Samsung System for Task 5 and Task 6 at SemEval 2019:

Linguistically enhanced deep learning offensive sentence classifier

Dr Alessandro Seganti, Dr Helena Sobol

May 13, 2019

IPI PAN

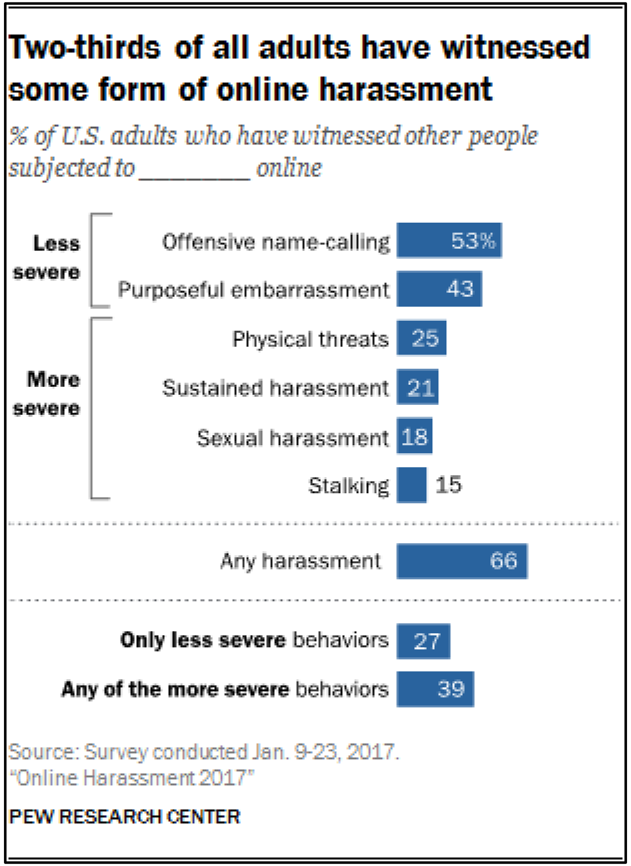
Contents

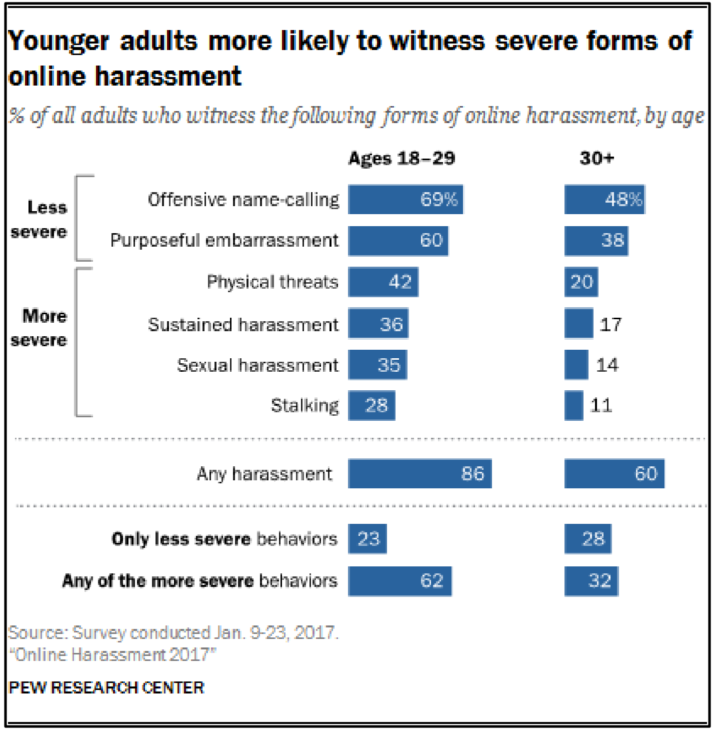
Contents

Introduction - Disclaimer

PARENTAL
ADVISORY
EXPLICIT CONTENT

Introduction – The problem





Pew Research Center – Online Harrassment 2017
http://www.pewinternet.org/2017/07/11/online-harassment-2017/

Introduction – SemEval 2019

Task 5 - hatEval

Multilingual detection of hate speech against immigrants and women in Twitter

- A: predict if a tweet is hateful against women or immigrants.
- B: (1) classify hateful tweets as aggressive or not aggressive; (2) classify the target as individual or group.

