Adobe

"To Target or Not to Target": Identification and Analysis of Abusive Text Using Ensemble of Classifiers

Gaurav Verma, Niyati Chhaya, Vishwa Vinay | Adobe Research, India

ICWSM-2020

DATA CHALLENGE: SAFETY

MAKE IT AN EXPERIENCE

Online Abuse and Hate: Personas

See no evil!

Speak no evil!

Hear no evil!

“Express yourself!”

Three Personas:

- Online Abusers/haters
- Those who want to stay away
- ***Moderators***

Content Moderators and Mental Health

The Guardian

Facebook to pay $52m for failing to protect moderators from 'horrors' of graphic content

BBC

Facebook and YouTube moderators sign PTSD disclosure

THE CONVERSATION

Jennifer Beckett
Lecturer in Media and Communications, University of Melbourne

We need to talk about the mental health of content moderators

September 27, 2018 3.13pm AEST

THE VERGE

THE TRAUMA FLOOR

The secret lives of Facebook moderators in America

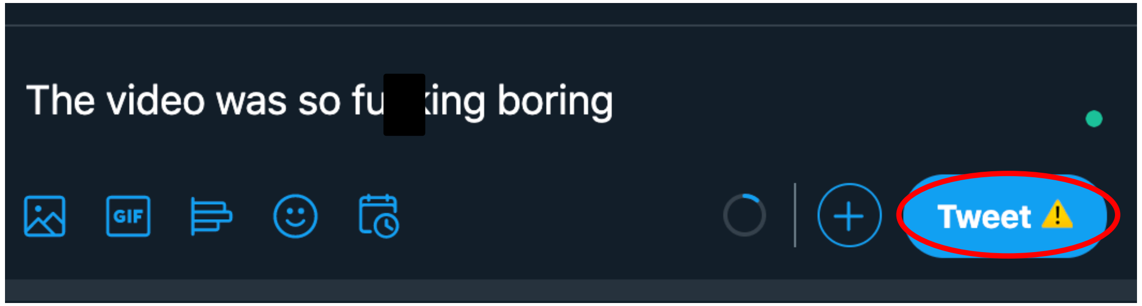
By Casey Newton | @CaseyNewton | Feb 25, 2019, 8:00am EST
Illustrations by Corey Brickley | Photography by Jessica Chou

Systems that can detect online hate and abuse more *accurately*

- *Lesser manual intervention → less impact on mental health of moderators*

Possible Ways to Moderate Content

1. Post-time Warnings



2. Consumption-time Adaptation and Warnings

3. Offline Moderation (after reporting)

Dashboard

23 reports, 3 past incidents, 7 past reports

Prediction: abusive (0.72 confidence score)

Offensive word(s): fu*k (0.92), pathetic (0.16)

> 80 % similarity to 4,320 other abusive posts

An Ideal Automated Moderation System

1. Reliable accuracy
2. Interpretable predictions
3. Human-in-the-loop
   - *Lesser cognitive load*
   - *Minimize exposure* to potentially harmful content

Classifiers that not only *perform well* in terms *of classification metrics*, but also provide *diverse*, yet, *coherent insights* into their predictions.

1. Logistic regression on LIWC features
2. N-gram based Classifier
3. Attention-based BiLSTM Classifier

4. Stacked Ensemble Classifier

LIWC: Linguistic inquiry and word count

Classification Task

- Classification task
  - Twitter Abusive Behavior dataset (Founta et al., 2018)
  - 4-class classification, ~100,000 examples, class imbalance
  - **normal** (53.85 %), **spam** (27.15 %), **abusive** (14.04 %), **hateful** (4.96 %)

