#HeOrShe? - Online Data & Gender Privacy Issues

Privacy Enhancing Technologies - Yang Zhang

Presented By - Aleena Thomas, Shubham Agarwal

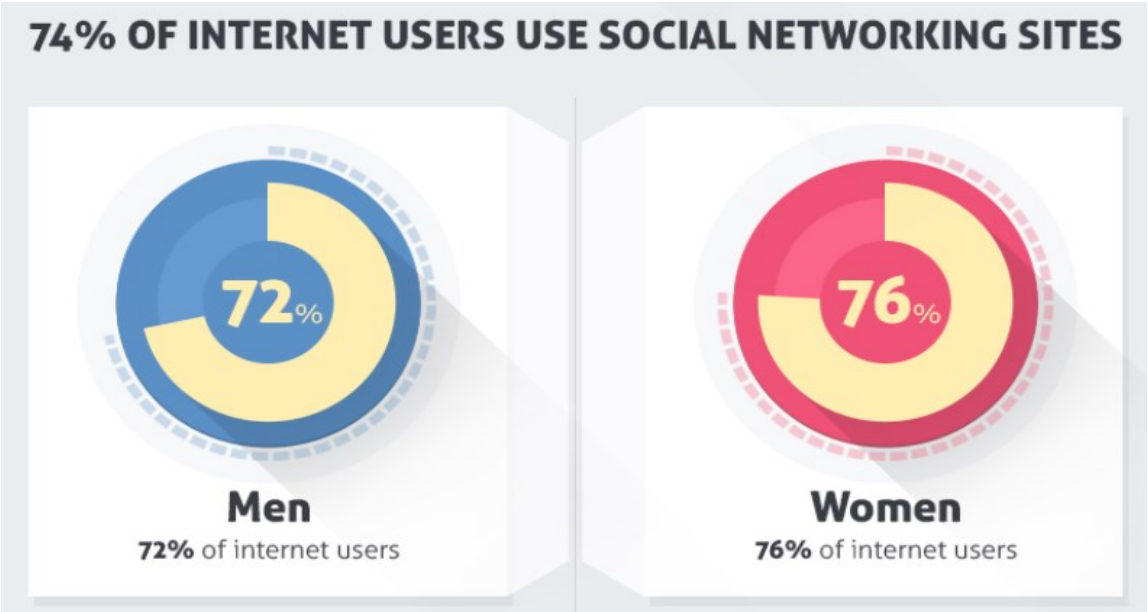
Motivation

Why Talk About Gender?

Gender-specific differences in product consumer behavior.

Why Talk About Gender?

Other Statistical Studies on Internet Users.



While at the same time…

Online Abuse Silences Women and Girls, Fuels Violence

By Nellie Peyton

A Violent Network – A discussion on online hate and sexual harassment

Technology-facilitated Gender-based Violence, Violence Against Women and Girls

And Unfortunately...

Goal

- Understand Gender-Based Classification Studies Performed by Research Community.
- Build simple experimental models to classify gender of online text data and identify associated challenges.

Base Study

- Motivation From “Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media Notebook for PAN at CLEF 2013”
- Extracted Dataset from PAN’13 Workshop website & Gender-based classification information to perform the experiment.

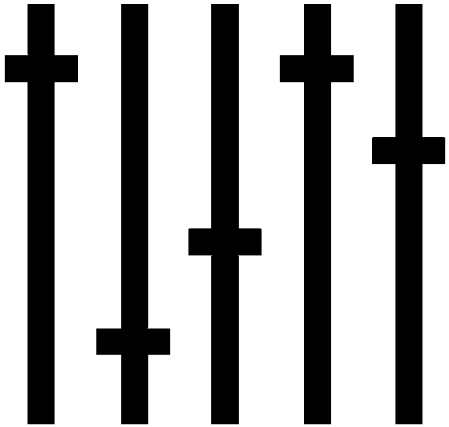
Relevant Discriminative Features Considered

- **Surface Features** - Length of Conversation & Words, use of special characters, emoticons and web links, etc.
- **Syntactical Features, Word N-gram Patterns & Punctuation** - grammar Metrics such as number of nouns, pronouns, etc; type & count of different punctuations used.
- **Semantic Features** - Relationship between Word n-grams and entities involved.
- Other content-based and linguistic features.