Task 6 - OffensEval

Identifying and Categorizing Offensive Language in Social Media

- A: Offensive language identification.
- B: Automatic categorization of offense types: Targeted Insults and Threats vs. Untargeted.
- C: Offense target identification: Individual, Group, Organization or Entity (e.g. a city, a location), Other (e.g. a situation, an event).

Contents

State of the art – Perspective API

- Perspective API made by Google is an API that developers can use to classify a comment's toxicity.

State of the art – Fooling Perspective API

Pros

- Bias mitigation through data preprocessing.
- Is able to recognize a large fraction of toxic comments.
- Around 95% accuracy.

Cons

- Model/Combination of models not available.
- Easily fooled with some tricks.

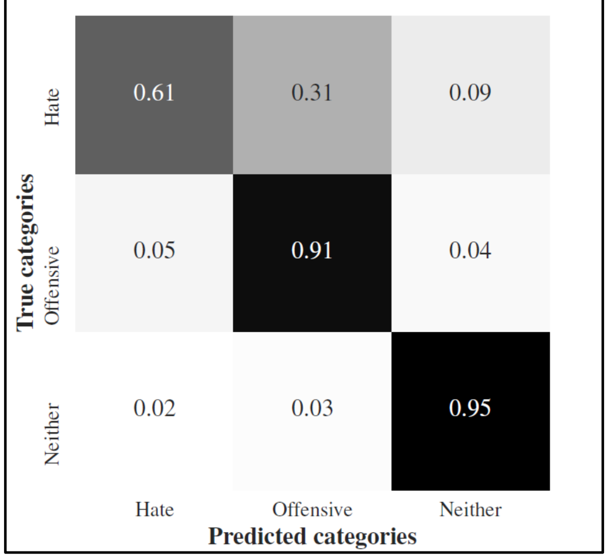
Original Phrase (Toxicity Score)	**Modified Phrase (Toxicity Score)**
Climate change is happening and it's not changing in our favor. If you think differently you're an **idiot**. (84%)	Climate change is happening and it's not changing in our favor. If you think differently you're an **idiiot**. (20%)
They're **stupid**, it's getting warmer, we should enjoy it while it lasts (86%)	They're **st.upid**, it's getting warmer, we should enjoy it while it lasts (2%)
They are liberal **idiots** who are **uneducated** (90%)	They are liberal **i.diots** who are **un.educated** (15%)
idiots. backward thinking people. **nationalists**. not accepting facts. susceptible to **lies**. (80%)	**idiiots**. backward thinking people. **nationaalists**. not accepting facts. susceptible to **l.ies**. (17%)
They are **stupid** and **ignorant** with no class (91%)	They are **st.upid** and **ig.norant** with no class (11%)
It's **stupid** and wrong (89%)	It's **stuipd** and wrong (17%)
If they voted for Hilary they are **idiots** (90%)	If they voted for Hilary they are **id.iots** (12%)
Anyone who voted for Trump is a **moron** (80%)	Anyone who voted for Trump is a **mo.ron** (13%)
Screw you trump supporters (79%)	**S c r e w** you trump supporters (17%)

Deceiving Google's Perspective API Built for Detecting Toxic Comments 2017 (https://arxiv.org/pdf/1702.08138.pdf)

State of the art – Hate sonar

- Lexicon from Hatebase.org
- Searched tweets using this lexicon.
- Selected ~25k tweets containing the lexicon and manually coded by crowd workers.
- Model:
  - 1) Logistic regression
  - 2) various models: logistic regression, naive Bayes, decision tree, SVM, …
  - → finally deciding for logistic regression with L2 regularization.

Figure 1: **True versus predicted categories**



Automated Hate Speech Detection and the Problem of Offensive Language, Davidson at al., 2017, https://arxiv.org/pdf/1703.04009.pdf