Text properties

Network properties

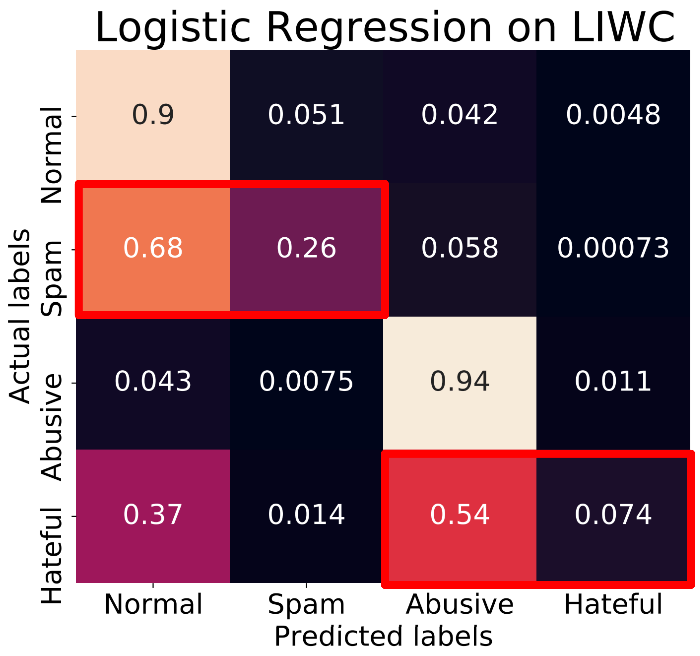
Logistic Regression with LIWC Features

- LIWC Features [1]
  - Categorization of words into psychologically meaningful categories
  - Capture *"attentional focus, emotionality, social relationships, thinking styles, and individual differences"* expressed in language [2]

- Train a logisitic regression classifier on these features and analyse the learned β-coefficients. Good practices:
  - Remove highly correlated features (Pearson correlation coefficient > 0.9); standardize the data; regularization, etc.

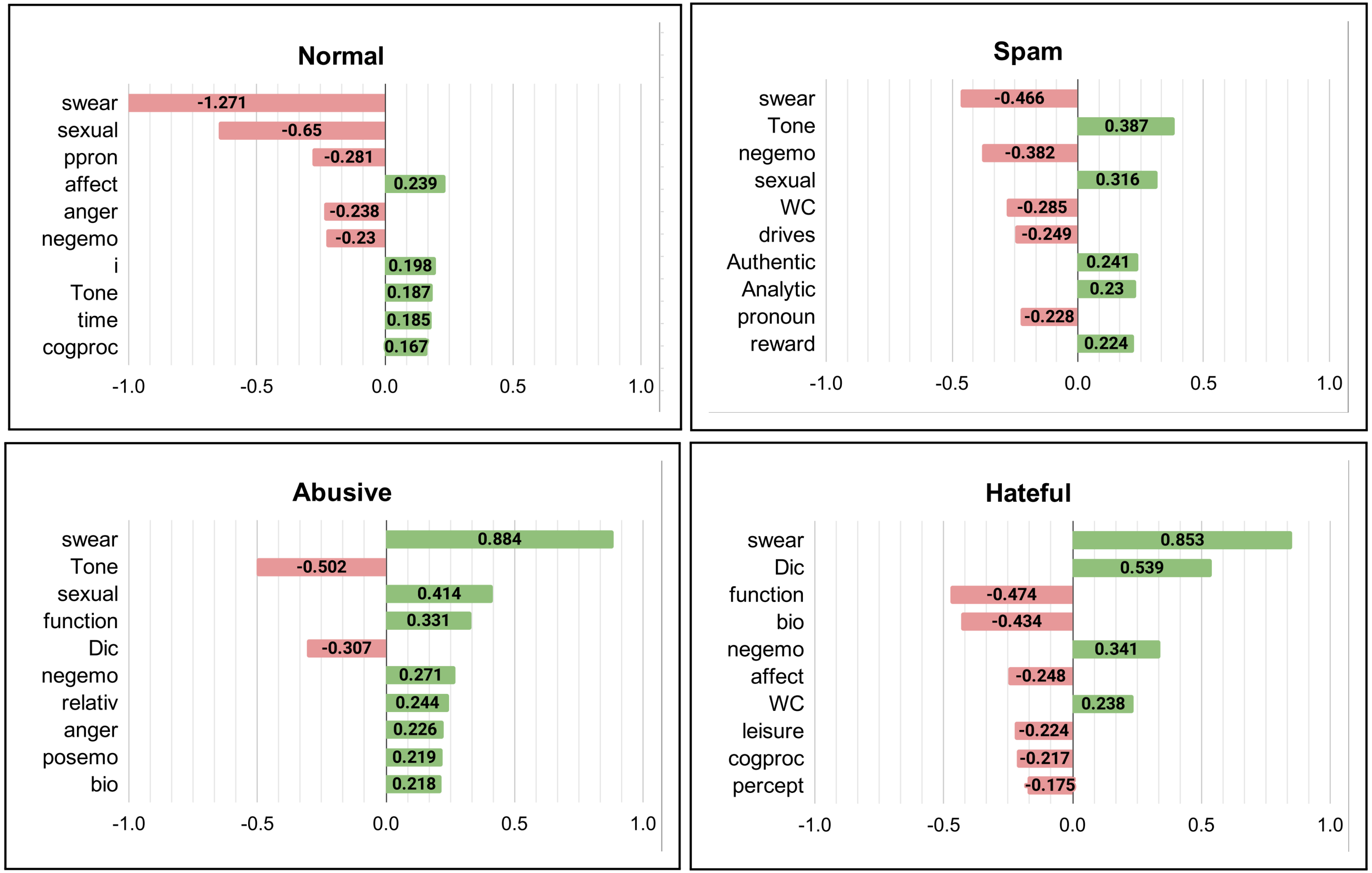
Model	**Accuracy**
LR on LIWC features	0.78