Experiments

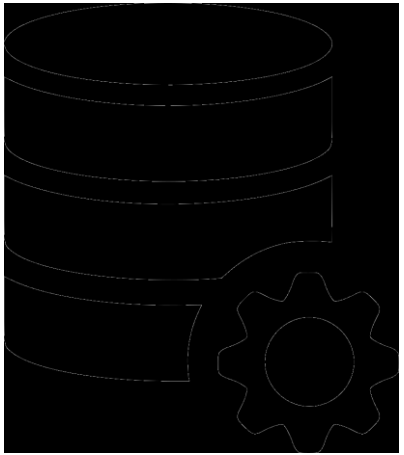
Experimental Models

- **Model 1** - Traditional Non-Neural Machine Learning Approach - SVM Classifier.
- **Model 2** - Neural Approach - CNN



Training Dataset

- **PAN’13 English Dataset** - Conversational Text Article - Posts/Opinion/Views on many different topics.
- **Size of Processed Dataset** - 166K
- **No. of Unique Tokens** - 77K



Training Dataset

Sample Document:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
- <author lang="en" age_group="30s" gender="female">
  - <conversations count="2">
      <conversation id="cded2472563e55c12d707557cea58bfa"> Though fiberglass  egarded a single of the sturdiest
        components available in the marketplace, numerous individuals... </conversation>
      <conversation id="d6a22c94da2767afb8b402f55f649340"> It sets hard, is less brittle than a lot of other products and is
        far cheaper to make. Its flexibility and strength... </conversation>
    </conversations>
</author>
```

Preprocessing

- Parse XML documents into text strings.
- Remove HTML Tags and unwanted characters such as digits, punctuations and other special symbols within each text string.
- Filter Word Tokens that are not present in WordNet.
- Use Stemming (PorterStemmer) & Lemmatizer, wherever needed, for generic token representation.

Experiment 1

Question:

How good a Bag of Words (BOW) approach is in gender classification ?

Experiment 1 - SVM

- **Idea:** traditional binary classifier
- **Input**: token vectors (by using Tf-idf Vectorizer)
- **Maximum Features** - Top 10,000 features
  - Variant 1 - RBF Kernel
  - Variant 2 - Linear Kernel

Experiment 1 - Results

RBF Kernel		
Iterations	**Hyperparameter, C**	**Accuracy**
2000	1000	51.4%
1500	1000	52.3%
1000	100	52.0%
500	1.0	52.0%

Experiment 1 - Results

Linear Kernel		
Iterations	**Hyperparameter, C**	**Accuracy**
200	100	52.4%
250	100	52.7%
300	100	52.8%
100	1	52.9%

Experiment 2

Questions:

- How to improve over BoW approach?
- Patterns?

Experiment 2 - CNN

- **Idea:** extract patterns in input text relevant for classification.
- **Maximum Features** - 50,000
- **3 Convolutional Layers** with Max Pooling
- **Regularization** - Dropout (*keep_prob* = 0.8/0.7)
- **Word Embeddings** - GloVe (100/300 dimensions)

Experiment 2 - Architecture

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 500, 100)	3762700
conv1d_1 (Conv1D)	(None, 500, 128)	64128
max_pooling1d_1 (MaxPooling1	(None, 100, 128)	0
dropout_1 (Dropout)	(None, 100, 128)	0
conv1d_2 (Conv1D)	(None, 100, 128)	82048
max_pooling1d_2 (MaxPooling1	(None, 20, 128)	0
dropout_2 (Dropout)	(None, 20, 128)	0
conv1d_3 (Conv1D)	(None, 20, 128)	82048
max_pooling1d_3 (MaxPooling1	(None, 4, 128)	0
dropout_3 (Dropout)	(None, 4, 128)	0
flatten_1 (Flatten)	(None, 512)	0
batch_normalization_1 (Batch	(None, 512)	2048
dropout_4 (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 128)	65664
batch_normalization_2 (Batch	(None, 128)	512
dense_2 (Dense)	(None, 2)	258

Total params: 4,059,406
Trainable params: 295,426
Non-trainable params: 3,763,980

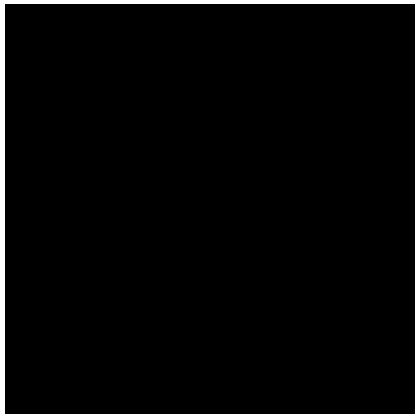
Experiment 2 - Results

Dropout Value	Epochs	Accuracy
0.2	11 - 20	58.7%
	40 -50	57.0%
0.3	21 - 30	59.0%
	40 - 50	59.3%

Conclusion

- BoW Approach using SVM - better than random classifier.
- Modelling the patterns in input text (using CNN) - comparatively better results than SVM.
- **Challenge** - Could not model long range dependencies in our experiment. Recent works like [1] give better results using RNN.

[1] Gender Classification with Deep Learning, Aric Bartle, Jim Zheng



Questions?