Perils and Promises of Automated Hate Speech Detection

07 Dec 2020

Roy Ka-Wei LEE & Rui CAO

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

SMU
SINGAPORE MANAGEMENT UNIVERSITY

WARNING: The following tutorial contain act of violence and discrimination that may be disturbing to some participants. Discretion is advised

Image from: http://clipart-library.com/clipart/8iG6GoEjT.htm

Scope

1. Introduction - 5W1H of Online Hate Speech
2. Data Annotation and Collection
3. Traditional Machine Learning Detection Techniques
4. Deep Learning Detection Techniques
5. Open Issues & Opportunities
6. Q&A Discussion

Resources

- Hate Speech Model Zoo Gitlab repo
  - This deck of slides!
  - Code implementation of hate speech detection methods
  - Open source datasets (pre-processed)
  - List of relevant readings

URL: https://gitlab.com/bottle_shop/safe/hate-speech-model-zoo

What is Hate Speech?

What is Hate Speech?

- *"Speech or writing that attacks or threatens a particular group of people, especially on the basis of race, religion or sexual orientation."* (Oxford Dictionary)
- *"We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability."* (Facebook Community Standards - Hate Speech)
- *"You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease."* (Twitter Hateful Conduct Policy)

What is Hate Speech?

- *"Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used."* (*Fortuna and Nunes, 2018*)

Table 2. Content Analysis of Hate Speech Definitions

Source	Hate speech is to incite violence or hate	Hate speech is to attack or diminish	Hate speech has specific targets	Humour has a specific status
EU Code of conduct	Yes	No	Yes	No
ILGA	Yes	No	Yes	No
Scientific paper	No	Yes	Yes	No
Facebook	No	Yes	Yes	Yes
YouTube	Yes	No	Yes	No
Twitter	Yes	Yes	Yes	No

Where to find Online Hate Speech?

- Hate speech can be found in all social platforms
  - Social media: Twitter, YouTube, Instagram, Facebook, etc.
  - Online forums: 4Chan, Reddit, etc.
  - Community Q&A: Quora, etc.
  - Chatapp communities: Telegram, Whatapps, etc.
  - Online games
- Data form: Text and Multimedia (memes, videos, audios etc)

Image from: https://researchoutreach.org/articles/hate-speech-regulation-social-media-intractable-contemporary-challenge/

Why should we be concern?

- A fast-growing and menacing problem
- Online hate speeches may lead to offline hate crimes

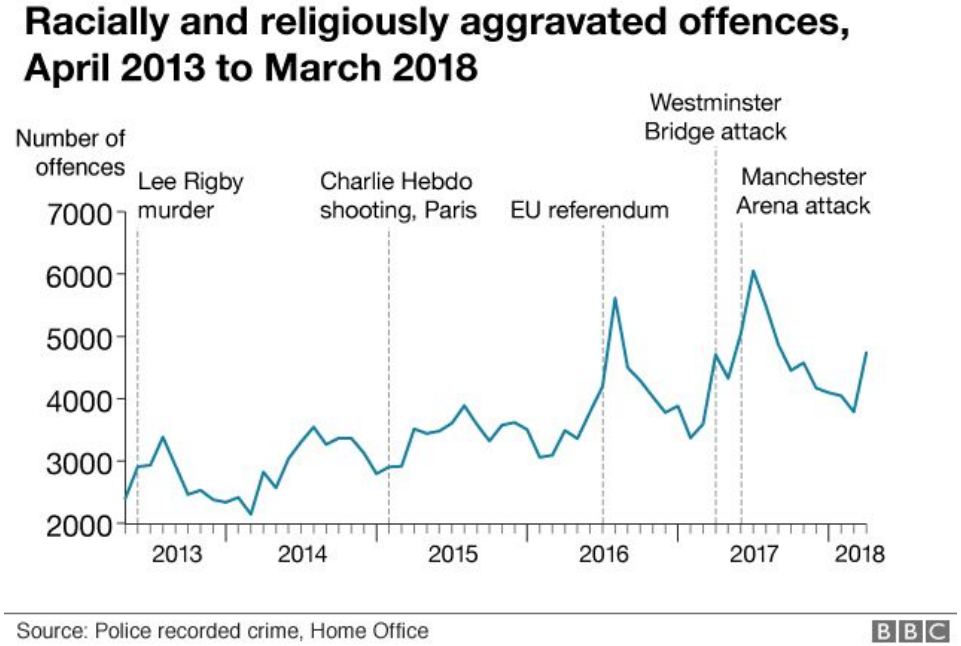


Image from: https://www.bbc.com/news/uk-wales-46552574

Mass shooting at an Orlando gay nightclub: https://theconversation.com/the-orlando-shooting-exploring-the-link-between-hate-crimes-and-terrorism-60992

Why should we be concern?

- Increased attention from governments and law-makers
  - Hate speech vs free speech

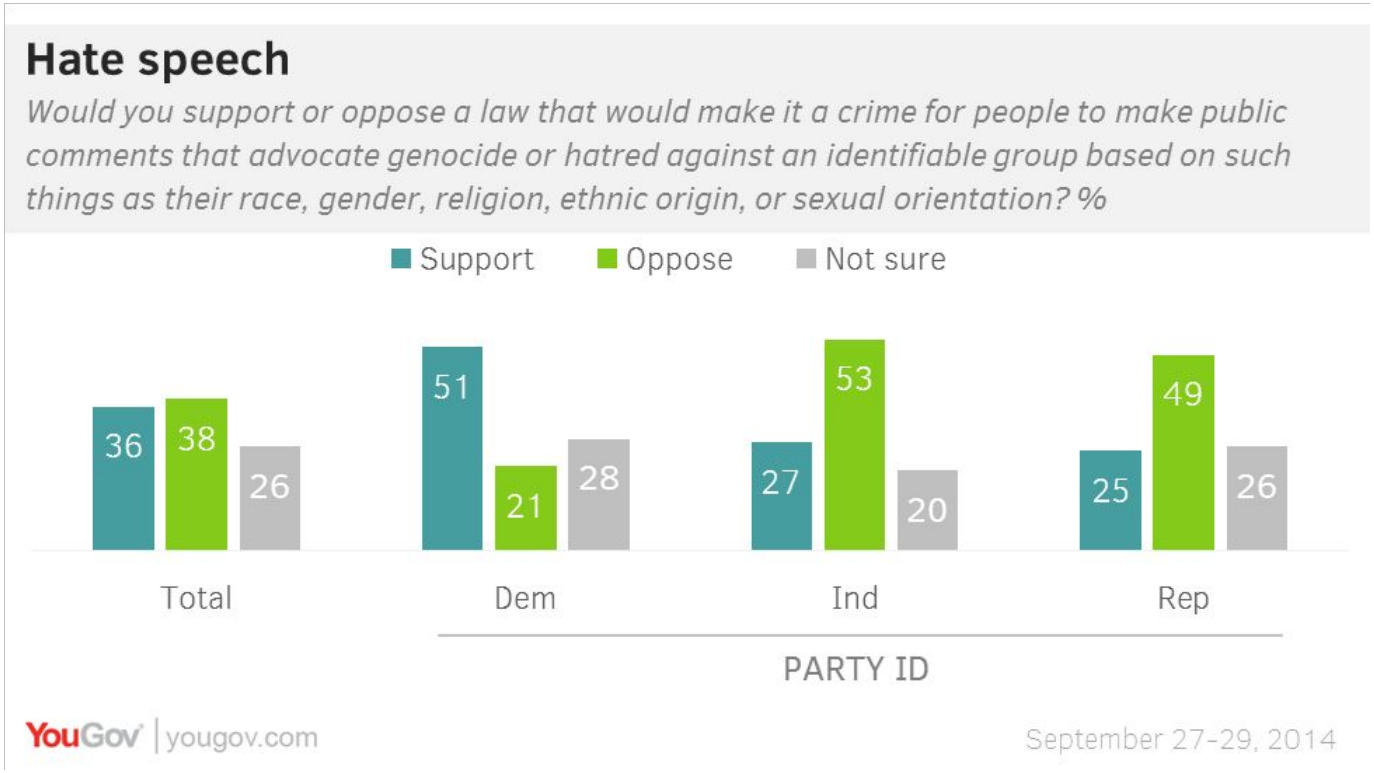


Image url:
https://today.yougov.com/topics/politics/articles-reports/2014/10/02/america-divided-hate-speech-laws

Why should we be concern?

- Companies are also working hard to combat the online hate speech - very taxing on the content moderators
- Lack automated methods!

Silicon UK

Facebook, Twitter, YouTube Agree To Hate Speech Audits

... platforms including Facebook, YouTube and Twitter have agreed with large advertisers to outside audits of their efforts to tackle hate speech.

