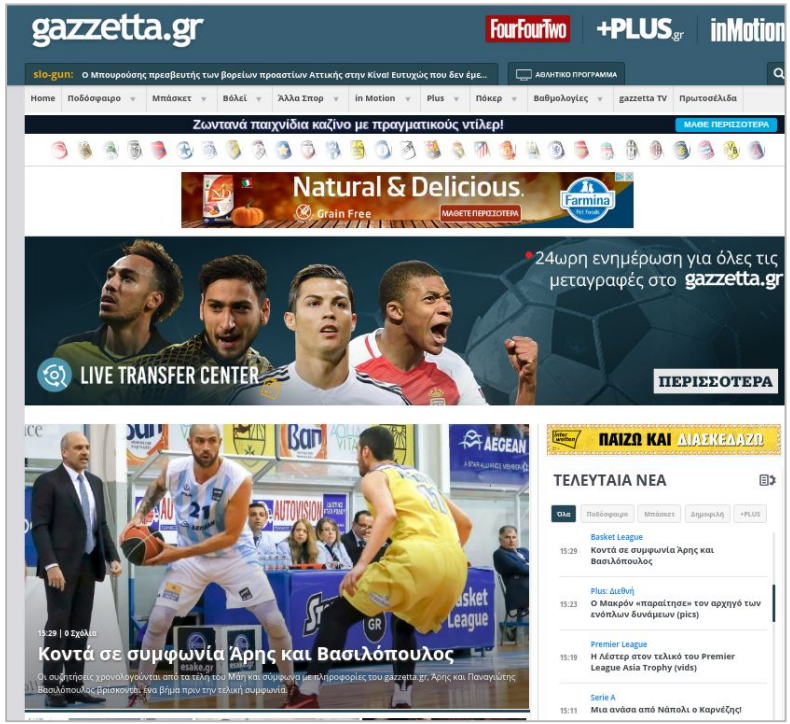
Improved Abusive Comment Moderation with User Embeddings

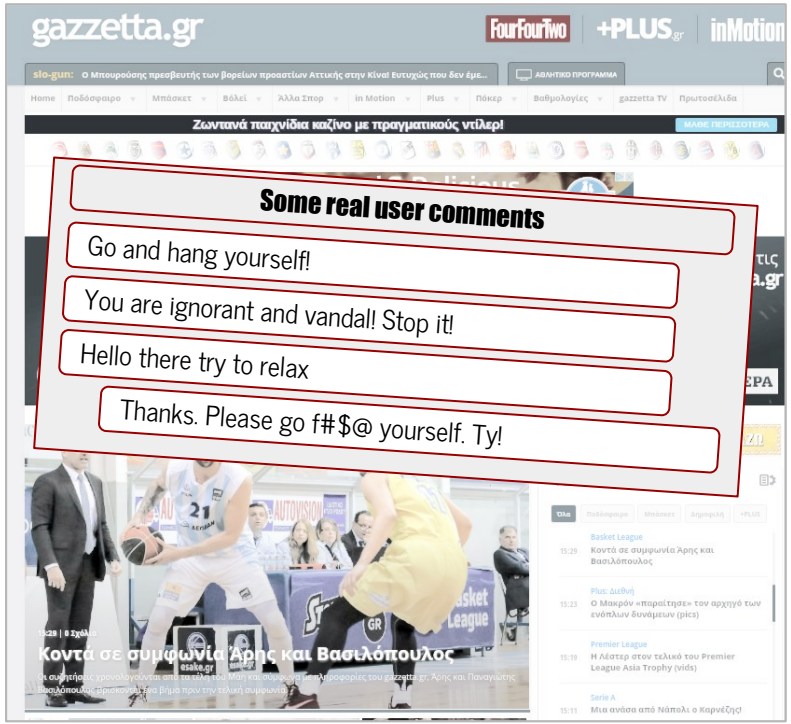
John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni, Ion Androutsopoulos

Straintek Straintek Straintek Athens University of Economics & Business

User comment moderation

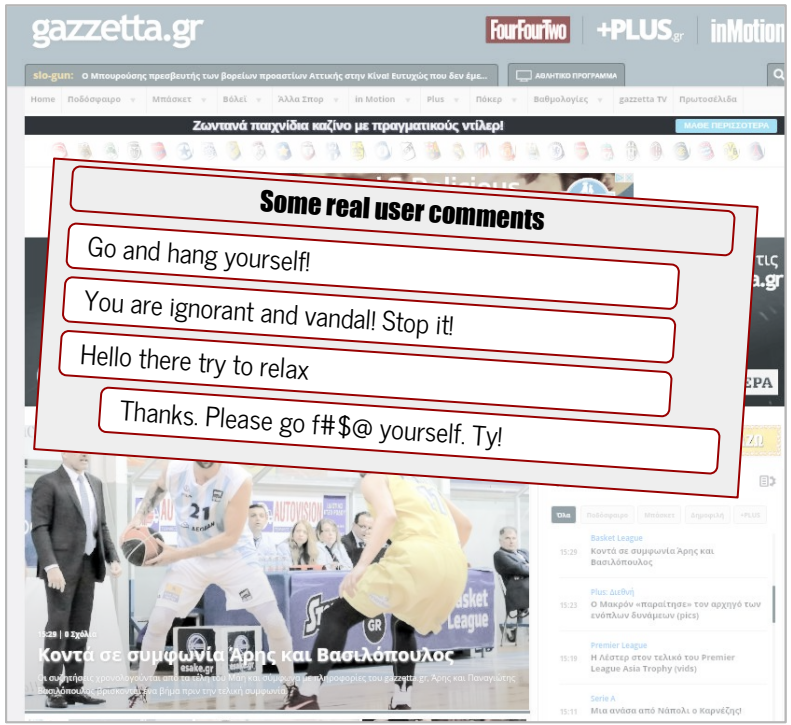


User comment moderation

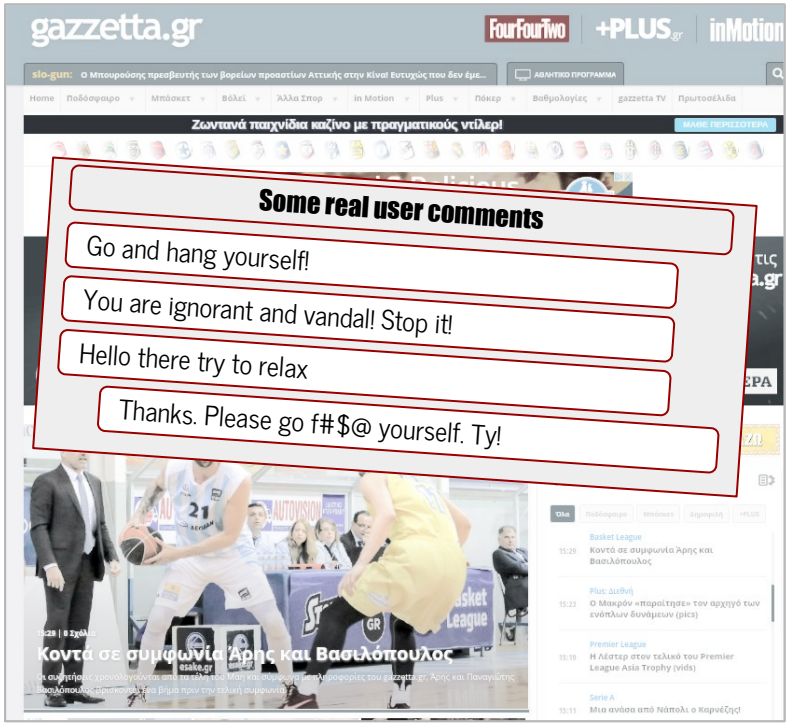


User comment moderation

Moderators to reject **abusive comments,** avoid **reputational damage, fines, putting off readers…**