Contents

Bad word embeddings

fastText

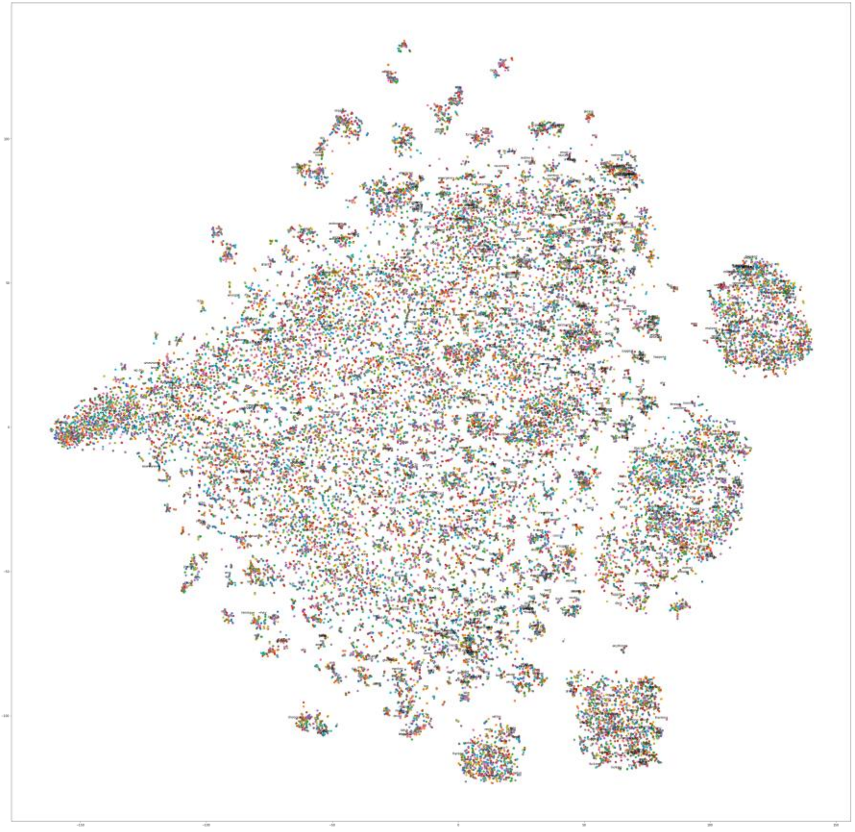
- Dataset: all sentences in the train and dev set.
- Preprocessing
  - Removing links, long words, nicknames, …
  - Tokenizing using nltk_tokenizer
- Learning a fastText embedding (using subwords).
- Embedding created: 300 dimensions, ~4k words.

ELMo

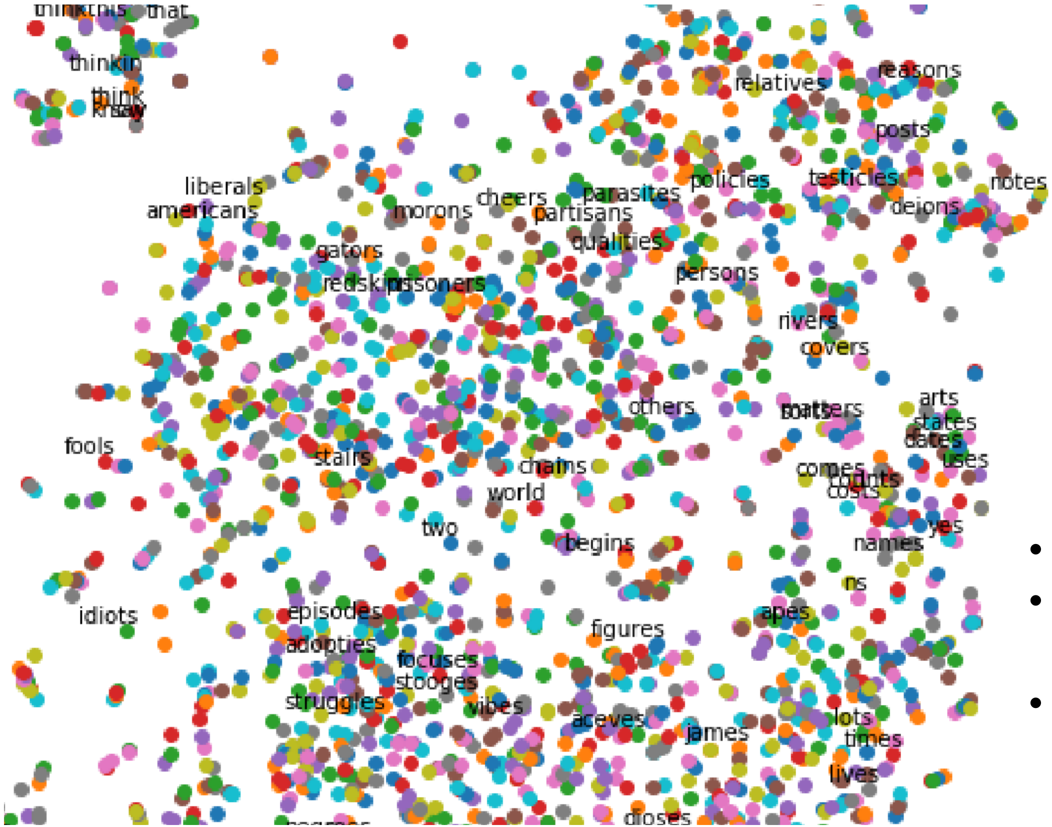
- Pretrained embeddings made from AllenNLP.
- Weighted sum of Bidirectional LSTM hidden state.
- Used through Tensorhub:
  - 3 LSTM layers.
  - Trained on 1 Billion Word benchmark.
  - 1024 dimensions.