Table 1: Classification accuracy on the test set.



[1] Pennebaker, J.; Francis, M.; and Booth, R. 1999. Linguistic inquiry and word count (LIWC)
[2] Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of Language and Social Psychology.

Logistic Regression with LIWC Features: Insights



Note: Interpret in conjunction with the model performance shown in confusion matrix earlier

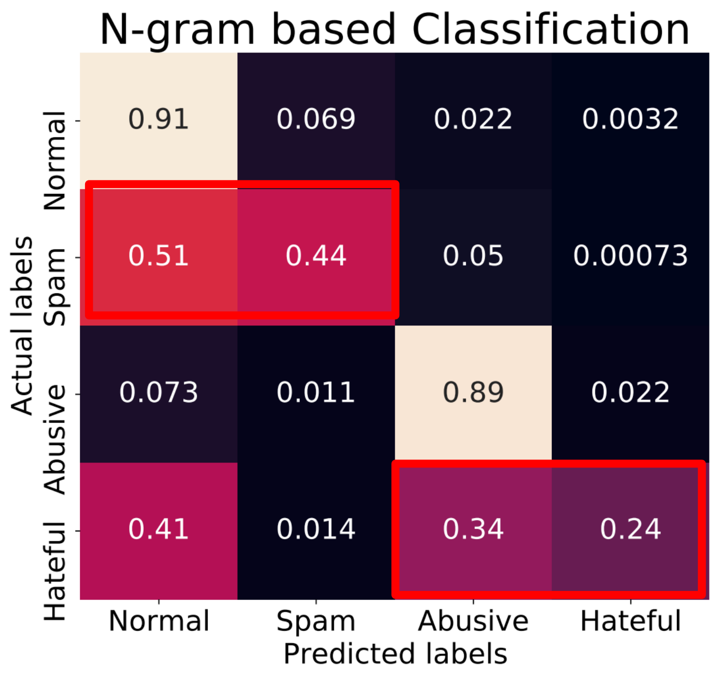
Top-10 learned coefficients based on their absolute values and the corresponding features.

N-gram based Classifier

Model	Accuracy
LR on LIWC features	0.78
N-gram based Classification	0.80

Table 1: Classification accuracy on the test set.