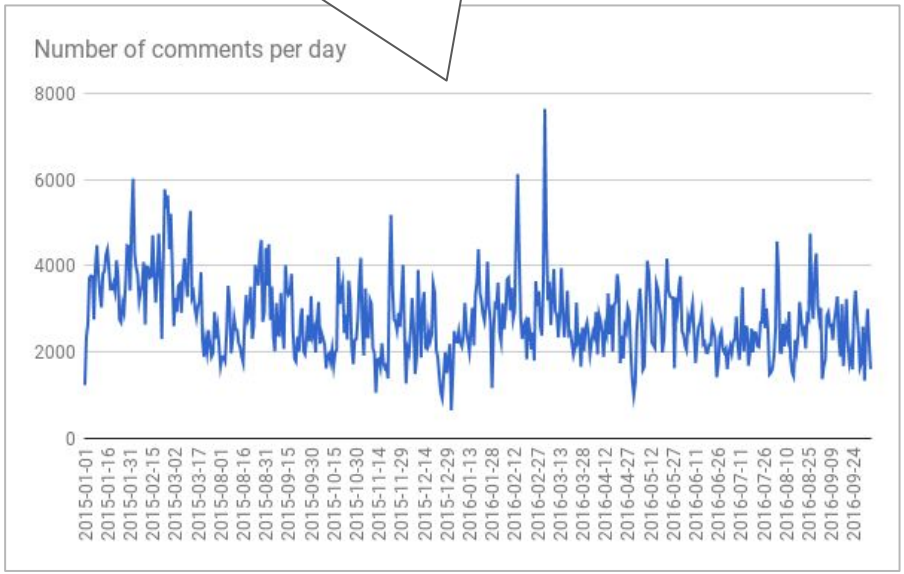


User comment moderation



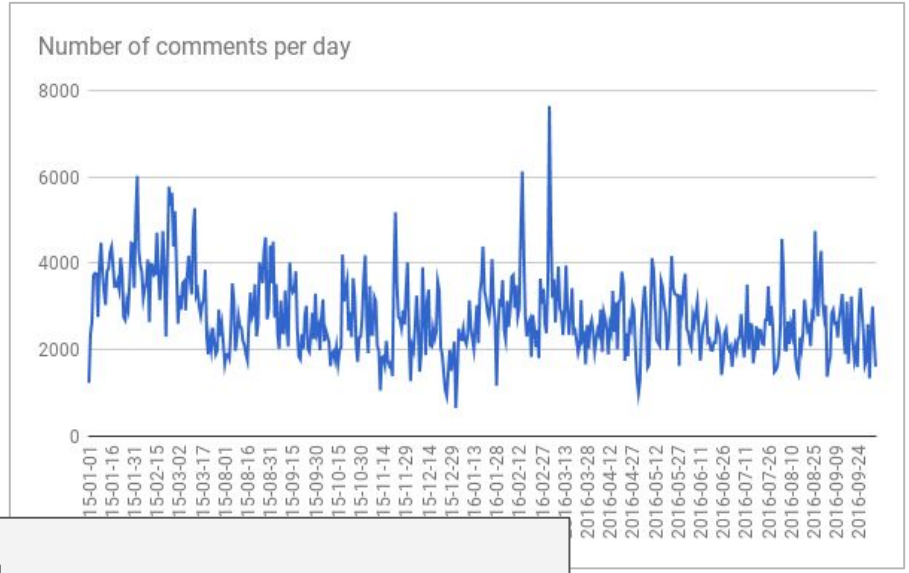
Moderators to reject **abusive comments,** avoid **reputational damage, fines, putting off readers...**

Thousands of comments per day require **moderation**. Some very **disturbing**...



User comment moderation

A better **moderation panel assists the moderators** to **detect abusive comments**, and leads to **quicker publication** of non-abusive comments.



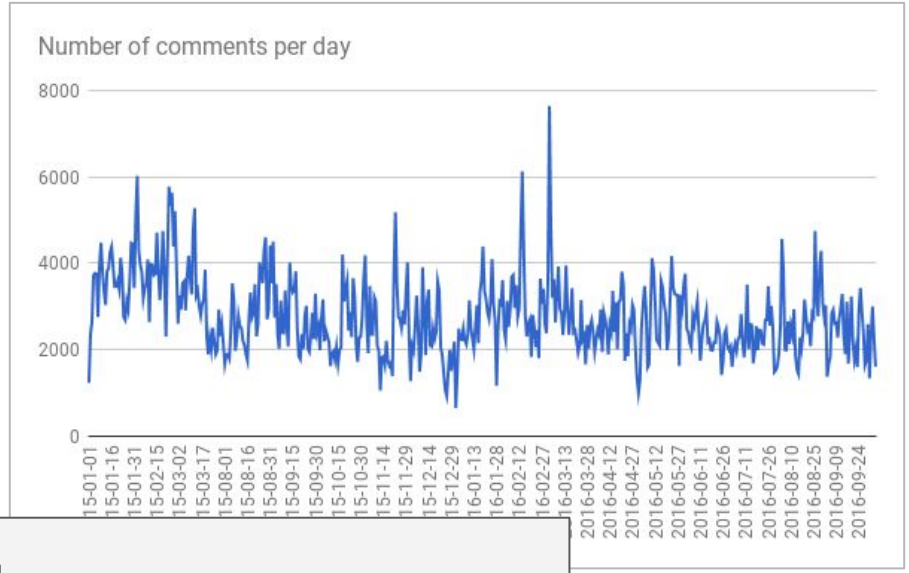
Moderation Panel

Go	and	hang	yourself	!				85%	
You	are	ignorant	and	vandal	!	Stop	it	!	88%
Hello	there	try	to	relax				0%	
Thanks	.	Please	go	f#$@	yourself	.	Ty	!	85%

User comment moderation

A better **moderation panel assists the moderators** to **detect abusive comments**, and leads to **quicker publication** of non-abusive comments.

Heatmaps show **suspicious words**: see our paper at **EMNLP-2017 (main)**.