Sep 23, 2020

Digital Information World

Facebook Published A Report On Its Evolving Efforts To Tackle Hate Speech; The Company Has Included The…

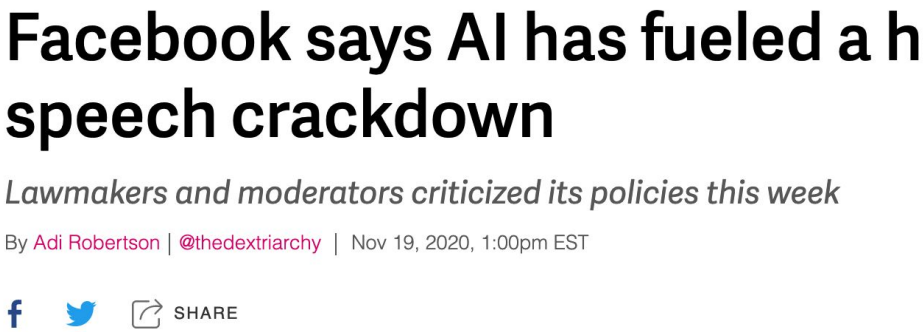
Facebook Published A Report On Its Evolving Efforts To Tackle Hate Speech; The Company Has Included The Prevalence of Hate Speech As ...

1 week ago

YouTube: Inside the traumatic life of a Facebook moderator

Automated Hate Speech Detection

- **Task**: *Given a text, we aim to predict if it contain hateful content automatically.*
- Challenging research problem:
  - Data scarcity - imbalance data
  - Subjectivity in data annotation
  - Rule-based methods - Not flexible, evolving hate speech
  - Supervised methods - Heavily depended on labelled data
  - …



Facebook says AI has fueled a hate speech crackdown

Lawmakers and moderators criticized its policies this week

By Adi Robertson | @thedextriarchy | Nov 19, 2020, 1:00pm EST

SHARE

News: https://www.theverge.com/2020/11/19/21575139/facebook-moderation-ai-hate-speech

Data Annotation and Collection

Hate Speech Datasets

- Open datasets included in Model Zoo

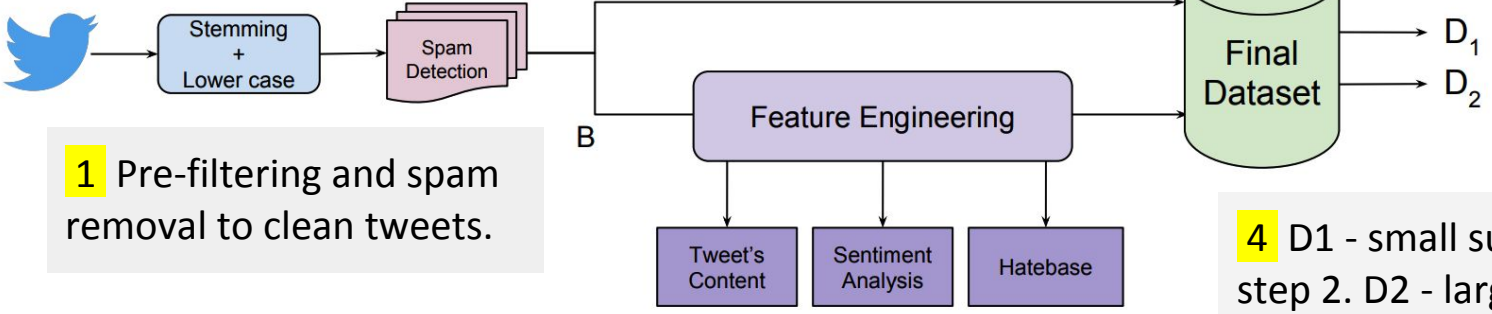
Dataset	#Tweets	Classes (#Tweets)
WZ-LS [Park et al, Waseem 1,2]	13,202	racism (82), sexism (3,332), both (21), neither (9,767)
DT [Davidson et al]	24,783	hate (1,430), offensive (19,190), neither (4,163)
FOUNTA [Founta et al]	89,990	normal (53,011), abusive (19,232), spam (13,840), hate (3,907)

- Other interesting dataset:
  - Facebook hateful memes dataset (~10K memes)
  - COVID-19 Anti-social dataset (~37M tweets)

Hate Speech Annotation

- *Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior*, Founta et al, ICWSM'18
- A crowdsourcing framework to annotate hate speech dataset

Step 1: Data Collection



2 A - Random set of unboosted tweets.

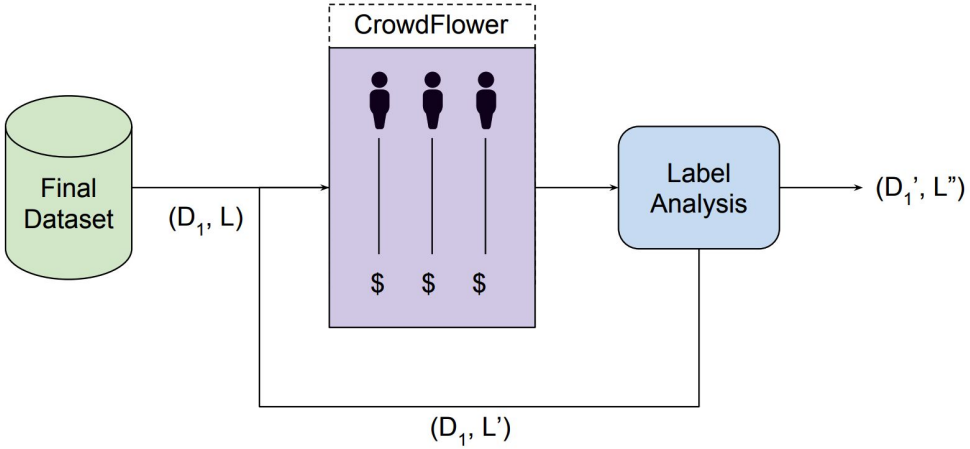
1 Pre-filtering and spam removal to clean tweets.

3 B - boosted sampling to produce a set of tweets biased towards abusive behavior

4 D1 - small subset for step 2. D2 - large subset for step 3.

Hate Speech Annotation

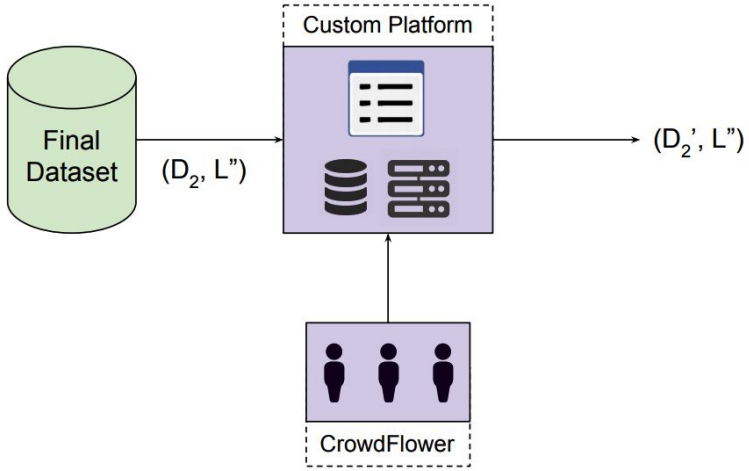
Step 2: Exploratory Analysis



- **Objective:** Use a small D1 subset (300 tweets) to iteratively find the optimal set of labels for full annotations.
- **Round 1**: Start with the most extensively used labels found in literature, crowdsource annotators to annotate the tweets. Each tweets is annotated by 5 annotators.
- **Round 2**: Merge and remove labels that are frequently confused by users, and focus on annotating tweets that are marked inappropriate in Round 1.
- **Round 3**: Validate the selected labels and confirm annotation agreement.