- Bag of n-gram features captures partial information about the local word order [3]
  - Computationally faster, better modelling than bag of words
  - Provides learned *embeddings for the words* in the vocabulary as well as *tweet embeddings*



[3] Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2017. Bag ́ of tricks for efficient text classification. In European Chapter of the ACL (EACL).

N-gram based Classifier: Insights

Nearest-neighbor (NN) querying using word embeddings: output remains offensive, yet *diverse*.

fu*king: as*holes, bullsh*t, su*ks, pen*s, dumba*s, sh*tty
[w2v [4] NN for fu*king: fu@kin, f_ck, f_*_cking, friggin, freakin, fu@ked]

Analogy operations using word embeddings: output has a clear shift from strictly inappropriate toward more acceptable words

(a) fu*king − abuse + normal = boring (w2v: f ** king)
(b) fata*s − hate + normal = pathetic (w2v: sh*thead)
(c) b*tch − hate + normal = nasty (w2v: haters)

[4] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In NeurIPS.

Adobe

N-gram based Classifier: Insights

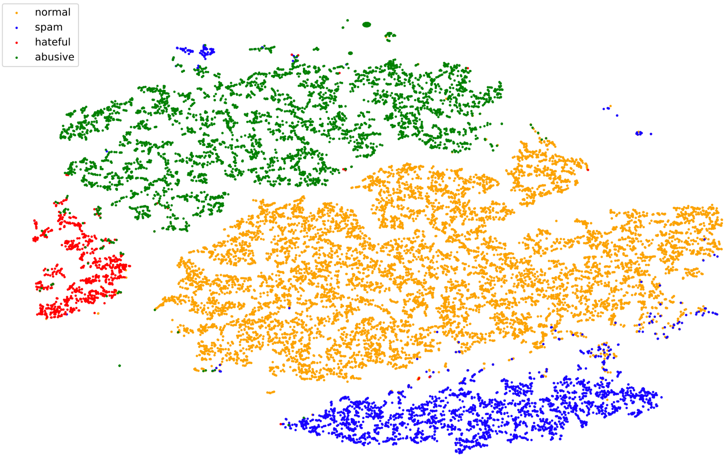
Nearest-neighbor (NN) querying using word embeddings: output remains offensive, yet *diverse*.

fu*king: as*holes, bullsh*t, su*ks, pen*s, dumba*s, sh*tty
[w2v [4] NN for fu*king: fu@kin, f_ck, f_*_cking, friggin, freakin, fu@ked]

Analogy operations using word embeddings: output has a clear shift from strictly inappropriate toward more acceptable words

(a) fu*king − abuse + normal = boring (w2v: f ** king)
(b) fata*s − hate + normal = pathetic (w2v: sh*thead)
(c) b*tch − hate + normal = nasty (w2v: haters)

