Ruddit: Norms of Offensiveness for English Reddit Comments

Rishav Hada[1], Sohi Sudhir[1], Pushkar Mishra[2], Helen Yannakoudakis[3], Saif M. Mohammad[4], Ekaterina Shutova[1]

[1] Institute of Logic, Language and Computation, University of Amsterdam
[2] Facebook AI, London
[3] Dept. of Informatics, King’s College London
[4] National Research Council Canada

The presentation includes some examples of offensive comments.

Offensive Language

- Offensive language has a wide range.
- Humans can distinguish the degrees of offensiveness at fine levels.
- Depends on context.
- Often associated with strong emotions (Jay and Janschewitz, 2008).

- In our work,
  - We focus on the entire spectrum of supportiveness–offensiveness.
  - We aim to find the **commonalities of what most people find offensive**.

Why detect it automatically?

- To study how people communicate offensiveness and supportiveness.
- Can help in developing better Human–Computer Interaction systems.
- Offensive language on social media platforms
  - negatively impacts the mental well-being of their users.
  - makes forums not conducive for a healthy discussion.

Challenges

- What is offensive language?
  - Categories have significant overlaps with each other, creating **ill-defined boundaries**, thus introducing ambiguity.
- Past work mostly uses **discrete labels.**
- Offensiveness is inherently **contextual** (Gao and Huang, 2017).
- Annotator **de-sensitization**.
- Skewed class distribution.

Our Work

- Dataset:
  - **6000 English language Reddit comments**
  - **fine-grained, real-valued scores**
  - between **-1 (maximally supportive) and 1 (maximally offensive)**.

- Used **Comparative annotation** setup (David, 1963)
  - Alleviates **annotation biases** present in standard **rating scales**.
  - Alleviates issues regarding **category definitions**.
  - Mitigates annotator **desensitization**.

- Contains **conversational context** for each comment.

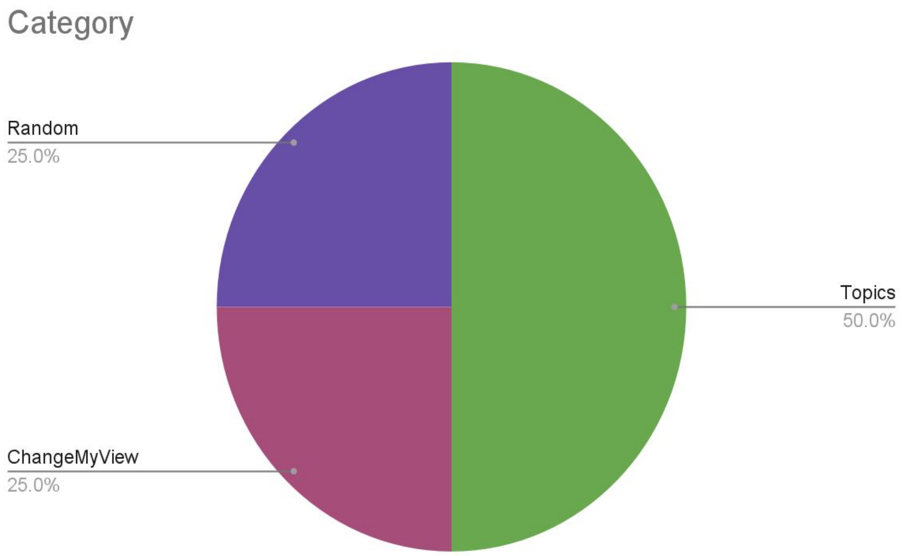
Emotions and Offensiveness

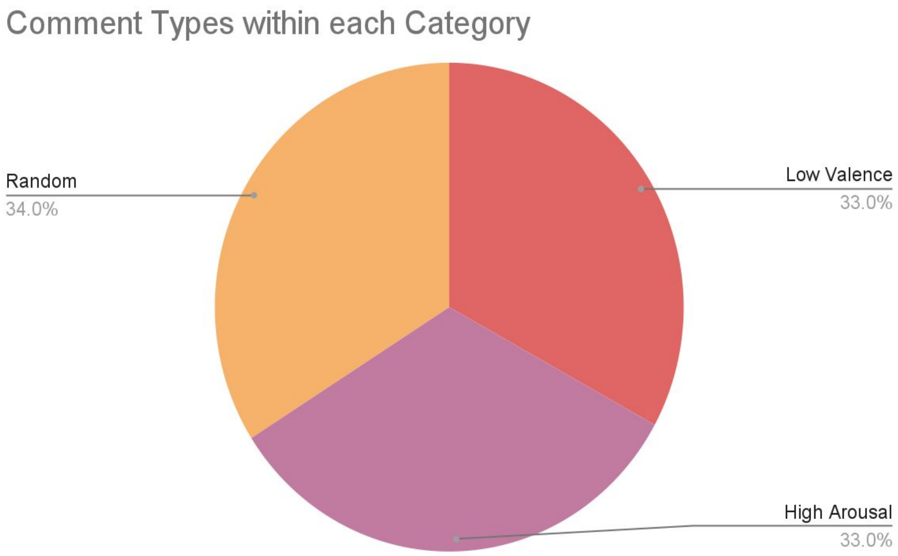
- Offensive behaviour is often associated with strong emotions.