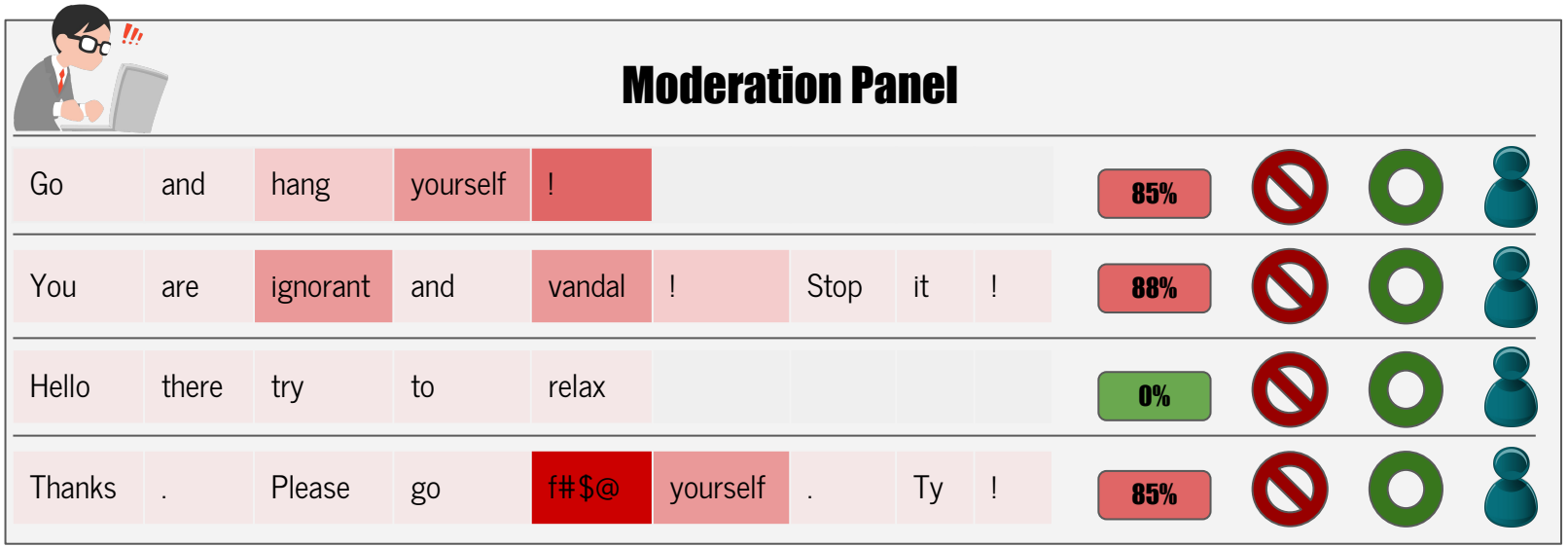


Moderation Panel

Go	and	hang	yourself	!					85%	
You	are	ignorant	and	vandal	!	Stop	it	!	88%	
Hello	there	try	to	relax					0%	
Thanks	.	Please	go	f#$@	yourself	.	Ty	!	85%	

Adding user-specific information

In **other work** (Abusive Language Workshop @ ACL-2017, EMNLP-2017) we obtained **SOTA results** using **RNN-based** methods, **considering only the text** of the comments.



Adding user-specific information

In **other work** (Abusive Language Workshop @ ACL-2017, EMNLP-2017) we obtained **SOTA results** using **RNN-based** methods, **considering only the text** of the comments.

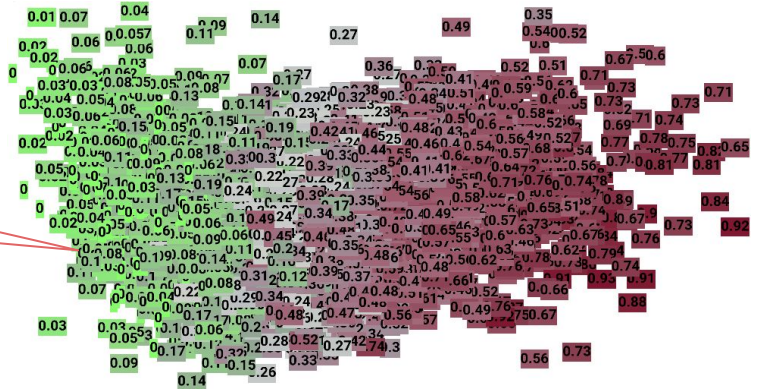
User-specific info (e.g., rejection rate, profanity in previous posts) may lead to **better moderation**. See Dadvar et al. (2013), Waseem et al. (2016), Cheng et al. (2015), Lee et al. (2014), Napoles et al. (2017).

Moderation Panel

Go	and	hang	yourself	!				85%	
You	are	ignorant	and	vandal	!	Stop	it	!	88%
Hello	there	try	to	relax					0%
Thanks	.	Please	go	f#$@	yourself	.	Ty	!	85%

Adding user-specific information

As a **first step** towards **user-specific** info, here we add **user embeddings**.



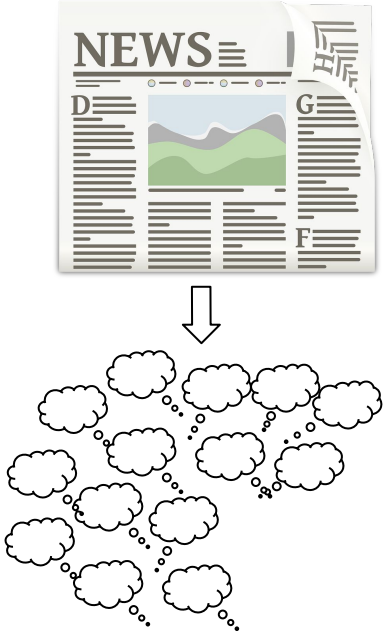
In **other work** (Abusive Language Workshop @ ACL-2017, EMNLP-2017) we obtained **SOTA results** using **RNN-based** methods, **considering only the text** of the comments.

User-specific info (e.g., rejection rate, profanity in previous posts) may lead to **better moderation**. See Dadvar et al. (2013), Waseem et al. (2016), Cheng et al. (2015), Lee et al. (2014), Napoles et al. (2017).

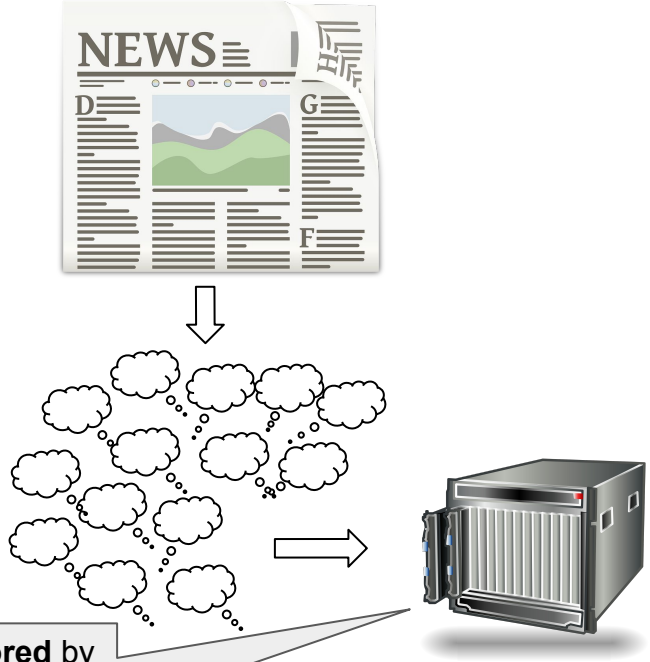
Moderation Panel

Go	and	hang	yourself	!				85%	
You	are	ignorant	and	vandal	!	Stop	it	!	88%
Hello	there	try	to	relax				0%	
Thanks	.	Please	go	f#$@	yourself	.	Ty	!	85%

Semi-automatic comment moderation



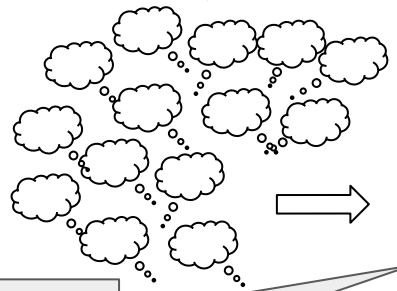
Semi-automatic comment moderation

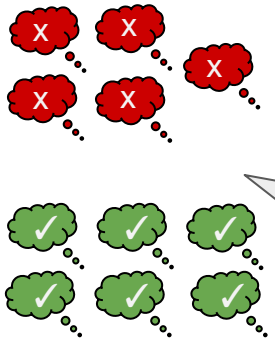


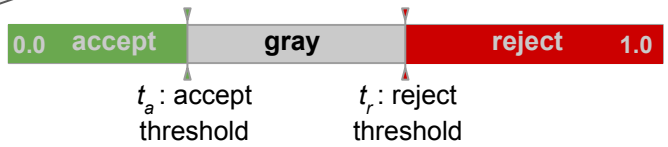
All comments are **scored** by a system.

Score: **how probable** it is for a comment to be **abusive**.

Semi-automatic comment moderation







All comments are **scored** by a system.

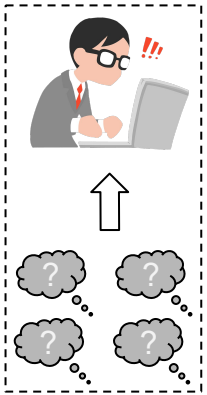
Score: **how probable** it is for a comment to be **abusive**.