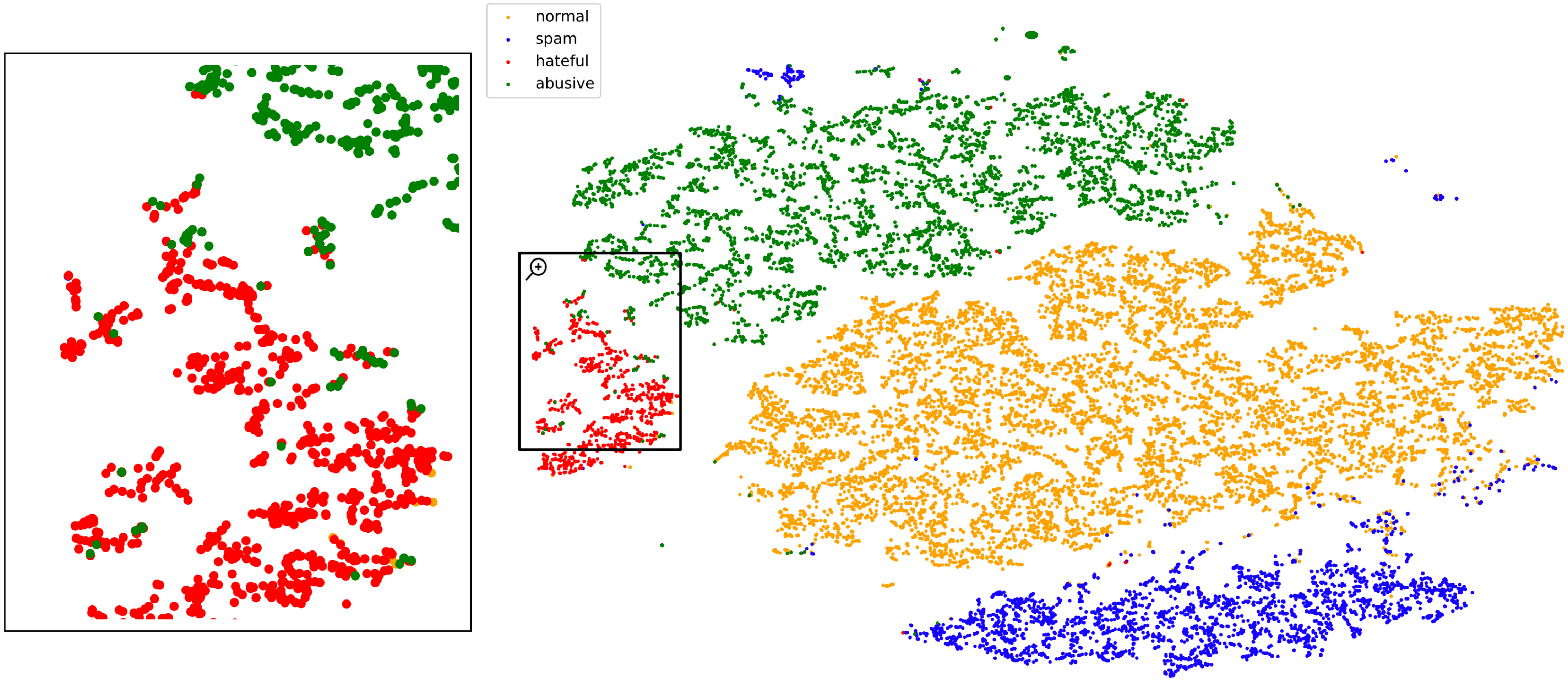
Tweet Embeddings



[4] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In NeurIPS.

N-gram based Classifier: Insights

Tweet Embeddings



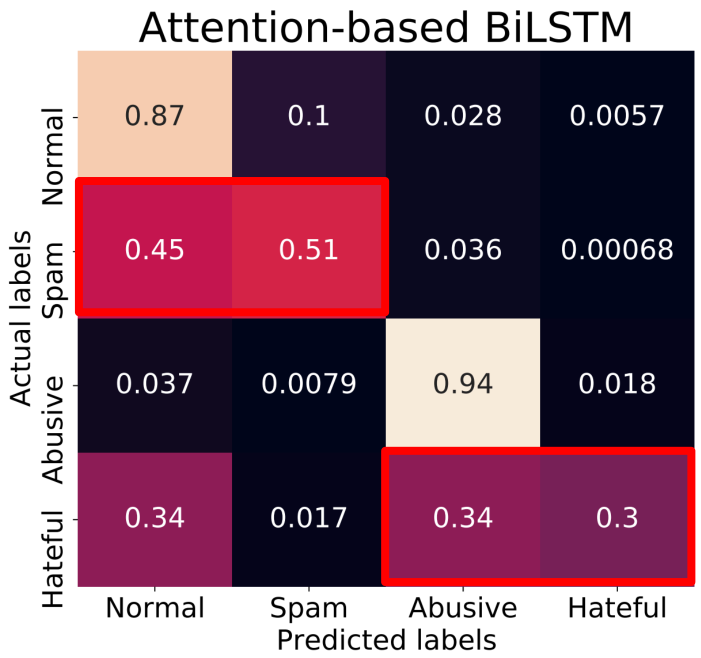
Many abusive tweets have similar embeddings as hateful tweets!