- Primary dimensions of emotion (Russell, 1980, 2003):
  - **Valence(V)** : positive/pleasure – negative/displeasure
  - **Arousal(A)**: excited/active – calm/passive
  - **Dominance(D)** : powerful/have full control – weak/have no control

- We up-sample comments with low-valence (highly negative) or high-arousal words
  - using the NRC VAD lexicon (Mohammad, 2018).
  - 20,000 English words with real-valued scores between 0 & 1 in V, A, D dimensions.

The Hybrid Approach for Data Sampling





The Annotation Task

- Annotations were crowd-sourced on **Amazon Mechanical Turk**.
- Steps to minimize the negative mental impact on the annotators.
- We annotated 5% data ourselves as **gold questions**.
- Worker annotations were discarded if their accuracy on gold questions was below **70%**.

Annotating with Best–Worst Scaling

- **Best–Worst Scaling** (Kiritchenko and Mohammad, 2016, 2017): An efficient form of comparative annotation.
- 2N 4-tuples, each comment seen in 8 different 4-tuples, no two 4-tuples had more than 2 items in common

Q. From the four comments below, choose the comment which is **least offensive** (most supportive) and the comment which is **most offensive** (least supportive).

Least Offensive	Comment	Most Offensive
○	It was a fun day!	○
○	Cool	○
○	Holy mother of God	○
○	You deserve to die!!	○

Annotating with Best–Worst Scaling

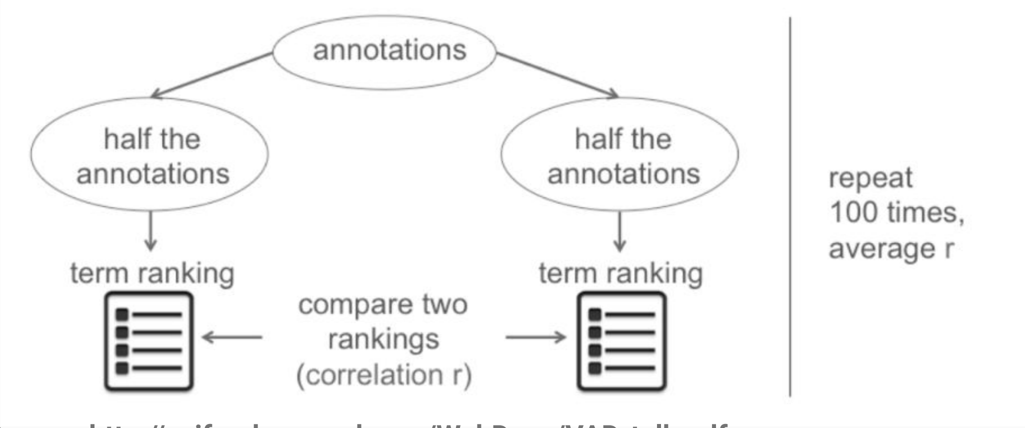
- Using these annotations we can calculate **real-valued scores** of association between the items and the property of interest.

Offensiveness score =

% times comment chosen as most offensive — % times comment chosen as least offensive

Annotation Reliability

Metric used: **Split-Half Reliability** (Cronbach, 1946)