(ELMO) Deep contextualized word representations Matthew E. Peters et al. NAACL 2018.

Bad word embeddings - fastText



Bad word embeddings - fastText



- Liberals, americans
- Parasites, partisans, policies
- Idiots, fools

Bias in Embeddings

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

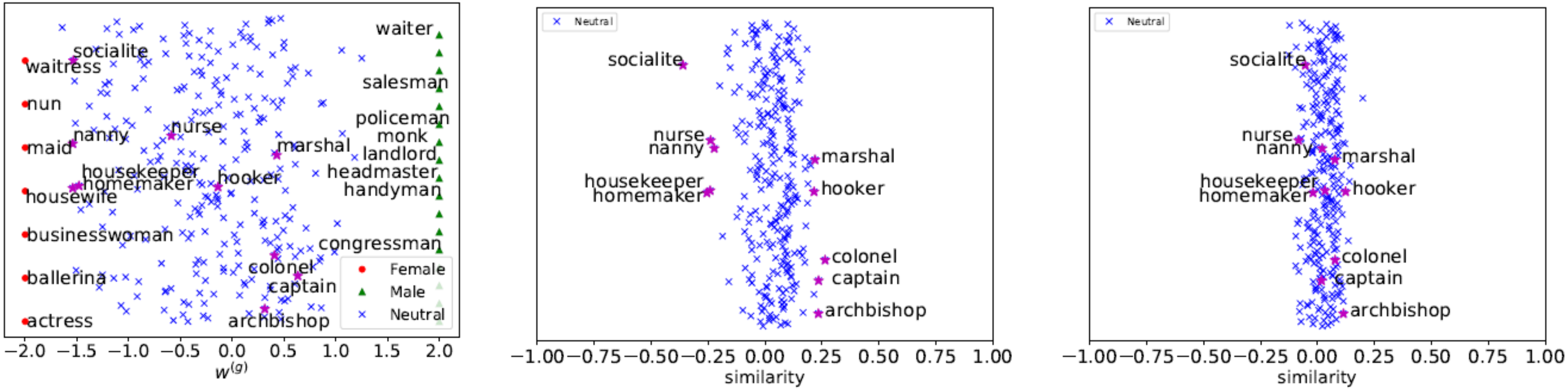
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay	3%	0.50%
transgender	0.04%	0.02%
lesbian	0.10%	0.04%
homosexual	0.80%	0.20%
feminist	0.05%	0.05%
black	0.70%	0.60%
white	0.90%	0.70%
heterosexual	0.02%	0.03%
islam	0.10%	0.08%
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

- Embeddings have biases due to the biases in the text used for learning.
- Toxic and non toxic comments can be distinguished by the model by recognizing determined words.