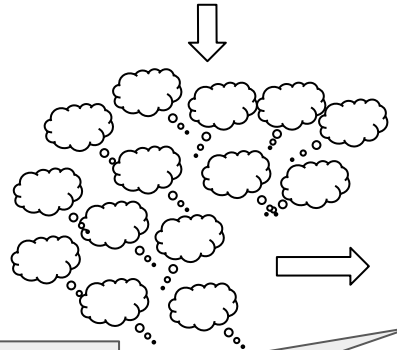
Automatically:

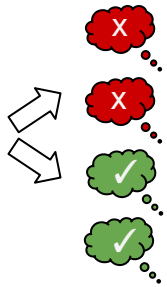
Comments with **high scores** get **rejected**.

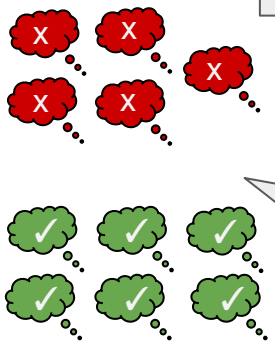
Comments with **low scores** get **accepted**.

Semi-automatic comment moderation









Gray comments, with **intermediate scores**, are sent to a human **moderator** to decide.

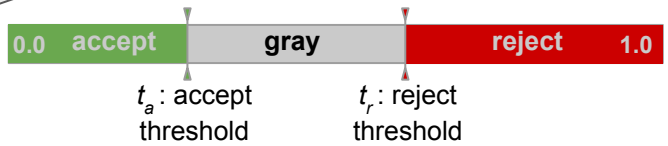
Automatically:

Comments with **high scores** get **rejected**.

Comments with **low scores** get **accepted**.

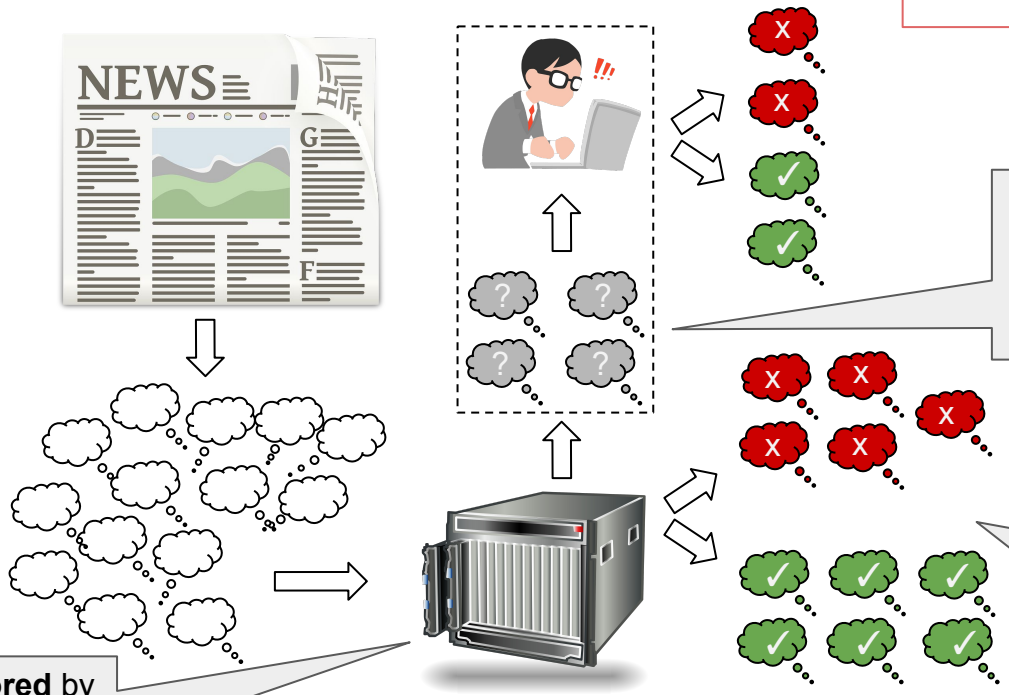
All comments are **scored** by a system.

Score: **how probable** it is for a comment to be **abusive**.



Semi-automatic comment moderation

For more **details** on **semi-automated moderation** see our paper at **EMNLP-2017 (main)**.



Gray comments, with **intermediate scores**, are sent to a human **moderator** to decide.

Automatically:

Comments with **high scores** get **rejected**.

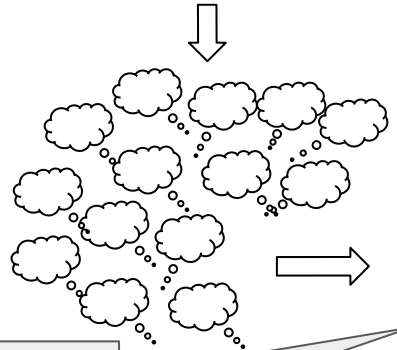
Comments with **low scores** get **accepted**.

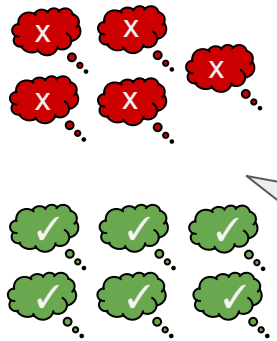
All comments are **scored** by a system.

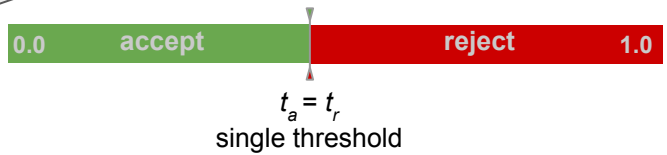
Score: **how probable** it is for a comment to be **abusive**.

Automatic comment moderation

For simplicity, **here** we consider **only fully automatic moderation** (no comments sent to human moderator), but **methods applicable** to **semi-automatic moderation** too.







All comments are **scored** by a system.

Score: **how probable** it is for a comment to be **abusive**.

Automatically:

Comments with **high scores** get **rejected**.

Comments with **low scores** get **accepted**.

Gazzetta dataset[1]

Approx. **1.6M user comments** (**accepted**, **rejected**) from the **Gazzetta Greek sports** news portal**.** Including **user ids**.

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments (accepted, rejected)** from the **Gazzetta Greek sports** news portal**.** Including **user ids**.

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments (accepted, rejected)** from the **Gazzetta Greek sports** news portal**.** Including **user ids**.

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments** (**accepted**, **rejected**) from the **Gazzetta Greek sports** news portal**.** Including **user ids**.

T(u) : Number of **training comments** posted **by user *u***
R(u) : **Rejection rate** of **user u** on training data

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments** (**accepted**, **rejected**) from the **Gazzetta Greek sports** news portal**.** Including **user ids**.

T(u) : Number of **training comments** posted **by user *u***
R(u) : **Rejection rate** of **user u** on training data
Green : *T(u)* >10, ***R(u)*** **≤ 0.33**

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments** (**accepted**, **rejected**) from the **Gazzetta Greek sports** news portal. Including **user ids**.

T(u) : Number of **training comments** posted **by user *u***
R(u) : **Rejection rate** of **user u** on training data
Green : *T(u)* >10, ***R(u)*** **≤ 0.33**
Yellow : *T(u)* >10, **0.33 < *R(u)*** **< 0.66**

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments** (**accepted**, **rejected**) from the **Gazzetta Greek sports** news portal. Including **user ids**.

T(u) : Number of **training comments** posted **by user *u***
R(u) : **Rejection rate** of **user u** on training data
Green : $T(u) > 10$, $R(u) \leq 0.33$
Yellow : $T(u) > 10$, $0.33 < R(u) < 0.66$
Red : $T(u) > 10$, $R(u) \geq 0.66$

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments** (**accepted**, **rejected**) from the **Gazzetta Greek sports** news portal. Including **user ids**.