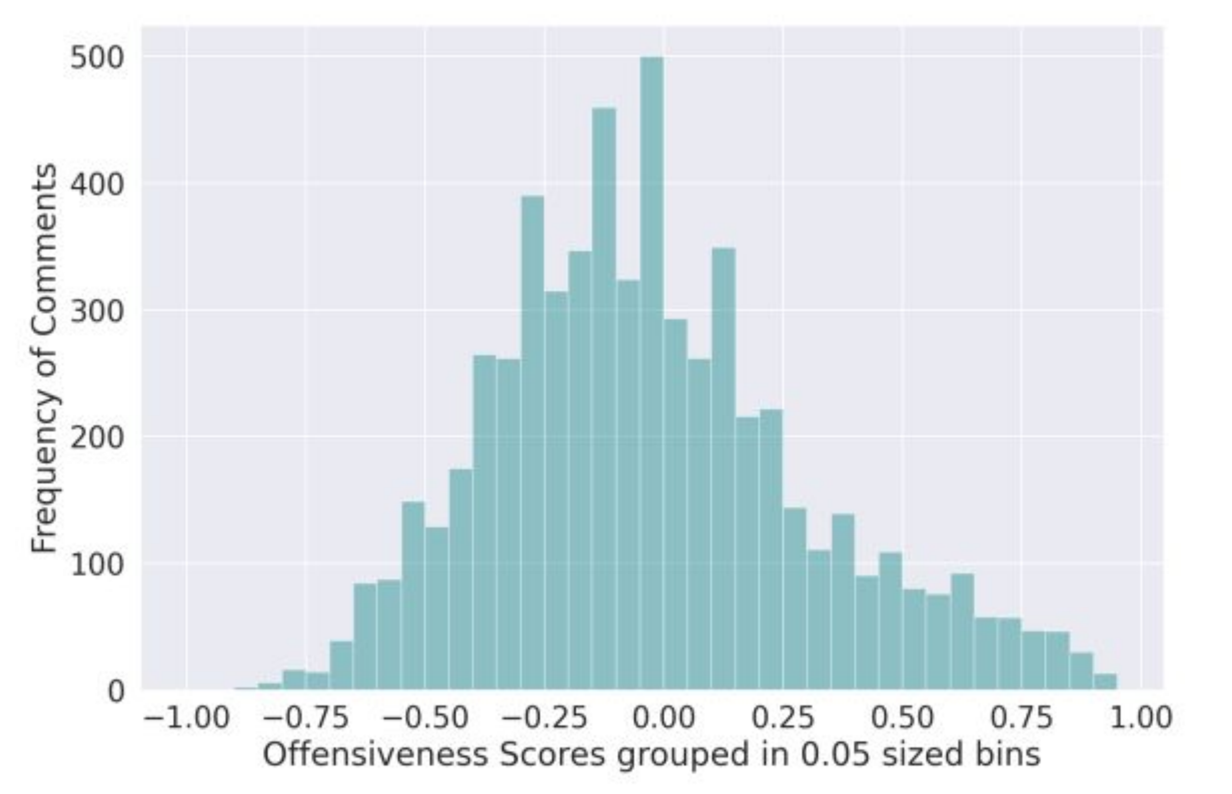


Source: http://saifmohammad.com/WebDocs/VAD-talk.pdf

# Comments	# Annotations per Tuple	# Annotations	# Annotators	SHR Pearson	SHR Spearman
6000	6	95,255	725	0.8818 ± 0.0023	0.8612 ± 0.0029

Data Analysis

Distribution of Scores



Sample Comments

Don’t worry, she’s going to be fine.	-0.75
I see you too are a man of culture;)	-0.604
This is so sexy! Love it!	-0.562
“I live with my ex, but it’s totally cool, we’re just friends”	-0.229
Not sure why Im being down voted? Why does the truth bother so many people?	-0.191
Oh man, that’s a great idea! Nuke a few large cities in the US, this will send the economy to the sky!	0.104
You look like a lesbian mechanic who has a shell collection.	0.438
Don’t forget Vaccines cause autism. And torture is awesome. We should murder the families of terrorists.	0.5
If you support trump kill yourself, painfully.	0.604
Shut the f*ck up bitch. It’s Bernie or Bust nobody is voting for Biden, now get the f*ck out of here you c*nt	0.958

Computational Modeling

- **Task**: Predict the offensiveness score of a given comment.

- Models:
  - **BiLSTM** with 300D GloVe word embeddings.
  - **BERT** (fine-tuning)
  - **HateBERT** (fine-tuning) : a version of BERT pretrained for abusive language detection in English. HateBERT was trained on RAL-E, a large dataset of English language Reddit comments from communities banned for being offensive or hateful.

- We performed **5-fold cross-validation** for each of the models.

Dataset Variations

We created variations of our dataset for a detailed analysis.

- **Ruddit:** The complete dataset.
- **Identity-agnostic:** to investigate the effect of identity terms.
  - Replaced *identity-term** in the comments with *[group]*.

* from list by: Dixon et al. (2018)

Dataset Variations

- **No-swearing:** to investigate the effect of swear words.
  - Removed comments with swear-words from the Cursing Lexicon (Wang et al., 2014).
- **Reduced-range:** to study the modeling of comments in the middle region of the offensiveness scale.
  - Comments from -0.5 to $+0.5$ offensiveness score range.