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings 2016 (https://arxiv.org/pdf/1607.06520.pdf)
Measuring and mitigating unintented bias in text classification 2017 – (link)

Bias in Embeddings



(a) $w^{(g)}$ dimension for all the professions

(b) Gender-neutral profession words projected to gender direction in GloVe

(c) Gender-neutral profession words projected to gender direction in GN-GloVe

Learning Gender-NeutralWord Embeddings (2018) - https://arxiv.org/pdf/1809.01496.pdf
Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them - https://arxiv.org/pdf/1903.03862.pdf

Contents

Dataset

Dictionary

- 4 pre-made wordlists combined (2,400 words)
- Checked by linguists to find offensive in context words
- Added spelling variants and even more swear words

Raw datasets

- Competition datasets
- Multiple corpora, pre-annotated (Hate Sonar, Wikipedia Detox project, Kaggle, Vulgar Twitter, Waseem & Hovy 2016 – Twitter)
- Custom corpus scraped from the Internet and annotated by linguists

Cleaned dataset

- Removed too long and too short sentences (3-30 words)
- Removed incomprehensible sentences (tagged by linguists as „nonsense")
- Preprocessing: substituting user names and URL's with <USER> and <URL> tags, normalizing words and letter case.
- Final corpus: 98k phrases, 49k not offensive and 49k offensive.

How to recognize offensive sentences?

- Contains expletives/swear-words/offensive terms

  GOD DAMN GOD DAMN GOD DAMN GOD DAMN GOD DAMN GOD DAMN

- Expresses rude meaning

  Before I accuse you of cringeworthy acts with donkeys, what does sprotected mean?

- Carries meaning that is harsh politically/ethically/emotionally and so expresses hate/disgust/disrespect.

  Only Americans are degenerate enough to 'honor' their war dead by having a barbecue. Anyone who 'grills out' for Memorial Day is trash.

- Raises uncomfortable topics related to the human genitals, such as sexual orientation, defecation, in a gross way

  My girlfriend was on her Period and forgot to tell me one night, I was rather drunk and so failed to notice the smell of a fish market coming from her lower regions.

- Contains hate speech/sarcasm/sexism/racism/violence/etc.

  Why keep your jewish mutt nose when you breathe out of your nigger mouth anyways?

- Discusses using drugs or performing other illegal actions

  JUST SMOEK WEEED TWICE AS HARD!!!!

Dataset – Difficult sentences

For linguists

- Sentences that caused conflict in linguistic assessment.
- Sentences regarding ethnicity, religion, political views.
- Linguists often find swear words non-offensive.
- Sentences containing bad words depending on context.

Dataset – Removing bias

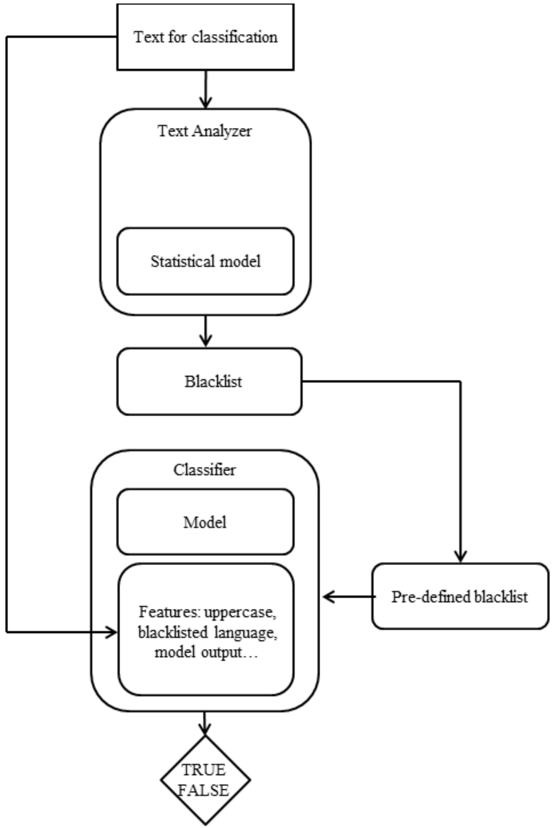
Bias removal?

Word	Toxic	Non-Toxic
atheist	0.009%	0.022%
arab	0.025%	0.013%
muslim	0.038%	0.060%
islam	0.003%	0.013%
queer	0.126%	0%
gay	0.445%	0.066%
transgender	0%	0%
lesbian	0.022%	0%
homosexual	0.009%	0%
feminist	0.006%	0.003%
black	0.644%	0.193%
white	1.133%	0.098%
heterosexual	0.003%	0%
bisexual	0.003%	0%

- Tried to make the dataset as balanced as possible.
- Checking the words contained in the sentence (based on the dictionary).
- In some sources less bias than in others.