T(u) : Number of **training comments** posted **by user *u***
R(u) : **Rejection rate** of **user u** on training data
Green : *T(u)* >10, ***R(u)*** ≤ **0.33**
Yellow : *T(u)* >10, **0.33** < ***R(u)*** < **0.66**
Red : *T(u)* >10, ***R(u)*** ≥ **0.66**
Unknown : ***T(u)*** ≤ **10**

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments (accepted, rejected)** from the **Gazzetta Greek sports** news portal. Including **user ids**.

T(u) : Number of **training comments** posted **by user *u***
R(u) : **Rejection rate** of **user u** on training data
Green : $T(u) > 10$, $R(u) \leq 0.33$
Yellow : $T(u) > 10$, $0.33 < R(u) < 0.66$
Red : $T(u) > 10$, $R(u) \geq 0.66$
Unknown : $T(u) \leq 10$

Most comments are posted by **Green users**.

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments (accepted, rejected)** from the **Gazzetta Greek sports** news portal. Including **user ids**.

T(u) : Number of **training comments** posted **by user *u***
R(u) : **Rejection rate** of **user u** on training data
Green : $T(u) > 10$, $R(u) \leq 0.33$
Yellow : $T(u) > 10$, $0.33 < R(u) < 0.66$
Red : $T(u) > 10$, $R(u) \geq 0.66$
Unknown : $T(u) \leq 10$

Most comments are posted by **Green users**.

A lot of comments are posted by **Yellow users**.

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments (accepted, rejected)** from the **Gazzetta Greek sports** news portal. Including **user ids**.

T(u) : Number of **training comments** posted **by user *u***
R(u) : **Rejection rate** of **user u** on training data
Green : $T(u) > 10$, $R(u) \leq 0.33$
Yellow : $T(u) > 10$, $0.33 < R(u) < 0.66$
Red : $T(u) > 10$, $R(u) \geq 0.66$
Unknown : $T(u) \leq 10$

Most comments are posted by **Green users**.

A lot of comments are posted by **Yellow users**.

Very few comments are posted by **Red users**.

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments (accepted, rejected)** from the **Gazzetta Greek sports** news portal. Including **user ids**.

T(u) : Number of **training comments** posted **by user *u***
R(u) : **Rejection rate** of **user u** on training data
Green : $T(u) > 10$, $R(u) \leq 0.33$
Yellow : $T(u) > 10$, $0.33 < R(u) < 0.66$
Red : $T(u) > 10$, $R(u) \geq 0.66$
Unknown : $T(u) \leq 10$

Dataset/Split	Individual Users Per User Type				Total
	Green	Yellow	Red	Unknown	
G-TRAIN	4,451	3,472	251	21,865 → 1	8,175
G-DEV	1,631	1,218	64	1,281 → 1	2,914
G-TEST	1,654	1,203	67	1,254 → 1	2,925

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

Approx. **1.6M user comments** (**accepted**, **rejected**) from the **Gazzetta Greek sports** news portal. Including **user ids**.

T(u) : Number of **training comments** posted **by user *u***
R(u) : **Rejection rate** of **user u** on training data
Green : $T(u) > 10$, $R(u) \leq 0.33$
Yellow : $T(u) > 10$, $0.33 < R(u) < 0.66$
Red : $T(u) > 10$, $R(u) \geq 0.66$
Unknown : $T(u) \leq 10$

Dataset/Split	Individual Users Per User Type				Total
	Green	Yellow	Red	Unknown	
G-TRAIN	4,451	3,472	251	21,865 → 1	8,175
G-DEV	1,631	1,218	64	1,281 → 1	2,914
G-TEST	1,654	1,203	67	1,254 → 1	2,925

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Gazzetta dataset[1]

For a more **detailed analysis** of the **dataset** and **additional datasets** see our paper at **EMNLP 2017** (**main**).

Approx. **1.6M user comments** (**accepted**, **rejected**) from the **Gazzetta Greek sports** news portal. Including **user ids**.

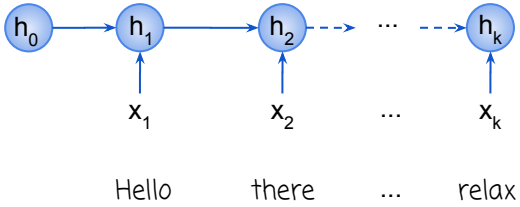
T(u) : Number of **training comments** posted **by user *u***
R(u) : **Rejection rate** of **user u** on training data
Green : $T(u) > 10$, $R(u) \leq 0.33$
Yellow : $T(u) > 10$, $0.33 < R(u) < 0.66$
Red : $T(u) > 10$, $R(u) \geq 0.66$
Unknown : $T(u) \leq 10$

Dataset/Split	Individual Users Per User Type				Total
	Green	Yellow	Red	Unknown	
G-TRAIN	4,451	3,472	251	21,865 → 1	8,175
G-DEV	1,631	1,218	64	1,281 → 1	2,914
G-TEST	1,654	1,203	67	1,254 → 1	2,925

Dataset/Split	Gold Label		Comments Per User Type				Total
	Accepted	Rejected	Green	Yellow	Red	Unknown	
G-TRAIN	960,378 (66%)	489,222 (34%)	724,247 (50%)	585,622 (40%)	43,702 (3%)	96,029 (7%)	1.45M
G-DEV	20,236 (68%)	9,464 (32%)	14,378 (48%)	10,964 (37%)	546 (2%)	3,812 (13%)	29,700
G-TEST	20,064 (68%)	9,636 (32%)	14,559 (49%)	10,681 (36%)	621 (2%)	3,839 (13%)	29,700

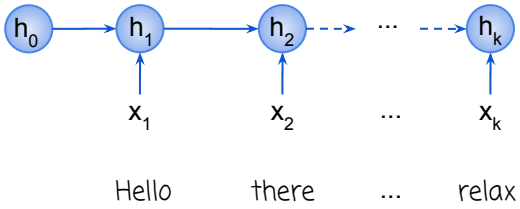
1: From http://www.gazzetta.gr/. Available at: http://www.straintek.com/.

Plain RNN-based moderation



Words are mapped to **embeddings** (word2vec, 300 dimensions).

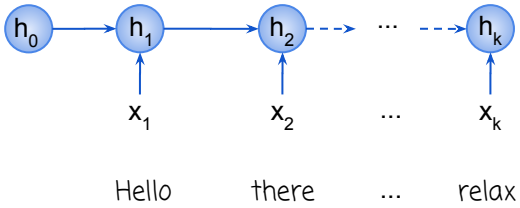
Plain RNN-based moderation



The **RNN states** “summarize” the **words seen**. We use **GRUs** (128 dims).

Words are mapped to **embeddings** (word2vec, 300 dimensions).

Plain RNN-based moderation

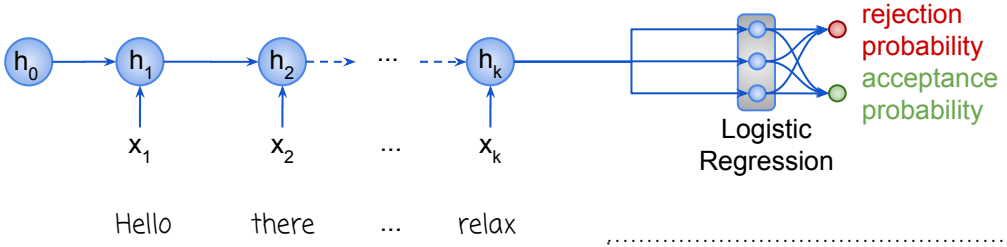


The **last state** of the RNN hopefully represents the **entire comment**.

The **RNN states** “summarize” the **words seen**. We use **GRUs** (128 dims).