Attention-based Bidirectional LSTM (BiLSTM)

- The attention module allows the model to "attend" to input words while performing classification tasks [5]
  - Learned weights are often used for interpretation

Model	Accuracy
LR on LIWC features	0.78
N-gram based Classification	0.80
Attention-based BiLSTM	0.81

Table 1: Classification accuracy on the test set.



[5] Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; and Xu, B. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In ACL.

Attention-based Bidirectional LSTM (BiLSTM)

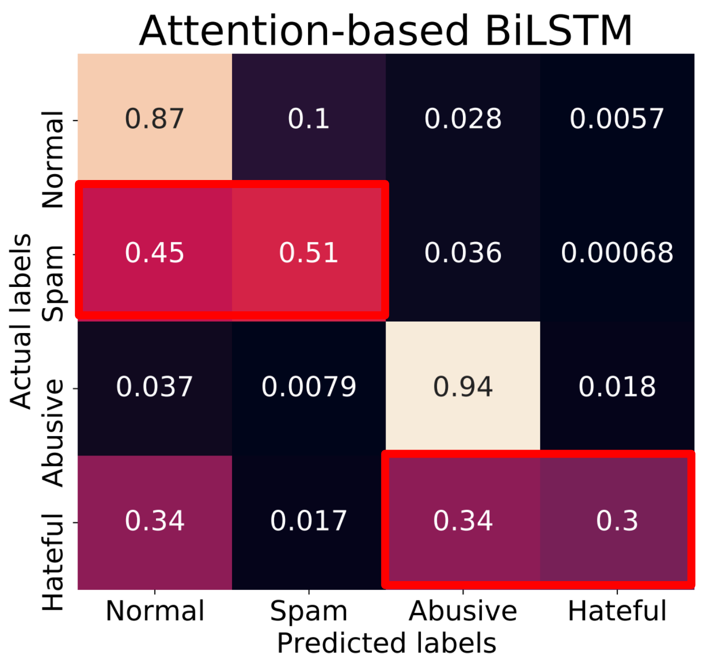
- The attention module allows the model to “attend” to input words while performing classification tasks [5]
    - Learned weights are often used for interpretation

Model	Accuracy
LR on LIWC features	0.78
N-gram based Classification	0.80
Attention-based BiLSTM	0.81

Table 1: Classification accuracy on the test set.

Normal	Spam	Abusive	Hateful
business	hoodies	jack*ss	ret*rds
gather	advertise	fu*king	spitt*ng
snapped	online	bruh	n*zi
holds	store	di*khead	ch*ke
freaking	horoscopes	fat*ss	b*tch

Some of the ***most-attended words*** *for each class*

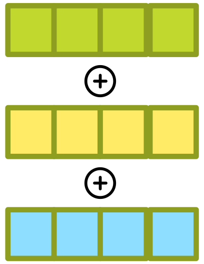


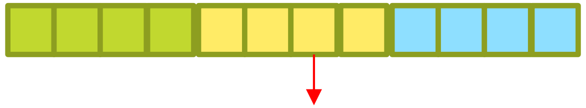
[5] Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; and Xu, B. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In ACL.

Stacked Ensemble

- **General intuition**
  - Take the ***predictions of sufficient diverse models*** (in terms of modelling assumptions), and
  - Train a ***meta model to interpret*** those predictions

- Base models
  - Logistic regression on LIWC features
  - N-gram based Classifier
  - Attention-based BiLSTM
- **Meta model**
  - A simple logistic regression classifier