Hate Speech Annotation

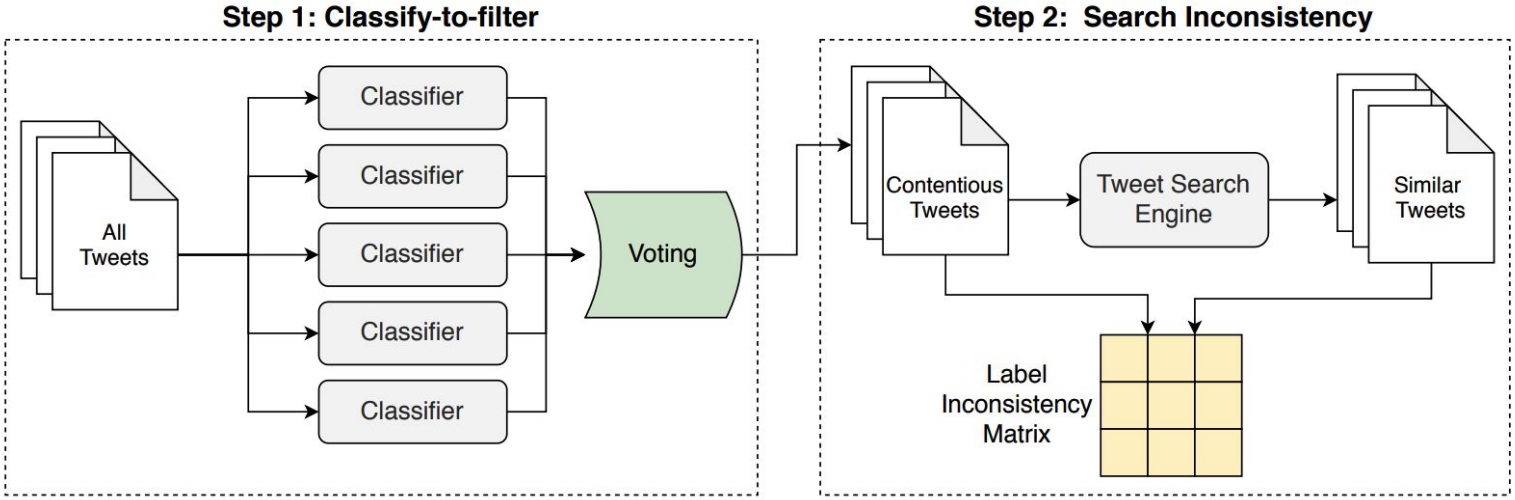
Step 3: Full Annotation



- **Objective:** Using the labels and parameters learn from Step 2, crowdsource and label the large D2 subset.

Inconsistency in Annotation

- Annotated by multiple annotators and inter-annotator agreement are computed, BUT...
  - Hate speech are still subjective and prone to bias
- Inconsistencies still found in the dataset!
  - *On Analyzing Annotation Consistency in Online Abusive Behavior Datasets*, Awal et al, ICWSM Data Challenge'20



Inconsistency in Annotation

Label Inconsistency Matrix for FOUNTA

		Contentious Tweet Label			
		Abusive	Hate	Spam	Normal
Similar Tweet Label	Abusive	491	1547	736	1062
	Hate	347	370	93	192
	Spam	109	62	790	1024
	Normal	758	1133	3170	915

Inconsistency annotation examples in FOUNTA

Id	Tweet	Label
C1	RT:[USER_1] How about we f**king hire trans boys to play trans boys	hate
S1	RT:[USER_1] How about we f**king hire trans boys to play trans boys	normal
C2	RT:[USER_2] I wish I wasn't so annoying like I even p*ss myself off	normal
S2	RT:[USER_2] I wish I wasn't so annoying like I even p*ss myself off	abusive
C3	RT:[USER_3] f**king faggot	hate
S3	[USER_4] f**king faggot	abusive

- Significant annotation inconsistency are found in FOUNTA
- The most similar tweets to the contentious tweets looks the same! - retweets
- The retweet may be annotated by different set of annotators, resulting in diffferent annotation
  - Subjectivity!

Dataset summary

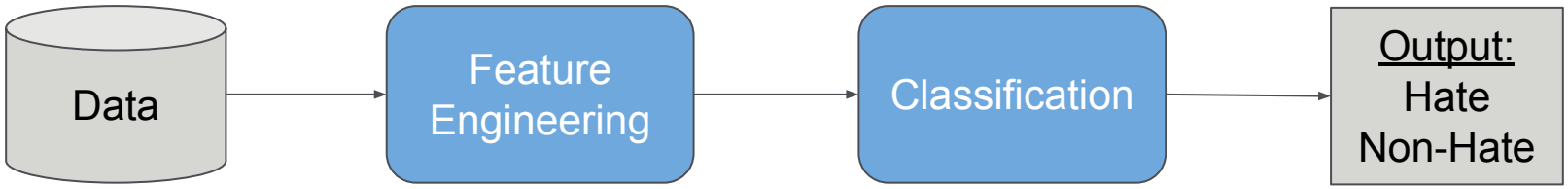
- Hate speech datasets are available , BUT…
  - Highly imbalanced (<15% of tweets are hateful)
  - Inconsistency in label - Annotator subjectivity
  - Mostly text-based content

- Downstream impact:
  - Introduce bias in supervised automated detection methods
  - Inaccurate classification

Traditional Machine Learning Detection Techniques

Traditional Machine Learning

- Build Machine Learning models for automated hate speech classification - text classification
  - Binary Classification (Hate vs Non-Hate)
  - Multi-label Classification (Sexism, Racism, Both, None, etc.)
- Two main components in Hate speech classification ML models:
  - Feature Engineering
  - Applying Classifiers



Feature Engineering

- Leveraging External Dictionaries/Lexicon
  - https://hatebase.org/ - multilingual hate speech lexicon
  - Apply to filter "*likely hateful*" contents
  - Used with rule-based detection models
  - Advantage: Simple and good starting point
  - Disadvantages:
    - Highly context dependent
    - Often up-to-date

Hatebase Example

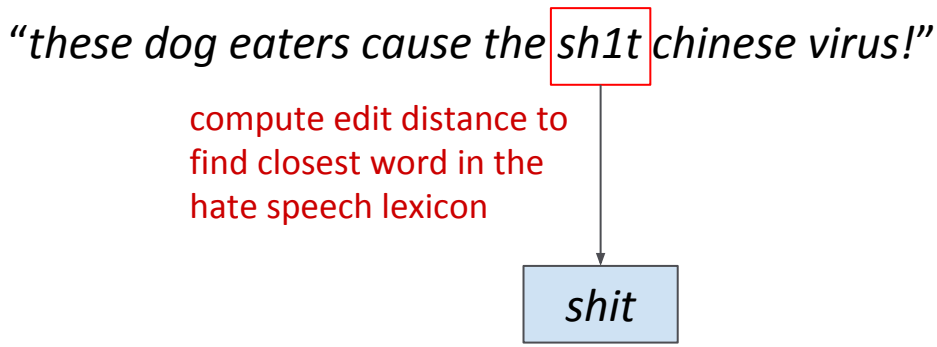
"Yankees" is considered highly offensive

New York Yankees

Yankee Candle

Feature Engineering

- Linguistics features derieve from text
  - Edit Distance Metrics
    - Catch un-normalized terms (e.g., "*@ss*", "*sh1t*", etc.)
    - Helpful for social media dataset - informal text
  - Bag-of-Words (BOW)
  - NGram
  - TF-IDF
  - Part-of-Speech (POS)



Feature Engineering

- Lingustics features derieve from text
  - Edit Distance Metrics
  - Bag-of-Words (BOW)
    - Frequency of the words
    - Ignore word sequences
  - NGram
  - TF-IDF
  - Part-of-Speech (POS)

"these dog eaters cause the sh1t chinese virus!"

these	dog	eaters	cause	the	sh1t	chinese	virus
1	1	1	1	1	1	1	1

Feature Engineering

- Lingustics features derieve from text
  - Edit Distance Metrics
  - Bag-of-Words (BOW)
  - NGram
    - Word-level or character-level
    - Useful and predictive feature
  - TF-IDF
  - Part-of-Speech (POS)

"these dog eaters cause the sh1t chinese virus!"

n=2 words

these dog	dog eaters	eaters cause	cause the	the shit	sh1t chinese	chinese virus
1	1	1	1	1	1	1

Feature Engineering

- Lingustics features derieve from text
  - Edit Distance Metrics
  - Bag-of-Words (BOW)
  - NGram
  - TF-IDF
    - Term frequency off-set by word frequency at corpus-level
    - Apply to word-level or ngram
  - Part-of-Speech (POS)

"these dog eaters cause the sh1t chinese virus!"

these	dog	eaters	cause	the	sh1t	chinese	virus
0.0001	0.09	0.1	0.003	0.0001	0.12	0.08	0.02

Feature Engineering

- Lingustics features derieve from text
  - Edit Distance Metrics
  - Bag-of-Words (BOW)
  - NGram
  - TF-IDF
  - Part-of-Speech (POS)
    - Type of words used in the text

"these dog eaters cause the sh1t chinese virus!"

these	dog	eaters	cause	the	sh1t	chinese	virus
DET	NOUN	NOUN	VERB	DET	NOUN	ADJ	NOUN

Feature Engineering

- Sentiment-based
  - Hate speech has negative polarity
  - Fine-grain affects: anger, disgust, fear, etc. [Mohammad et al]

- Topic Modeling
  - Contextual information (e.g., COVID-19, Border Wall, etc.)
  - Use with lexicon approach

Classifiers

- Logistic Regression
- SVM
- Decision Tree
- Random Forest
- Naive Bayes
- Ensemble Models

Sample Classification Results

- *One-step and Two-step Classification for Abusive Language Detection on Twitter*, Park et al, WOAH'17
  - Feature used: Character ngram

	None			Racism			Sexism			Total		
Method	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
LR	.824	.945	.881	.810	.598	.687	.835	.556	.668	.825	.824	.814
SVM	.802	**.956**	.872	**.815**	.531	.643	**.851**	.483	.616	.814	.808	.793
FastText	.828	.922	**.882**	.759	.630	.685	.777	.557	.648	.810	.812	.804
CharCNN	.861	.867	.864	.693	.746	.718	.713	.666	.688	.801	.811	.811
WordCNN	.870	.868	.868	.704	.762	.731	.712	**.686**	.694	.818	.816	.816
HybridCNN	**.872**	.882	.877	.713	**.766**	**.736**	.743	.679	**.709**	**.827**	**.827**	**.827**

- Traditional machine learning models can perform on-par or better than deep learning models!