Words are mapped to **embeddings** (word2vec, 300 dimensions).

Plain RNN-based moderation



$$P_{\text{RNN}}(reject|c) = \sigma(W_p h_k + b)$$

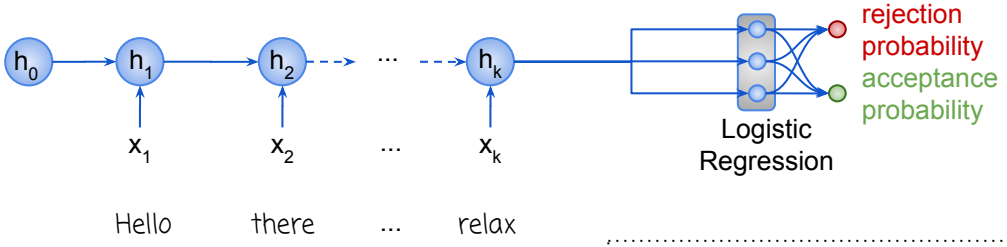
A **Logistic Regression (LR) layer** uses the **last state** of the RNN as a feature vector.

The **last state** hopefully represents the **entire comment**.

The **RNN states** "summarize" the **words seen**. We use **GRUs** (128 dims).

Words are mapped to **embeddings**.

Plain RNN-based moderation



$$P_{\text{RNN}}(reject|c) = \sigma(W_p h_k + b)$$

The **last state** hopefully represents the **entire comment**.

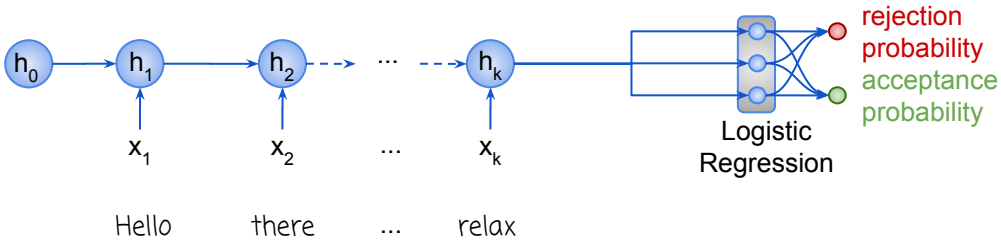
The **RNN states** "summarize" the **words seen**. We use **GRUs** (128 dims).

Words are mapped to **embeddings**.

A **Logistic Regression (LR) layer** uses the **last state** of the RNN as a feature vector.

See our **papers** at **EMNLP-2017** (**main**) and the **Abusive Language Online workshop** of ACL-2017 for **more variants** of the plain **RNN-based** method, including **RNNs with deep self-attention**.

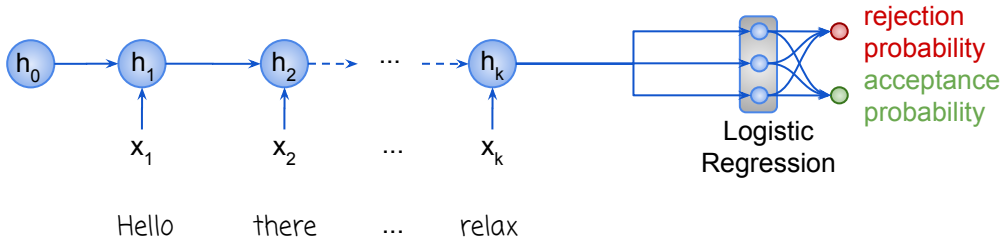
Adding user-specific information



$$P_{\text{RNN}}(reject|c) = \sigma(W_p h_k + b)$$

	User specific	User type
Biases		
Embeddings		

Adding user-specific info



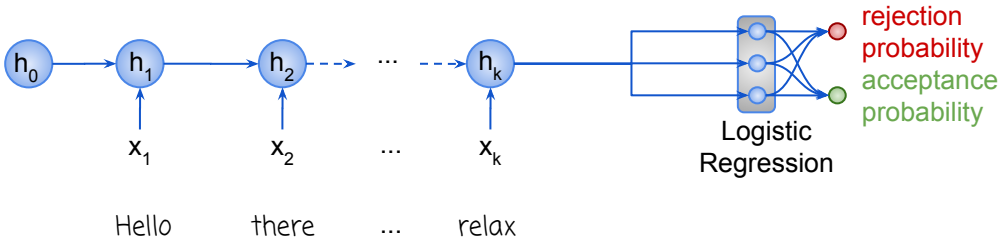
$$P_{\text{RNN}}(reject|c) = \sigma(W_p h_k + b)$$

$$P_{ub\text{RNN}}(reject|c) = \sigma(W_p h_k + \boxed{b_u})$$

User-specific biases (1 dim)

	User specific	User type
Biases	**ubRNN**	
Embeddings		

Adding user-specific information



$$P_{\text{RNN}}(reject|c) = \sigma(W_p h_k + b)$$

$$P_{ub\text{RNN}}(reject|c) = \sigma(W_p h_k + b_u)$$

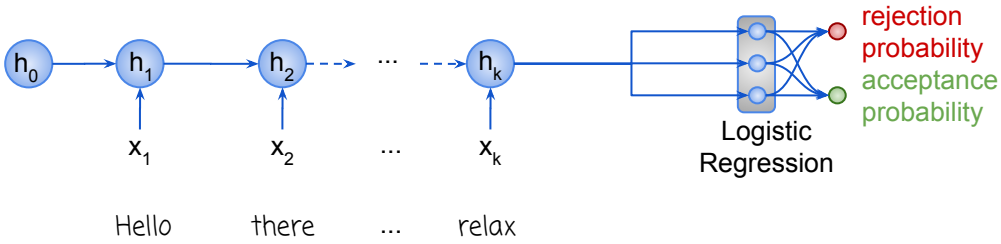
$$P_{ue\text{RNN}}(reject|c) = \sigma(W_p h_k + \boxed{W_v v_u + b})$$

User-specific biases (1 dim)

User-specific embeddings (300 dimensions)

	User specific	User type
Biases	ubRNN	
Embeddings	**ueRNN**	

Adding user-specific information



$$P_{\text{RNN}}(reject|c) = \sigma(W_p h_k + b)$$

$$P_{ub\text{RNN}}(reject|c) = \sigma(W_p h_k + b_u)$$

$$P_{ue\text{RNN}}(reject|c) = \sigma(W_p h_k + W_v v_u + b)$$

$$P_{te\text{RNN}}(reject|c) = \sigma(W_p h_k + \boxed{W_v v_t} + b)$$

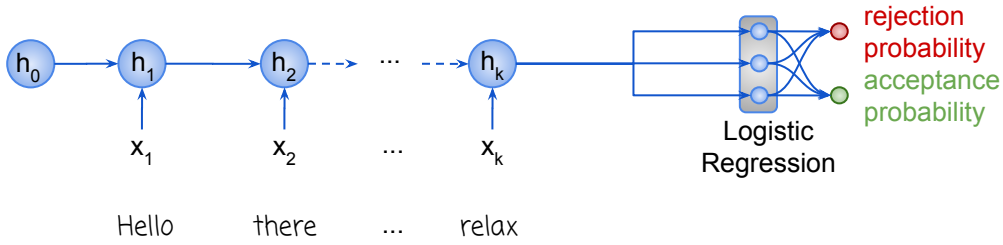
User-specific biases (1 dim)

User-specific embeddings (300 dimensions)

User-type embeddings (300 dimensions)