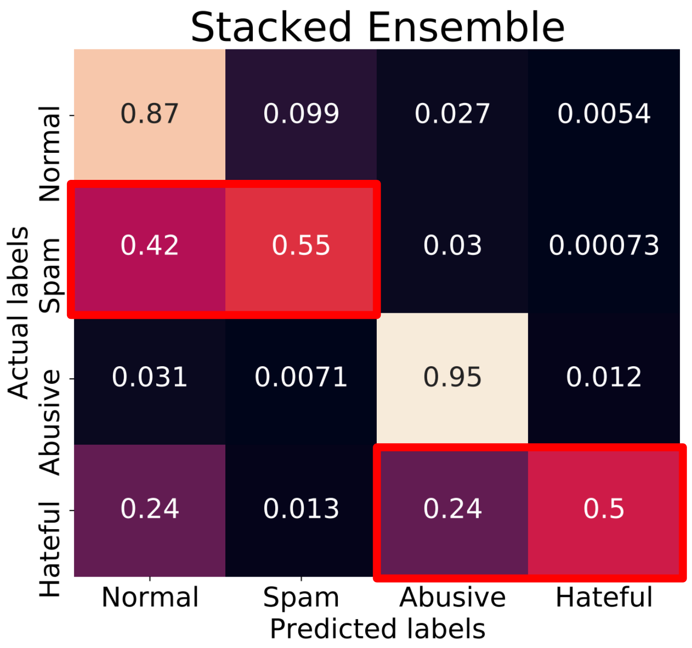




$$y = normal, spam, abusive, hateful$$

Model	**Accuracy**
LR on LIWC features	0.78
N-gram based Classification	0.80
Attention-based BiLSTM	0.81
Stacked Ensemble	0.83

Table 1: Classification accuracy on the test set.



Adobe

Stacked Ensemble: Key Points

- **Overall Performance:**
  - *Comparable performance* to Founta et al. (2019) [6] *without* using user or network-related information
  - Ensemble performs *better than all base models*
  - ***Alleviates*** spam and normal confusion
  - ***Alleviates*** abusive and hateful confusion
- **BUT**, the performance on these fronts is still ***not*** "reliable"

1. Spam and normal confusion

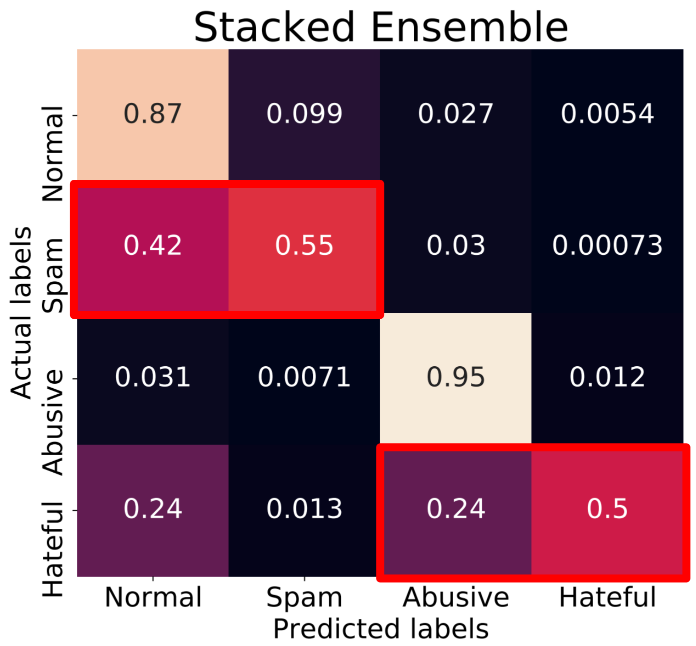
- Can be handled well by incorporating user or network information – bot accounts spam repeatedly, lesser engagement

2. Abusive and hateful confusion

- Differences are more linguistic in nature. Let's discuss more!

Model	**Accuracy**
LR on LIWC features	0.78
N-gram based Classification	0.80
Attention-based BiLSTM	0.81
Stacked Ensemble	0.83

Table 1: Classification accuracy on the test set.



[6] Founta, A. M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; and Leontiadis, I. 2019. A unified deep learning architecture for abuse detection. In ACM Conference on Web Science.

Abusive or Hateful?