Traditional ML Summary

- Dictionary/Lexicon with rule-based methods are commonly used in hate speech classification
- Most effort are in data-preprocessing and feature engineering
- Traditional ML models are able to perform on-par or better than deep learning models

Deep Learning Detection Techniques

Deep Learning

- Two Steps in Hate Speech Detection:
  - Understand the sentence
  - Classify the sentence
- How to Apply Deep Learning to Hate Speech Detection
  - Extract sequential information
    - LSTM (Long-Short Term Memory Networks), CNN (Convolutional Neural Networks), etc, are powerful to learn sentence representation
  - Hate speech detection: a classification task
    - Binary classification: between hate & non-hate
    - Multi-class classification: between hate and other classes
    - Feed sentence representation to a classification layer

Baseline of Deep Learning

- The Most Straightforward Way
  - Two components in the model
    - Sentence representation learner
    - Classfication layer
  - LSTM or CNN to learn sentence-level representation
  - Fully-connected (FC) layer for classification
  - Two baselines:
    - LSTM + FC
    - CNN + FC

Preprocessing Steps

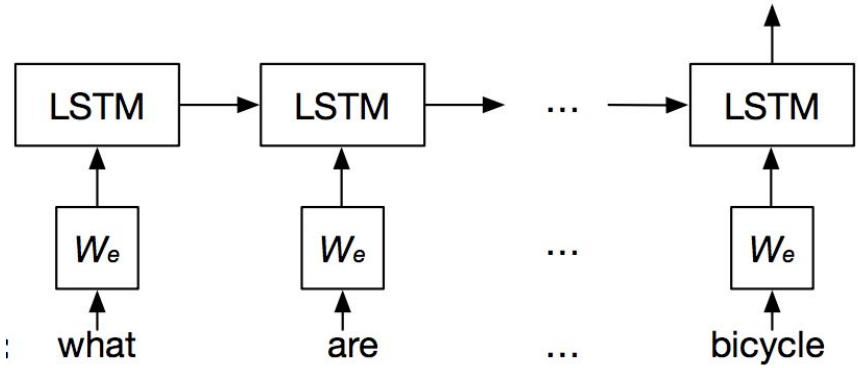
- Why Preprocessing
  - Data from online platforms involve some unexpected symbols or tokens
  - Makes it easier for models to understand texts
- Preprocessing of texts:
  - Remove punctuations, emoticons, stopwords, etc.
  - Apply lowercase or stemming
  - Remove some infrequent words
  - Nomalize of hashtags into words

Results from Preprocessing

Before Preprocessing	After Preprocessing
RT @Fewjr: @1MarKus_A @MakEitSndGoOd she curved regular ni*gas daily. Especially us Austin ni*gas. D*mn yellow ppl	RT USER she curved regular ni*gas daily Especially us Austin ni*gas D*mn yellow people
RT @MadPatsFan1954: .@FR_INC "c*on meat?!" Made me ill to even type that. Racist much? @lybr3	RT USER USER c*on meat Made me ill to even type that Racist much USER
RT @ESPNSecondTake_: If this ugly to you, you either a g*y ni*ga or a hating a*s bi*ch http://t.co/H0wAPpsmgg	RT USER If this ugly to you you either a g*y ni*ga or a hating a*s bi*ch
RT @EnglandBailey: Happy birthday to the nicest fa*got ever 💁 http://t.co/vZygNI9qtQ	RT USER Happy birthday to the nicest fa*got ever

LSTM + FC

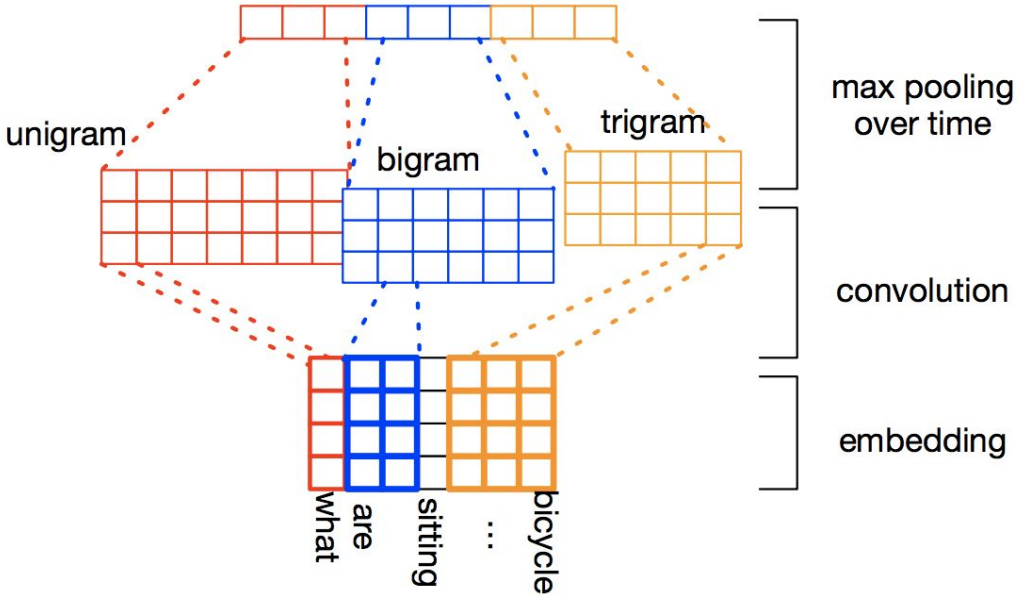
- Word Embeddings
  - Transform each word into a vector
  - Use either pretrained embeddings (GloVe, Fast-Text, etc) or without pretraining
- LSTM for Sequential Information Extraction
  - Use last hidden state as sentence representation



- Classification
  - Feed the sentence representation to FC to get confidence scores for each class

CNN + FC

- Word Embeddings & Classification
  - Similar to LSTM + FC
- CNN for Sequential Information Extraction
  - Convolution using different kernel sizes
  - Max pooling over each output from convolution
  - Concatenate outputs from each max pooling layer