	User specific	User type
Biases	ubRNN	
Embeddings	**ueRNN**	**teRNN**

Adding user-specific information



$$P_{\text{RNN}}(reject|c) = \sigma(W_p h_k + b)$$

$$P_{ub\text{RNN}}(reject|c) = \sigma(W_p h_k + b_u)$$

$$P_{ue\text{RNN}}(reject|c) = \sigma(W_p h_k + W_v v_u + b)$$

$$P_{te\text{RNN}}(reject|c) = \sigma(W_p h_k + W_v v_t + b)$$

$$P_{tb\text{RNN}}(reject|c) = \sigma(W_p h_k + \boxed{b_t})$$

User-specific biases (1 dim)

User-specific embeddings (300 dimensions)

User-type embeddings (300 dimensions)

User type biases (1 dim)

	User specific	User type
Biases	ubRNN	**tbRNN**
Embeddings	**ueRNN**	**teRNN**

Baselines

T(u)	: Number of **training comments** posted **by user *u***
R(u)	: **Rejection rate** of **user u** on training data
t	: The **type** of **user *u***

uBase

$$P_{u\text{BASE}}(reject|c) = \begin{cases} R(u), \text{ if } T(u) > 10 \\ 0.5, \text{ if } T(u) \leq 10 \end{cases}$$

Baselines

T(u)	: Number of **training comments** posted **by user *u***
R(u)	: **Rejection rate** of **user u** on training data
t	: The **type** of **user *u***

uBase

$$P_{u\text{BASE}}(reject|c) = \begin{cases} R(u), \text{ if } T(u) > 10 \\ 0.5, \text{ if } T(u) \leq 10 \end{cases}$$

tBase

$$P_{t\text{BASE}}(reject|c) = \begin{cases} 1, \text{if } t \text{ is Red} \\ 0.5, \text{if } t \text{ is Yellow} \\ 0.5, \text{if } t \text{ is Unknown} \\ 0, \text{if } t \text{ is Green} \end{cases}$$

Results

AUC of **ROC** (std error of 3 repetitions in brackets), considering **multiple classification thresholds**.

System	G-DEV	G-TEST
*ue*RNN	**80.68 (±0.11)**	**80.71 (±0.13)**
*ub*RNN	80.54 (±0.09)	80.53 (±0.08)
*te*RNN	80.37 (±0.05)	80.41 (±0.09)
*tb*RNN	80.33 (±0.12)	80.32 (±0.05)
RNN	79.40 (±0.08)	79.24 (±0.05)
*u*BASE	67.61	68.57
*t*BASE	63.16	63.82

Results

AUC of **ROC** (std error of 3 repetitions in brackets), considering **multiple classification thresholds**.

System	G-DEV	G-TEST
*ue*RNN	**80.68 (±0.11)**	**80.71 (±0.13)**
*ub*RNN	80.54 (±0.09)	80.53 (±0.08)
*te*RNN	80.37 (±0.05)	80.41 (±0.09)
*tb*RNN	80.33 (±0.12)	80.32 (±0.05)
RNN	79.40 (±0.08)	79.24 (±0.05)
*u*BASE	67.61	68.57
*t*BASE	63.16	63.82

RNN is **always improved** when **user information** is **added**.

Results

User-specific info is **better** than **user type** info.

RNN is **always improved** when **user information** is **added**.

AUC of **ROC** (std error of 3 repetitions in brackets), considering **multiple classification thresholds**.

System	G-DEV	G-TEST
*ue*RNN	**80.68 (±0.11)**	**80.71 (±0.13)**
*ub*RNN	80.54 (±0.09)	80.53 (±0.08)
*te*RNN	80.37 (±0.05)	80.41 (±0.09)
*tb*RNN	80.33 (±0.12)	80.32 (±0.05)
RNN	79.40 (±0.08)	79.24 (±0.05)
*u*BASE	67.61	68.57
*t*BASE	63.16	63.82

Results

User-specific info is **better** than **user type** info.

User-specific or type-specific **embeddings** are **better** than user-specific or type-specific **biases**.

RNN is **always improved** when **user information** is **added**.

AUC of **ROC** (std error of 3 repetitions in brackets), considering **multiple classification thresholds**.

System	G-DEV	G-TEST
*ue*RNN	**80.68 (±0.11)**	**80.71 (±0.13)**
*ub*RNN	80.54 (±0.09)	80.53 (±0.08)
*te*RNN	80.37 (±0.05)	80.41 (±0.09)
*tb*RNN	80.33 (±0.12)	80.32 (±0.05)
RNN	79.40 (±0.08)	79.24 (±0.05)
*u*BASE	67.61	68.57
*t*BASE	63.16	63.82

Results

User-specific info is **better** than **user type** info.

User-specific or type-specific **embeddings** are **better** than user-specific or type-specific **biases**.

RNN is **always improved** when **user information** is **added**.

The **baselines** are **much worse**.

AUC of **ROC** (std error of 3 repetitions in brackets), considering **multiple classification thresholds**.

System	G-DEV	G-TEST
*ue*RNN	**80.68 (±0.11)**	**80.71 (±0.13)**
*ub*RNN	80.54 (±0.09)	80.53 (±0.08)
*te*RNN	80.37 (±0.05)	80.41 (±0.09)
*tb*RNN	80.33 (±0.12)	80.32 (±0.05)
RNN	79.40 (±0.08)	79.24 (±0.05)
*u*BASE	67.61	68.57
*t*BASE	63.16	63.82

Biases and embeddings learned

Biases of Green users
decrease rejection **probability**

User Type	b_t of tbRNN	average b_u of ubRNN
Green	$-0.471\ (\pm 0.007)$	$-0.180\ (\pm 0.024)$
Yellow	$0.198\ (\pm 0.015)$	$0.058\ (\pm 0.022)$
Unknown	$0.256\ (\pm 0.021)$	$0.312\ (\pm 0.011)$
Red	$1.151\ (\pm 0.013)$	$0.387\ (\pm 0.023)$

Biases and embeddings learned

Biases of Green users **decrease** rejection **probability**

Biases of Red users **increase** rejection **probability**

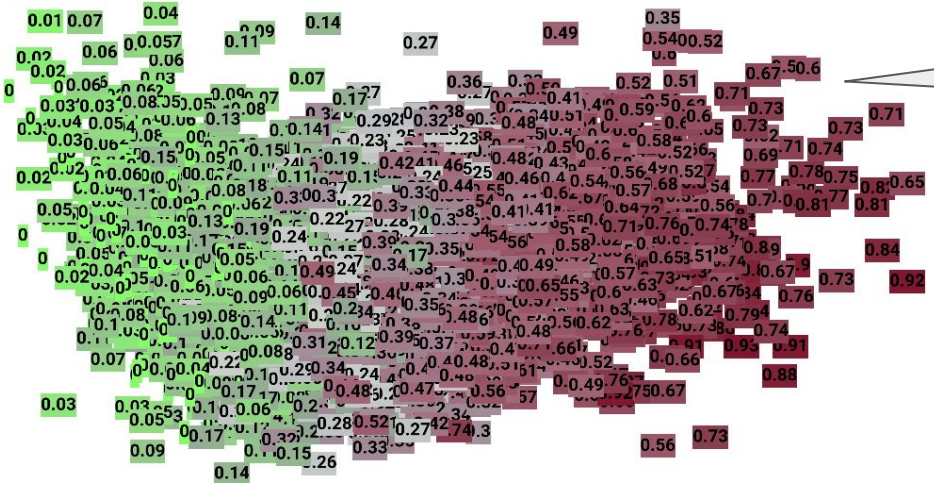
User Type	b_t of tbRNN	average b_u of ubRNN
Green	$-0.471\ (\pm 0.007)$	$-0.180\ (\pm 0.024)$
Yellow	$0.198\ (\pm 0.015)$	$0.058\ (\pm 0.022)$
Unknown	$0.256\ (\pm 0.021)$	$0.312\ (\pm 0.011)$
Red	$1.151\ (\pm 0.013)$	$0.387\ (\pm 0.023)$

