

Data The dataset and annotations are provided by the publisher of 24sata.hr, Styria Media Group, for research purposes and deposited in the CLARIN repository. The authors of the comments are anonymised. The researchers used the data as-is and did not modify or add annotations.

Intended Use The models we present here are intended to assist comment moderators in their work. We do not recommend that the model be deployed in the moderation process without a human-in-the-loop.

Potential Misuse The models and the analysis of their performance we provide in this paper could be used by malicious actors to gain an insight into the comment moderation process and find loopholes in the process. However, we think such a risk is unlikely and the impact it might have outweighs the potential benefits of models intended to assist human moderators such as the ones we present here.

References

Areej Alokaili, Nikolaos Aletras, and Mark Stevenson. 2020. Automatic generation of topic labels. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1965–1968.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2016. Automatic labelling of topics with neural embeddings. *arXiv preprint arXiv:1612.05340*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Yatin Chaudhary, Hinrich Schütze, and Pankaj Gupta. 2020. Explainable and discourse topic-aware neural language understanding. In *International Conference on Machine Learning*, pages 1479–1488. PMLR.

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

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Adji B Dieng, Chong Wang, Jianfeng Gao, and John Paisley. 2016. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702*.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *International Conference on Text, Speech and Dialogue*, pages 395–402. Springer.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR.

Emiel van Miltenburg, Ruud Koolen, and Emiel Krahmer. 2018. Varying image description tasks: spoken versus written descriptions. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 88–100.

Chikashi Nobata, J. Tetreault, A. Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017a. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni, and Ion Androutsopoulos. 2017b. Improved abusive comment moderation with user embeddings. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 51–55, Copenhagen, Denmark. Association for Computational Linguistics.

Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

Senja Pollak, Marko Robnik-Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjić, Salla Salmela, Ivar Krustok, Tarmo Paju, Carl-Gustav Linden, Leo Leppänen, Elaine Zosa, Matej Ulčar, Linda Freienthal, Silver Traat, Luis Adrián Cabrera-Diego, Matej Martinc, Nada Lavrač, Blaž Škrlj, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podpečan, Janez Kranjc, Shane Sheehan, Emanuela Boros, Jose G. Moreno, Antoine Doucet, and Hannu Toivonen. 2021. EMBEDDIA tools, datasets and challenges: Resources and hackathon contributions. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 99–109, Online. Association for Computational Linguistics.

Cristian Popa and Traian Rebedea. 2021. Bart-tl: Weakly-supervised topic label generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1418–1425.

Dietmar Schabus and Marcin Skowron. 2018. Academic-industrial perspective on the development and deployment of a moderation system for a newspaper website. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1241–1244.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Ravi Shekhar, Marko Pranjić, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. Automating News Comment Moderation with Limited Resources: Benchmarking in Croatian and Estonian. *Journal for Language Technology and Computational Linguistics (JLCL)*, 34(1).

Jan Šnajder. 2014. DerivBase.hr: A high-coverage derivational morphology resource for Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3371–3377, Reykjavik, Iceland. European Language Resources Association (ELRA).

Andrej Švec, Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2018. Improving moderation of online discussions via interpretable neural models. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 60–65, Brussels, Belgium. Association for Computational Linguistics.

Fei Tan, Yifan Hu, Changwei Hu, Keqian Li, and Kevin Yen. 2020. TNT: Text normalization based pre-training of transformers for content moderation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4735–4741, Online. Association for Computational Linguistics.

Thanh Tran, Yifan Hu, Changwei Hu, Kevin Yen, Fei Tan, Kyumin Lee, and Se Rim Park. 2020. HABERTOR: An efficient and effective deep hatespeech detector. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7486–7502, Online. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Wenlin Wang, Zhe Gan, Wenqi Wang, Dinghan Shen, Jiaji Huang, Wei Ping, Sanjeev Satheesh, and Lawrence Carin. 2018. Topic compositional neural language model. In *International Conference on Artificial Intelligence and Statistics*, pages 356–365. PMLR.