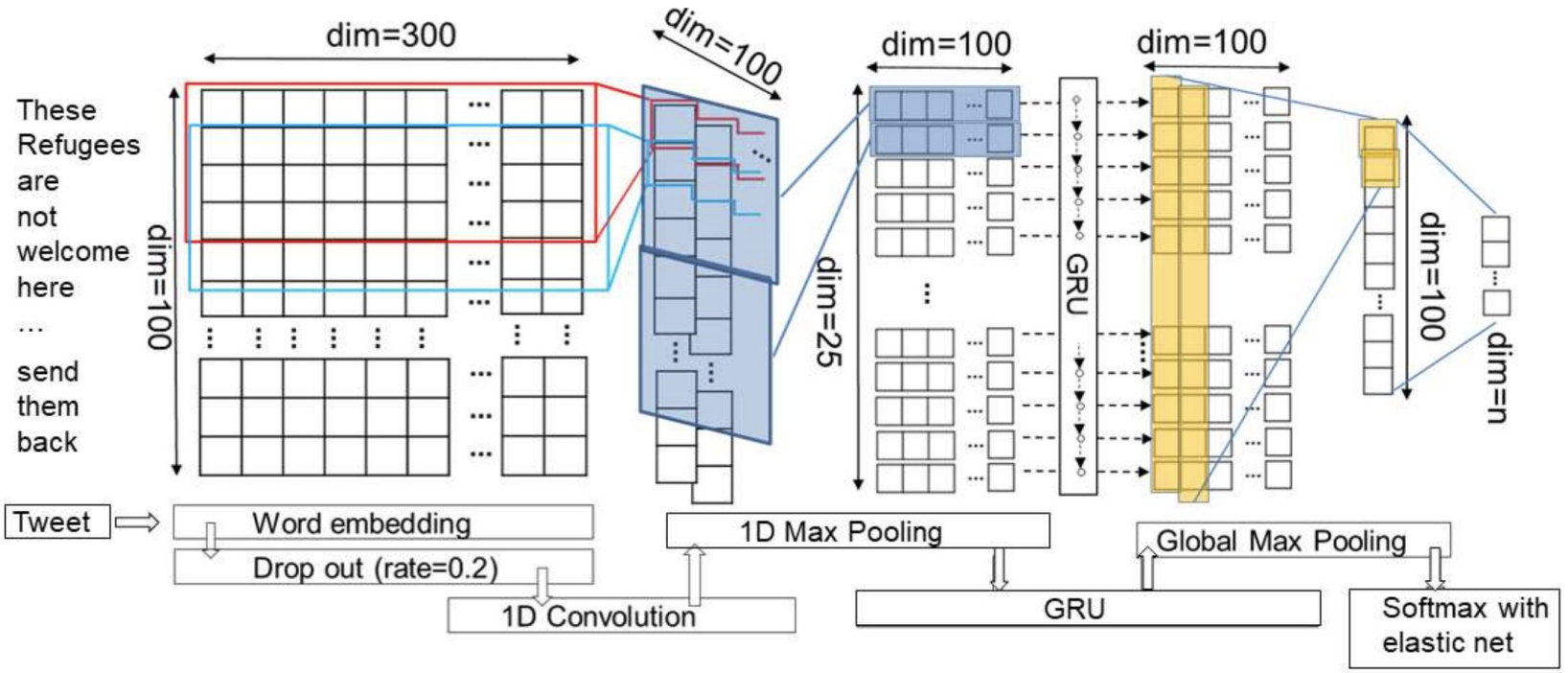


Comparison between Baselines

- Main difference
  - The way to extract sentence representation
- Question
  - Are there other ways to learn sentence-level representation?
  - Yes!
    - CNN-GRU: combine both CNN and GRU (another form of RNN)
    - Bert: a pretrained model powerful for several natural language processing tasks

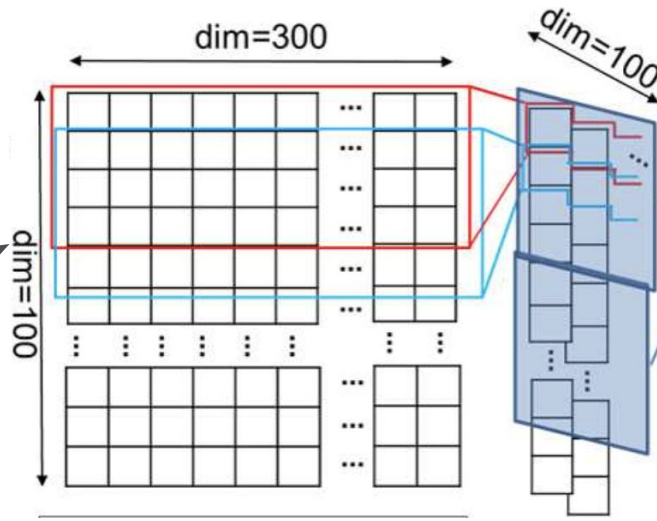
CNN-GRU

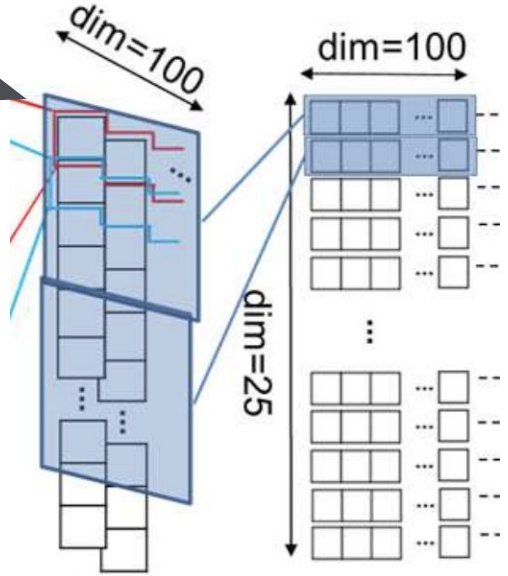
- Detecting hate speech on twitter using a convolution-gru based deep neural network, Zhang et al, ESWC'18
- Motivation: Take merits from both CNN and GRU
- Encode Words into Embeddings: Similar to LSTM-FC
- Main Difference: how to encode the sentence

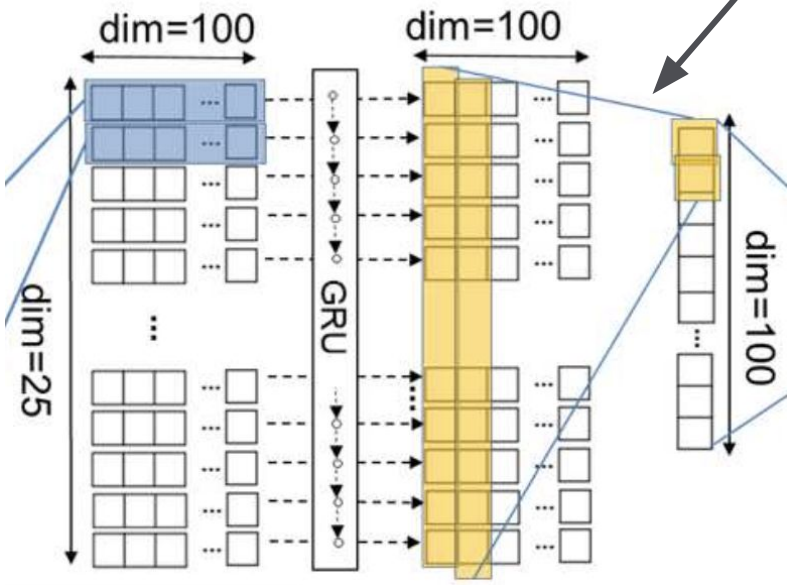


CNN-GRU

- CNN-GRU Sentence Encoder:
  - Convolution over sentence
    - Kernel size is 4, 100 kernels in total
  - Max Pooling with Pooling Size 4
  - Feed Output from Max Pooling to GRU
    - Max pooling over hidden states







Experimental Results

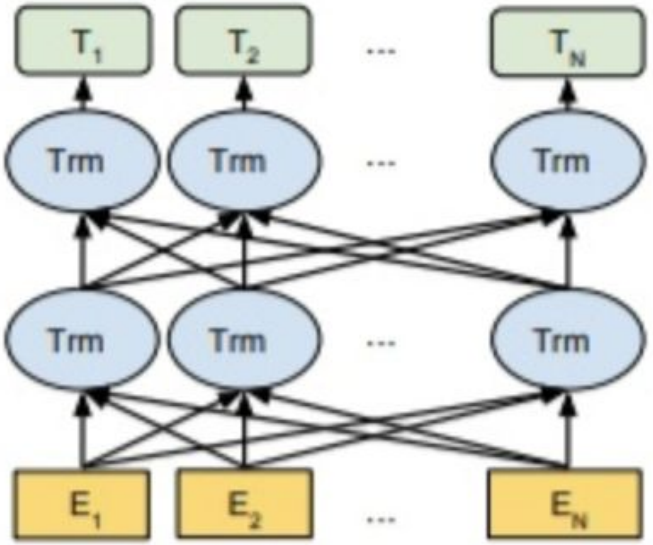
Dataset	SVM	SVM+	CNN	CNN+ GRU_B	CNN+ GRU	State of the art
WZ-L	0.74	0.74	0.80	0.81	**0.82**	0.74 Waseem [26], best F1
WZ-S.amt	0.86	0.87	0.91	**0.92**	**0.92**	0.84 Waseem [25], Best features
WZ-S.exp	0.89	0.90	0.90	0.91	**0.92**	0.91 Waseem [25], Best features
WZ-S.gb	0.86	0.87	0.91	0.92	**0.93**	0.90 Gamback [10], best F1
WZ-LS	0.72	0.73	0.81	0.81	0.82	0.82 Park [20], WordCNN 0.81 Park [20], CharacterCNN **0.83** Park [20], HybridCNN
DT	0.87	0.89	**0.94**	**0.94**	**0.94**	0.87 SVM, Davidson [7]
RM	0.86	0.89	0.90	0.91	**0.92**	0.86 SVM, Davidson [7]

Bert

- What is Bert?
  - Bidirectional Encoder Representations from Transformers
  - A Pretrained Deep Neural Network Model
    - Pretrained on *Masked LM* and *Next Sentence Prediction*
    - Powerful on several NLP tasks
    - Can be easily fine-tuned for other supervised downstream tasks