Results and Analysis

Dataset	HateBERT		BERT		BiLSTM	
	r	MSE	r	MSE	r	MSE
a. Ruddit	0.886 ± 0.003	0.025 ± 0.001	0.873 ± 0.005	0.027 ± 0.001	0.831 ± 0.005	0.035 ± 0.001
b. *Identity-agnostic*	0.883 ± 0.006	0.025 ± 0.001	0.869 ± 0.007	0.027 ± 0.001	0.824 ± 0.007	0.036 ± 0.001

- **HateBERT** outperforms other models.
- Slight drop in performance for identity-agnostic.
  - Not learning to benefit from the association of certain identity terms with specific ranges of offensiveness scores.

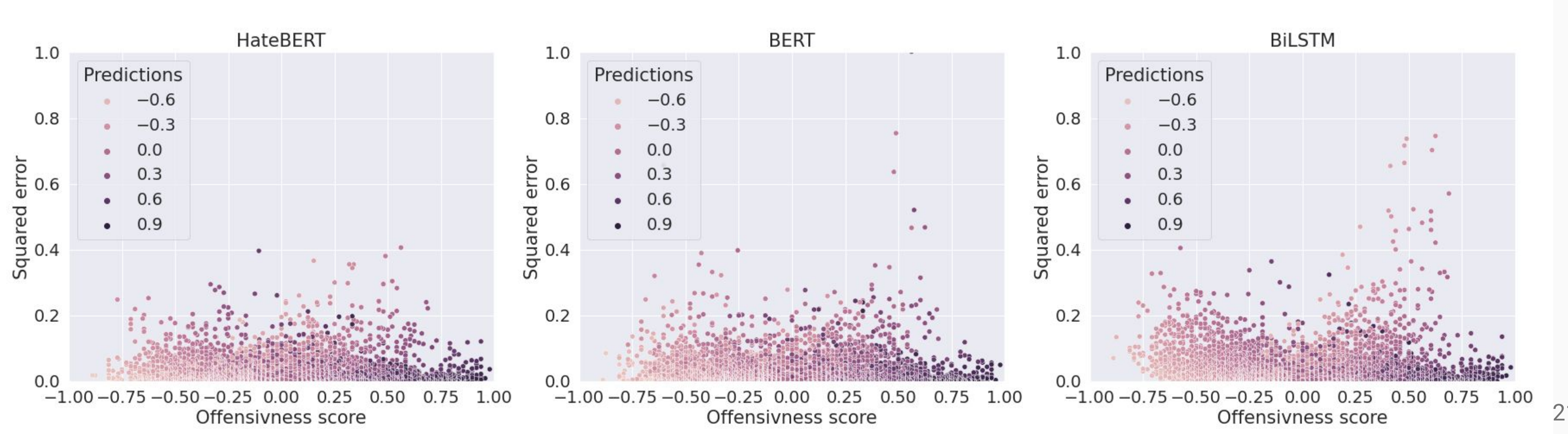
Results and Analysis

Dataset	HateBERT		BERT		BiLSTM	
	r	MSE	r	MSE	r	MSE
c. *No-swearing*	0.808 ± 0.013	0.023 ± 0.001	0.783 ± 0.012	0.027 ± 0.001	0.704 ± 0.014	0.036 ± 0.002
d. *Reduced-range*	0.781 ± 0.014	0.022 ± 0.001	0.757 ± 0.011	0.025 ± 0.001	0.659 ± 0.008	0.033 ± 0.001

- **HateBERT** outperforms other models.
- Drop in performance for no-swearing:
  - Swear words are important indicators but there are other features being learnt!
- Reduced-range: Still an **interesting** and **feasible** task
- Task not just predicting a discrete label but **assigning an offensiveness score.**

Find out more in the paper!

- Best–Worst Scaling procedure
- Sampling and scoring method for the dataset
- The complete annotation procedure
- Data analysis in depth
- Error analysis of the models



Conclusion

- First dataset of online comments annotated for their **degree of offensiveness.**
- Using **BWS** addresses the limitations of traditional rating scales.
- Ratings obtained are **highly reliable (SHR Pearson r≈0.88)**
- We show that **low valence** and **high arousal** comments have a higher correlation with the offensiveness scores.
- We present **benchmark experiments** to **predict offensiveness scores** on our dataset.

Future Work

- More **context dependent annotations.**
- **Use of conversational context** in computational modeling of offensiveness.
- Studying the interaction between offensiveness and emotions in more depth.
- Conducting functional tests on models trained on Ruddit.
    - **HATECHECK**: Functional Tests for Hate Speech Detection Models (Rottger et. al., 2021)

Code and Data available at:

SCAN ME

https://github.com/hadarishav/Ruddit

rishavhada@gmail.com
sohigre@gmail.com

@rishanky