Data-specific limitations:
Average number of agreed annotators (out of 5)

- Normal (53.85 %): 3.90
- Spam (27.15 %): 3.47
- Abusive e (14.04 %): 3.53
- Hateful (4.96 %): 2.95

Linguistic Challenges:
Hateful tweets contain specific mention of *targeted groups(s)* [7, 8], whereas abusive tweets do not.

- "some women need to grow the hell up. it's so pathetic." (**hateful**)
- "some people are so pathetic and need to grow the fu*k up!" (**abusive**);

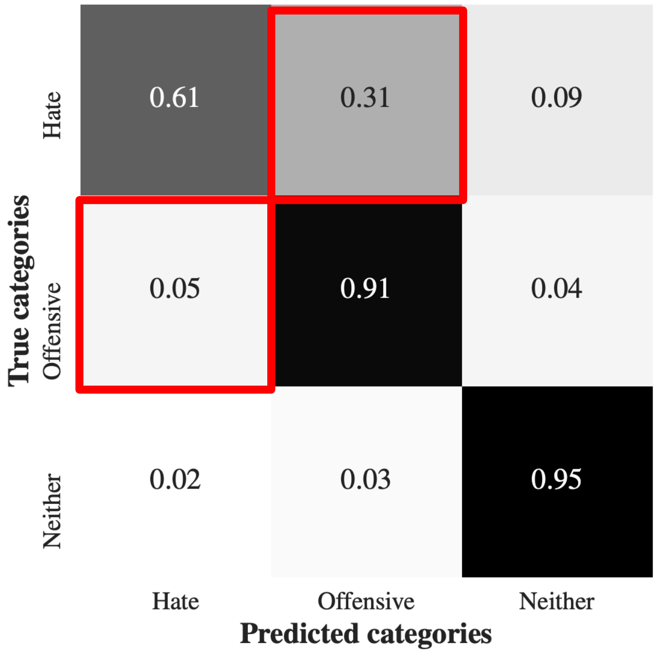


Figure 1: True versus predicted categories

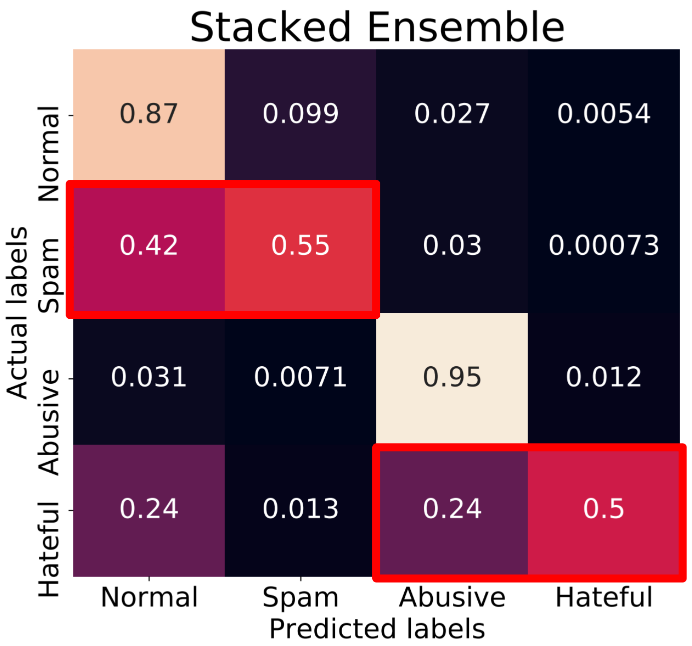
From Davidson et al., 2017 [8]

[7] Founta, A. M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; and Leontiadis, I. 2019. A unified deep learning architecture for abuse detection. In ACM Conference on Web Science.
[8] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

Open Questions

Q1: How to make *language classifiers aware of target group(s)* to allow better distinction between abusive and hateful content?

Q2: How does the *incorporation of user or network-related information* influence classification performance?



{gaverma, nchhaya, vinay}@adobe.com