Representation for each word considering its contexts

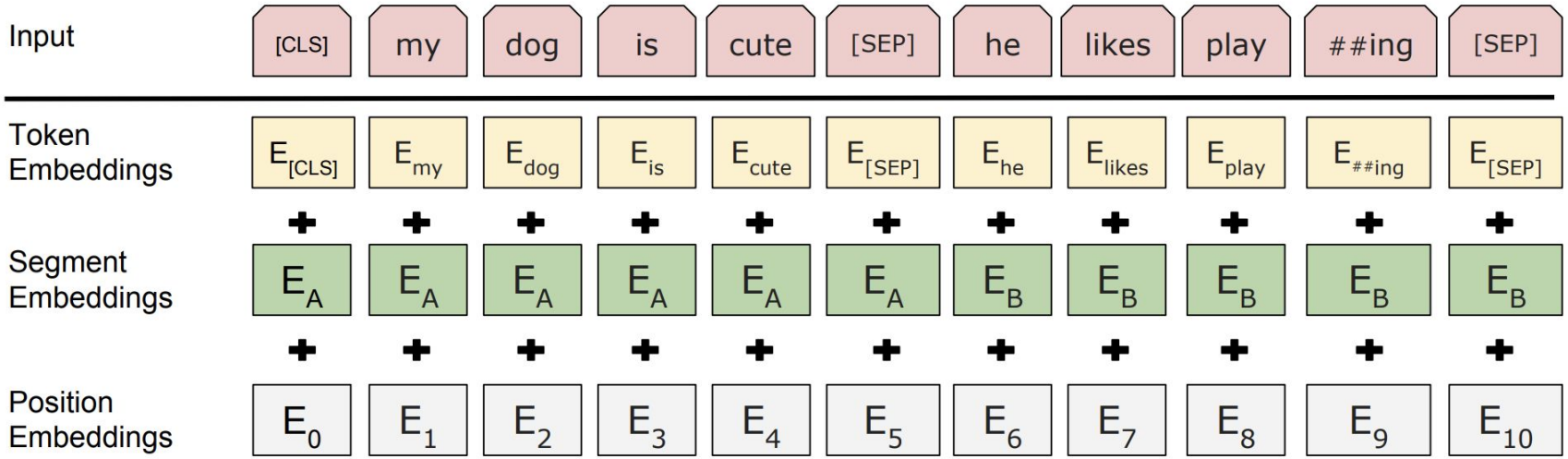
Stacks of Transformer layers



Inputs to Bert model

Bert for Hate Speech Detection

- Details of Using Bert to Hate Speech Detection
  - Preprocessing texts and tokenize sentence using default tokenizer from Bert
  - Send tokenized words into Bert model
    - Initialized with pre-trained weights
  - Feed output from [CLS] token to FC for classification

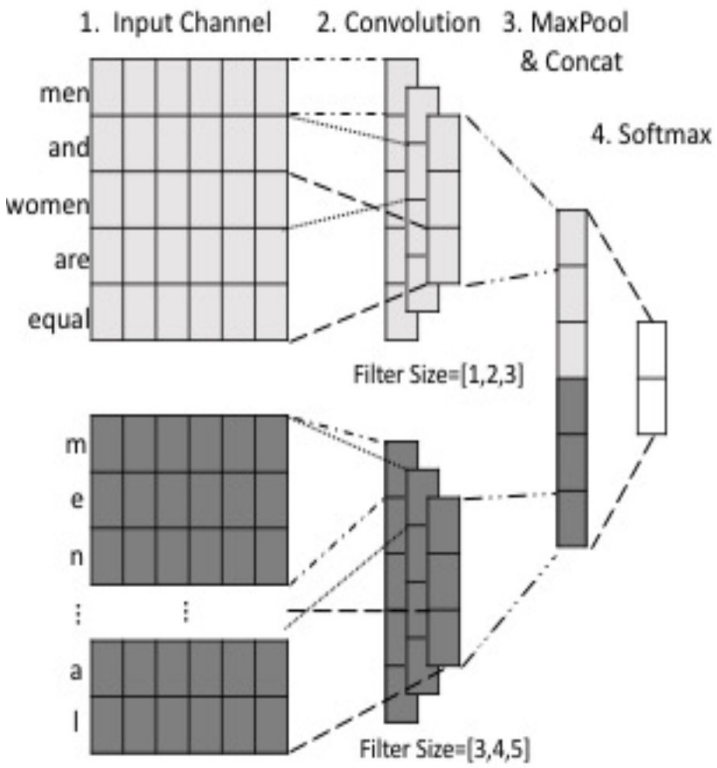


Hate Speech is beyond Words

- Not Only Word-Level Semantic Information in Texts
  - Character-level information
  - Sentiment information
  - Topic information
- Models Considering Other Types of Information
  - HybridCNN
    - Utilize both word and character level features
  - DeepHate
    - Utilize semantic, sentiment and topic representations

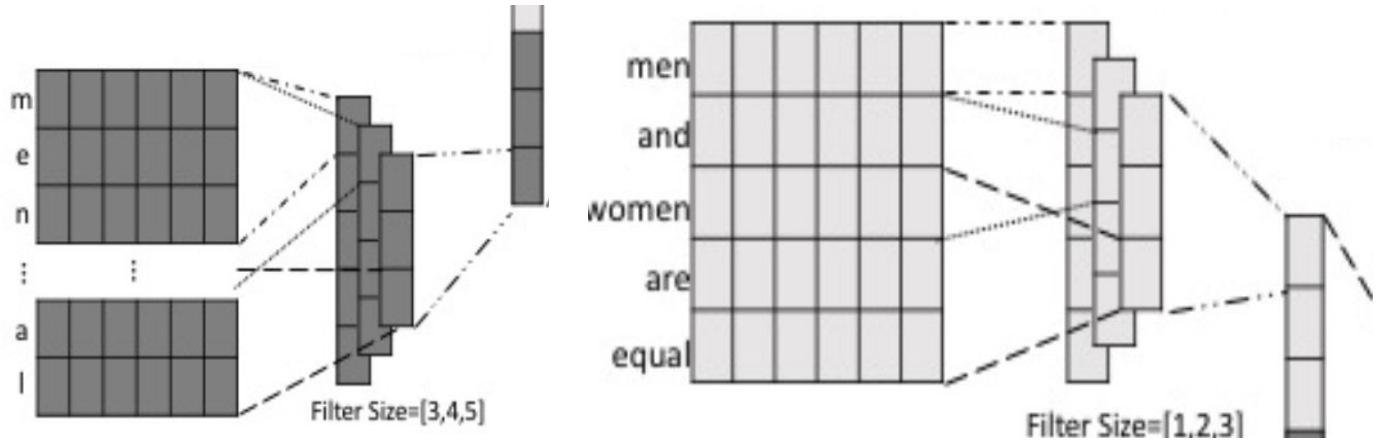
HybridCNN

- *One-step and Two-step Classification for Abusive Language Detection on Twitter*, Park & Fung, WOAH'17
- Motivation:
  - Abusive language often contains either purposely or mistakenly mis-spelled words
  - Capture features from both word and character level inputs



HybridCNN

- Two Main Components:
  - CharCNN:
    - Encode characters into embedding
    - Apply convolution and max pooling over character embeddings
  - WordCNN
    - Encode words into embedding
    - Apply convolution and max pooling over word embeddings
  - Concatenation of two levels of representation