Biases and embeddings learned

Biases of Green users **decrease** rejection **probability**

Biases of Red users **increase** rejection **probability**

User Type	b_t of *tb*RNN	average b_u of *ub*RNN
Green	$-0.471\ (\pm 0.007)$	$-0.180\ (\pm 0.024)$
Yellow	$0.198\ (\pm 0.015)$	$0.058\ (\pm 0.022)$
Unknown	$0.256\ (\pm 0.021)$	$0.312\ (\pm 0.011)$
Red	$1.151\ (\pm 0.013)$	$0.387\ (\pm 0.023)$



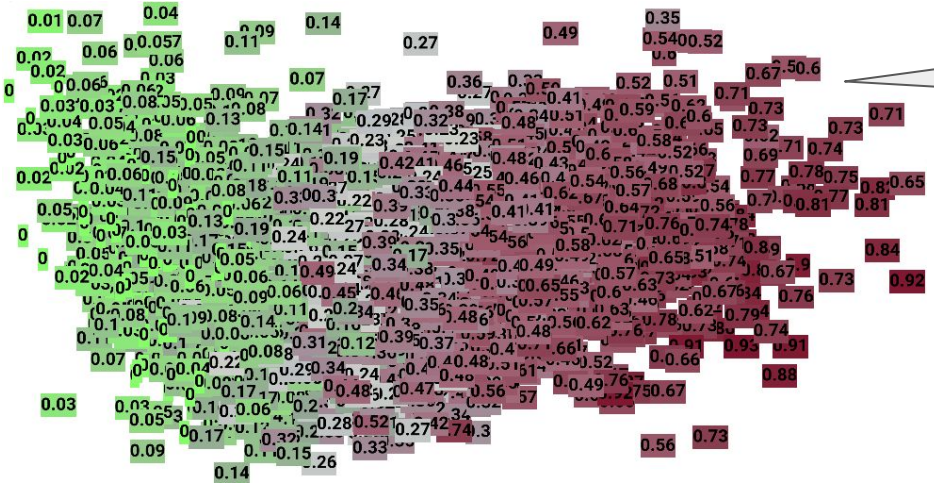
User-specific embeddings (PCA, 2 principal components): **mostly** capture **rejection rates** (but better than simple user-specific biases).

Biases and embeddings learned

Biases of Green users **decrease** rejection **probability**

Biases of Red users **increase** rejection **probability**

User Type	b_t of *tb*RNN	average b_u of *ub*RNN
Green	$-0.471\ (\pm 0.007)$	$-0.180\ (\pm 0.024)$
Yellow	$0.198\ (\pm 0.015)$	$0.058\ (\pm 0.022)$
Unknown	$0.256\ (\pm 0.021)$	$0.312\ (\pm 0.011)$
Red	$1.151\ (\pm 0.013)$	$0.387\ (\pm 0.023)$



User-specific embeddings (PCA, 2 principal components): **mostly** capture **rejection rates** (but better than simple user-specific biases).

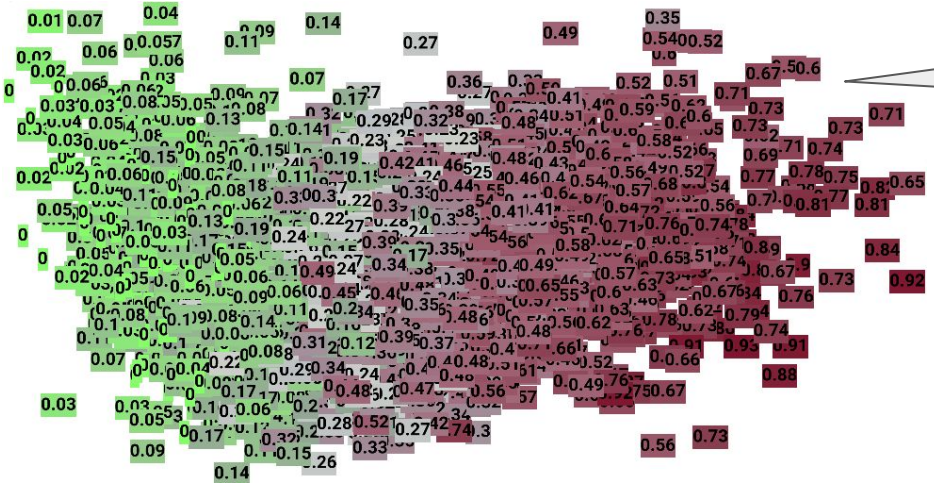
Text of comment	P_{RNN}	P_{ueRNN}
"Ooooh, down to Pireaus…"	0.34	0.72
"Indeed, I know nothing about the filth of Greek soccer."	0.57	0.15

Biases and embeddings learned

Biases of Green users **decrease** rejection **probability**

Biases of Red users **increase** rejection **probability**

User Type	b_t of *tb*RNN	average b_u of *ub*RNN
Green	$-0.471\ (\pm 0.007)$	$-0.180\ (\pm 0.024)$
Yellow	$0.198\ (\pm 0.015)$	$0.058\ (\pm 0.022)$
Unknown	$0.256\ (\pm 0.021)$	$0.312\ (\pm 0.011)$
Red	$1.151\ (\pm 0.013)$	$0.387\ (\pm 0.023)$



User-specific embeddings (PCA, 2 principal components): **mostly** capture **rejection rates** (but better than simple user-specific biases).

Text of comment	P_{RNN}	P_{ueRNN}
"Ooooh, down to Pireaus…"	0.34	0.72
"Indeed, I know nothing about the filth of Greek soccer."	0.57	0.15

Further work

MLP instead of LR to learn **non-linear combinations of RNN states** and **user embeddings**. Like Amir et al. (2016), but they use CNN instead of RNN, and detect sarcasm in tweets.

Further work

MLP instead of LR to learn **non-linear combinations of RNN states** and **user embeddings**. Like Amir et al. (2016), but they use CNN instead of RNN, and detect sarcasm in tweets.

Character-based layers (e.g., for unknown, obfuscated words).

Further work

MLP instead of LR to learn **non-linear combinations of RNN states** and **user embeddings**. Like Amir et al. (2016), but they use CNN instead of RNN, and detect sarcasm in tweets.

Character-based layers (e.g., for unknown, obfuscated words).

Consider **entire threads** and the **original article**, instead of individual comments.

Further work

MLP instead of LR to learn **non-linear combinations of RNN states** and **user embeddings**. Like Amir et al. (2016), but they use CNN instead of RNN, and detect sarcasm in tweets.

Character-based layers (e.g., for unknown, obfuscated words).

Consider **entire threads** and the **original article**, instead of individual comments.

RNN with deep self-attention, **ablation testing**, experiments against **DETOX** (Wulczyn et al. 2017), **other datasets**: see our papers at **EMNLP-2017 (main)** and **Abusive Language Online workshop** of ACL-2017.

Further work

MLP instead of LR to learn **non-linear combinations of RNN states** and **user embeddings**. Like Amir et al. (2016), but they use CNN instead of RNN, and detect sarcasm in tweets.

Character-based layers (e.g., for unknown, obfuscated words).

Consider **entire threads** and the **original article**, instead of individual comments.

RNN with deep self-attention, **ablation testing**, experiments against **DETOX** (Wulczyn et al. 2017), **other datasets**: see our papers at **EMNLP-2017 (main)** and **Abusive Language Online workshop** of ACL-2017.

Highlighting suspicious words (EMNLP-2017 main).

Go	and	hang	yourself	!				
You	are	ignorant	and	vandal	!	Stop	it	!
Thanks	.	Please	go	fuck	yourself	.	ty	!

Thank you! Any questions?

Check http://www.straintek.com/
for papers, data, demos!

sTrainTek
Reader Engagement & Brand Safety

DIGITAL
NEWS
INITIATIVE