Contents

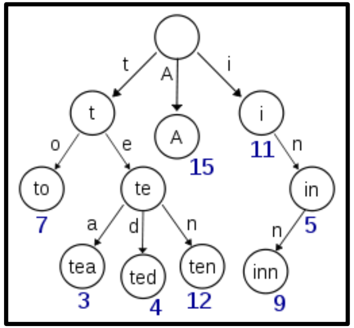
Pipeline


Text for classification
Text Analyzer
Statistical model
Blacklist
Classifier
Model
Features: uppercase, blacklisted language, model output…
Pre-defined blacklist
TRUE
FALSE

Model

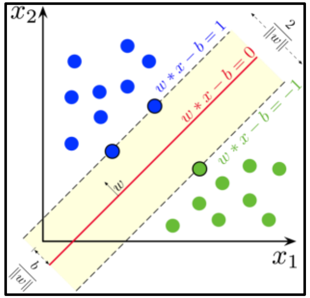
Features

- Linguistic features



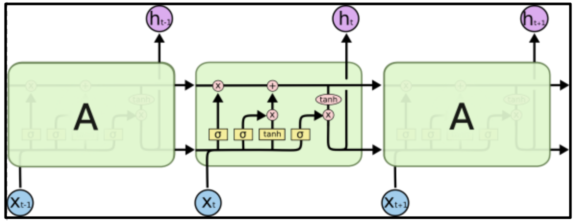
Machine learning

- Sentence classifier (SVM, Random forest)
  - NLTK tokenizer
  - Tf-Idf
  - Classifier



Deep learning

- LSTM + Attention
- Transformer (OpenAI GPT, Custom implementation)
- Embeddings: ELMo, fastText 1M, Universal Sentence Encoder, custom embedding



Contents

Results – SemEval tasks

Task	Macro F1	Placement
6 A Offensive/Not Offensive	0.80	8th position
6 B Targeted/Untargeted	0.69	9th position
6 C Individual/Group/ Organization/Other	0.63	**2nd position**
5 A Hate speech against women o r immigrants	0.51	8th position

Results – Internal

	Test			
Model	Accuracy	Precision	Recall	F-score
SVM	0.93	0.93	0.93	0.93
Lstm (Custom)	0.90	0.90	0.90	0.90
Lstm (Random)	0.89	0.89	0.89	0.89
Lstm (Elmo)	0.91	0.91	0.91	0.91
Lstm (Fastext-1M)	0.89	0.89	0.89	0.89
Finetune	0.95	0.95	0.95	0.95

Results – Internal, Difficult

	Test difficult			
Model	Accuracy	Precision	Recall	F-score
SVM	0.59	0.62	0.59	0.59
Lstm (Custom)	0.58	0.60	0.58	0.58
Lstm (Random)	0.60	0.61	0.60	0.60
Lstm (Elmo)	0.61	0.63	0.61	0.62
Lstm (Fastext-1M)	0.59	0.62	0.59	0.59
Finetune	0.61	0.65	0.61	0.62

Difficult examples

	Expected	SVM	Lstm Custom	Lstm Random	Lstm ELMO	Lstm 1M	Finetune
quit using song lyrics queer	**FILTER**	FILTER	FILTER	CLEAN	CLEAN	FILTER	FILTER
Matt your so luckey to be going out with jade"	**CLEAN**	CLEAN	CLEAN	CLEAN	FILTER	CLEAN	FILTER
You were an adopted failure	**FILTER**	CLEAN	CLEAN	FILTER	CLEAN	CLEAN	FILTER
Yeah he went of to trojan	**CLEAN**	CLEAN	FILTER	CLEAN	FILTER	FILTER	CLEAN
on church they told me that Jesus can walk on water, and I told them that Chuck Norris can walk on Jesus	**FILTER**	CLEAN	CLEAN	FILTER	FILTER	FILTER	FILTER
i hate being fat	**CLEAN**	FILTER	FILTER	FILTER	FILTER	FILTER	FILTER

Contents

Conclusions

- A classifier for detecting offensive sentences.
- Our architecture is suitable for mulitple (related) offensive sentence classification tasks.
- Our models and datasets.
- Still many challenges to solve.
- Socially debiasing Deep learning/ML model is difficult (if not impossible).
- Offensive sentence classification is a problem of classifying „another language”.

Further research

- Wider set of features.
- Using both a blacklist and a whitelist.
- Linguistic analysis of the corpus.

Thank You

SAMSUNG