Experimental Results

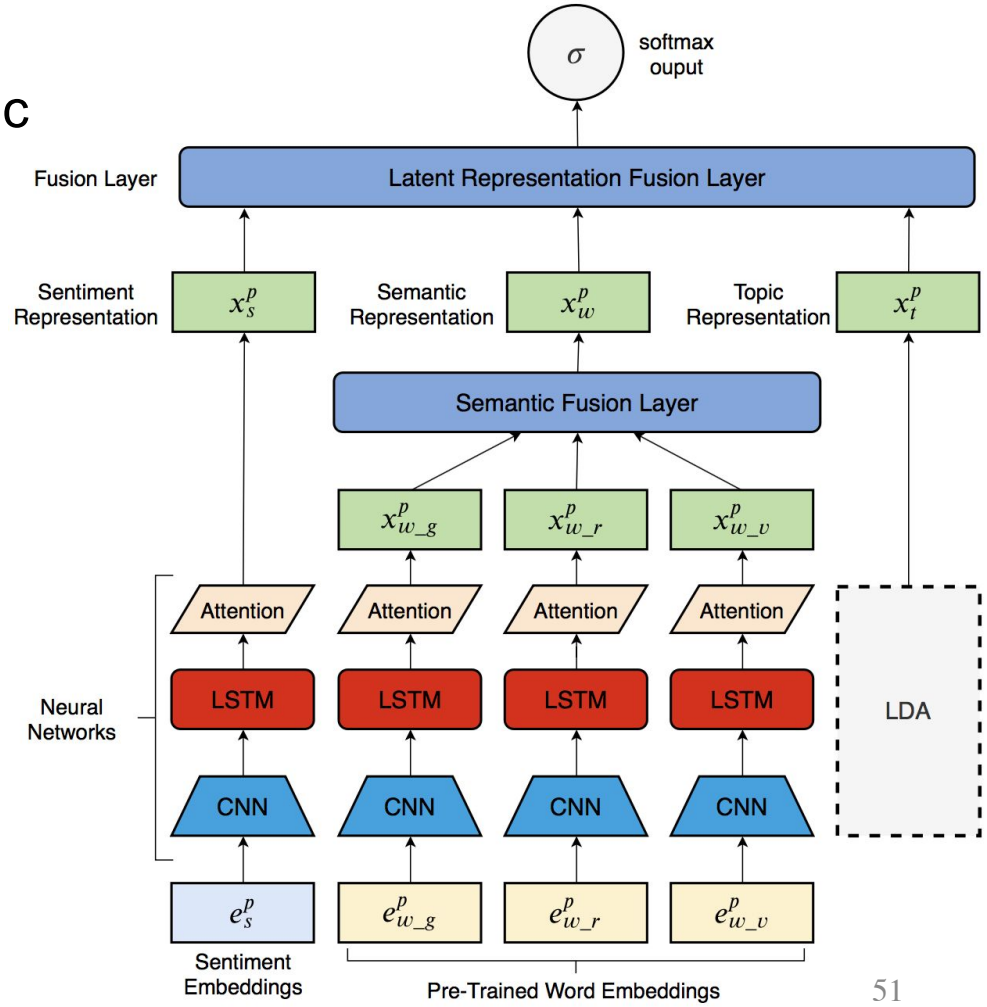
	None			Racism			Sexism			Total		
Method	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
LR	.824	.945	.881	.810	.598	.687	.835	.556	.668	.825	.824	.814
SVM	.802	**.956**	.872	**.815**	.531	.643	**.851**	.483	.616	.814	.808	.793
FastText	.828	.922	**.882**	.759	.630	.685	.777	.557	.648	.810	.812	.804
CharCNN	.861	.867	.864	.693	.746	.718	.713	.666	.688	.801	.811	.811
WordCNN	.870	.868	.868	.704	.762	.731	.712	**.686**	.694	.818	.816	.816
HybridCNN	**.872**	.882	.877	.713	**.766**	**.736**	.743	.679	**.709**	**.827**	**.827**	**.827**
LR (two)	.841	.933	**.895**	.800	.664	.731	.809	.590	.683	**.828**	**.831**	**.824**
SVM (two)	.816	**.945**	.876	**.811**	.605	.689	**.823**	.511	.630	.816	.815	.803
HybridCNN (two)	.877	.864	.869	.690	**.759**	.721	.705	.701	**.699**	.807	.809	.807
HybridCNN + LR(two)	**.880**	.859	.869	.722	.751	**.735**	.683	**.717**	**.699**	.821	.817	.818

DeepHate

- *DeepHate: Hate Speech Detection via Multi-Faceted Text Representations*, Rui et al, WEBSCI'20
- Use Multi-faceted Textual Representations
  - Semantic representation
    - Extract contents of a post: semantics; Obtain expressive word-level representations
  - Sentiment representation
    - Incorporate attitude and emotion of the writer for hate speech detection
  - Topic representation
    - Incorporates information of posts about a certain topic or issue

DeepHate

- Main Components
  - CNN-LSTM-Att encoder
  - Semantic, sentiment, topic representation leaner
  - Representation Fusion layer



Information Redundancy

- Not All Words are Equally Important
  - Some words are essential for hate speech detection while others are not
- Solution
  - Attention mechanism
  - Force model to put more attention to words important for making predictions
  - Self-Attention over words
    - CNN-LSTM-Att encoder in DeepHate
    - Illustration of visualization using attention

Label	Example Post
Offensive	Bitch ill fuck you up
Hate	Those guys are the definition of white trash

CNN-LSTM-Att in DeepHate

- Function:
  - Obtain sentence representation based on different embeddings
  - Attend to task related words
- Details:
  - CNN-LSTM: similar to CNN-GRU
    - Difference: no max pooling over outputs from LSTM
  - Apply self-attention over all hidden states from LSTM
    - Relevance of each state to sentence-level information
    - Weighted average over all states

$$M = tanh(W_H H + W_h h_L + b_h)$$

$$\alpha = softmax(w^T M)$$

$$x^p = H\alpha^T$$

DeepHate

- Semantic Representation
  - Encode words using three kinds of pretrained embeddings
  - Feed each word embeddings into a CNN-LSTM-Att encoder
- Sentiment Representation
  - Train sentiment embeddings under task of sentiment analysis
  - Encode words using sentiment embeddings and feed embeddings into a CNN-LSTM-Att encoder
- Topic Representation
  - Generate topic representation using LDA
- Representation Fusion
  - Use gate attention to fuse different information

Experimental Results

Table 2: Experiment results of DeepHate and baselines on WZ-LS dataset

Model	Micro-Prec	Micro-Rec	Micro-F1
CNN-W	75.95	78.57	75.54
CNN-C	54.77	74.01	62.95
CNN-B	76.30	79.08	74.78
LSTM-W	75.39	**79.52**	74.52
LSTM-C	74.82	78.13	71.95
LSTM-B	54.77	74.01	62.95
HybridCNN	76.35	78.93	75.98
CNN-GRU	75.33	79.27	74.42
DeepHate	**77.95**	79.48	**78.19**

Table 3: Experiment results of DeepHate and baselines on DT dataset

Model	Micro-Prec	Micro-Rec	Micro-F1
CNN-W	87.88	88.65	87.95
CNN-C	60.53	77.43	67.60
CNN-B	78.02	80.33	77.01
LSTM-W	88.08	89.08	87.87
LSTM-C	77.21	79.88	76.47
LSTM-B	59.97	77.44	67.60
HybridCNN	88.33	88.96	88.07
CNN-GRU	87.60	88.24	87.23
DeepHate	**89.97**	**90.39**	**89.92**

Table 4: Experiment results of DeepHate and baselines on FOUNTA dataset.

Model	Micro-Prec	Micro-Rec	Micro-F1
CNN-W	78.26	79.71	78.27
CNN-C	69.66	70.15	64.40
CNN-B	52.01	58.41	50.64
LSTM-W	78.54	79.87	78.48
LSTM-C	70.15	70.89	66.30
LSTM-B	55.22	62.71	54.52
HybridCNN	78.34	79.24	77.73
CNN-GRU	78.62	80.17	78.39
DeepHate	**78.95**	**80.43**	**79.09**

Table 5: Experiment results of DeepHate and baselines on COMBINED dataset.

Model	Micro-Prec	Micro-Rec	Micro-F1
CNN-W	91.86	91.77	91.72
CNN-C	79.56	79.06	78.50
CNN-B	42.26	58.63	43.67
LSTM-W	91.54	91.54	91.52
LSTM-C	82.46	80.64	79.70
LSTM-B	64.93	65.50	63.85
HybridCNN	91.77	91.72	91.67
CNN-GRU	91.63	91.40	91.31
DeepHate	**92.48**	**92.45**	**92.43**

Impact of Different Information

Table 6: Performance of various DeepHate components on WZ-LS dataset

Model	Micro-Prec	Micro-Rec	Micro-F1
Semantic	77.00	78.75	77.04
Topic+Semantic	**77.98**	79.32	77.98
Sentiment+Semantic	77.09	78.62	77.35
DeepHate	77.95	**79.48**	**78.19**

Table 7: Performance of various DeepHate components on DT dataset

Model	Micro-Prec	Micro-Rec	Micro-F1
Semantic	89.44	90.24	89.49
Topic+Semantic	89.56	90.28	89.64
Sentiment+Semantic	89.59	**90.39**	89.64
DeepHate	**89.97**	**90.39**	**89.92**

Table 8: Performance of various DeepHate components on FOUNTA dataset

Model	Micro-Prec	Micro-Rec	Micro-F1
Semantic	78.68	80.40	78.57
Topic+Semantic	78.77	80.45	78.62
Sentiment+Semantic	78.88	**80.53**	78.79
DeepHate	**78.95**	80.43	**79.09**

Table 9: Performance of various DeepHate components on COMBINED dataset

Model	Micro-Prec	Micro-Rec	Micro-F1
Semantic	92.26	92.23	92.20
Topic+Semantic	92.33	92.30	92.27
Sentiment+Semantic	92.32	92.28	92.25
DeepHate	**92.48**	**92.45**	**92.43**

Demo

Deep Learning for Hate Speech Detection

Open Issues & Opportunities

Open Issues

- Datasets issues:
  - Inconsistency in data annotation
  - Imbalance dataset
  - Restricted to text-based

- Model issues
  - Mostly supervised - depend heavily on labels
  - Bias in the models

Inconsistency of Datasets

- Annotation of Hate Speech Datasets: Difficult
  - Semantic differences are blur (hate vs. offensive)
  - Annotation is subjective

- Analysis of Inconsistency of Datasets
  - *On Analyzing Annotation Consistency in Online Abusive Behavior Datasets*, Awal et al, ICWSM Data Challenge'20
  - *In Data We Trust- A Critical Analysis of Hate Speech Detection Datasets*, Madukwe et al, WOAH'20

Inconsistency of Datasets

- Results for Analysis of Inconsistency
  - Datasets: WZ, DT and FOUNTA
  - Retweets with different labels from FOUNTA

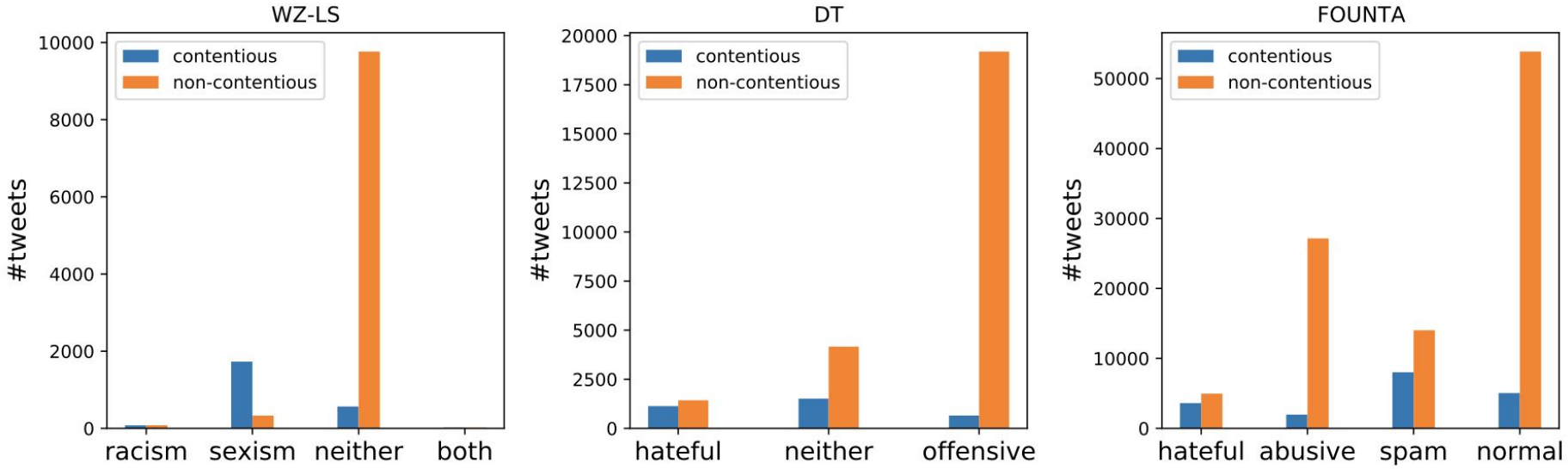


Figure 2: Breakdown distributions of contentious and non-contentious tweets from **WZ** (left), **DT** (middle), and **FOUNTA** (right) retrieved in step 1 of the annotation consistency analysis framework.

Imbalance of Datasets

- Hate Texts are Fewer Compared with Texts in Other Classes
- Solutions:
  - Add more data
    - Expensive and time consumimg
  - Automantic generate hate speech

Dataset	#Tweets	Classes (#Tweets)
WZ-LS [Park et al, Waseem 1,2]	13,202	racism (82), sexism (3,332), both (21), neither (9,767)
DT [Davidson et al]	24,783	hate (1,430), offensive (19,190), neither (4,163)
FOUNTA [Founta et al]	89,990	normal (53,011), abusive (19,232), spam (13,840), hate (3,907)

Automantic Generation

- Language Model
  - Few generated tweets are hate
  - Confusion to models
- Swap or Substitution of Words
  - Detrimental to fluency of sentences
- HateGAN
  - *HateGAN: Adversarial Generative-Based Data Augmentation for Hate Speech Detection*, Rui & Lee, COLING'20
  - A deep generative reinforcement learning model
  - Augment the datasets with generated hate tweets

Bias of Models

- Reason: datasets are imbalanced
- Models tend to give biased predictions
  - Predict hate when seeing offensive words
  - Predict hate when seeing group target words
- Example: bias of BERT model for hate speech detection
  - See words of group identifier and give hate predictions

Demo

Bias of Model: Bert

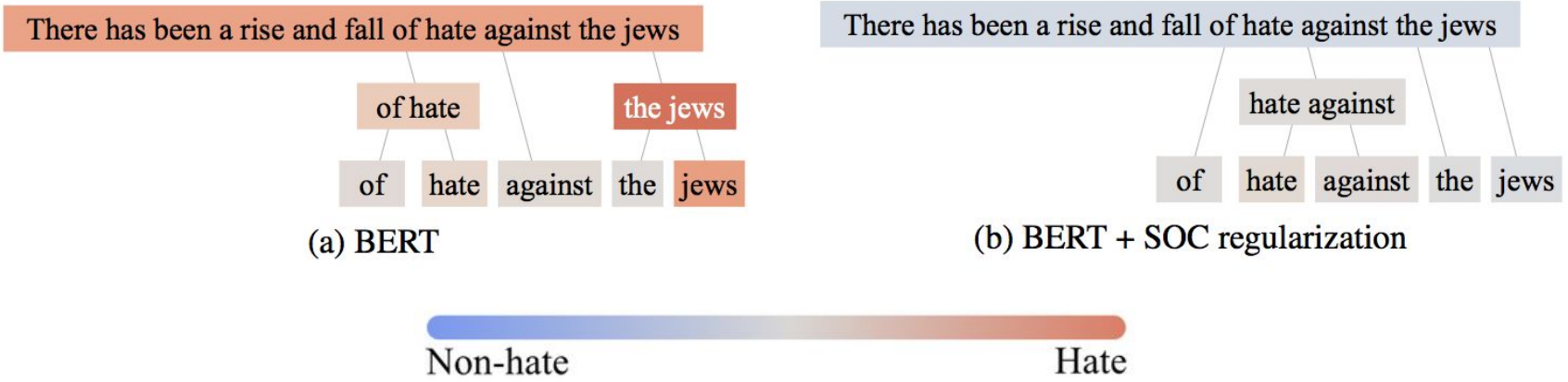
Debias of Models

- *Incorporating Priors with Feature Attribution on Text Classification*, Liu & Acvi, ACL'19
  - For toxic comment detection
  - Mitigating unintended bias in text classifiers by neutralizing identity terms
  - Use Integrated Gradients to calculate attribution scores of words
  - Punish over emphasised words

Method	Sentence	Probability
Baseline	I am *gay*	0.915
	I am straight	0.085
Our Method	*I am* gay	0.141
	I am straight	0.144

Debias of Models

- *Contextualizing Hate Speech Classifiers with Post-hoc Explanation*, Kennedy et al, ACL'20
  - Pre define a set of group target words
  - Use Sampling Occlusion (SOC) to score these words under hate speech detection models
  - Punish words with high scores



Results of Debias

- Top 20 Word by Mean SOC Before and After Debias
  - Bias to group identifiers: eased

BERT	Δ Rank	Reg.	Δ Rank
ni**er	+0	ni**er	+0
ni**ers	-7	fag	+35
kike	-90	traitor	+38
mosques	-260	faggot	+5
ni**a	-269	bastard	+814
jews	-773	blamed	+294
kikes	-190	alive	+1013
nihon	-515	prostitute	+56
faggot	+5	ni**ers	-7
nip	-314	undermine	+442
islam	-882	punished	+491
homosexuality	-1368	infection	+2556
nuke	-129	accusing	+2408
niro	-734	jaggot	+8
muhammad	-635	poisoned	+357
faggots	-128	shitskin	+62
nitrous	-597	ought	+229
mexican	-51	rotting	+358
negro	-346	stayed	+5606
muslim	-1855	destroys	+1448

Research Opportunities

- Multilingual and multi-culture dataset
  - Detecting hate speech in low resource languages
- Going beyond text
  - Detecting hateful multi-media content (memes, videos, etc.)
- Propagation of hate speech
  - Who are spreading hate speech and why?
- Improving annotation framework
  - Handle cultural context and bias in crowdsource annotation

Discussion & Conclusion

Conclusion

- Hate speech is a challenging research problem
- Existing automatic hate speech detection methods have good performance but limitations
- Lots of research opportunities in this domain!

Recruitment & Collaboration

- If you are interested in online misbehaviors research,
  - Chat with us for collaboration!
  - Join us as a Research Assistant or Research Fellow!

Roy Ka-Wei LEE | roy_lee@sutd.edu.sg

Rui CAO | ruicao.2020@phdcs.smu.edu.sg

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

SMU
SINGAPORE MANAGEMENT UNIVERSITY