Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm

Xue Yuan
yx9485@gmail.com
Outline

• Background
• Introduction
• Method
• Experiments
• More analysis to the model
• Conclusion
Background

- **Current situation**
  Natural language processing is becoming more and more popular. But AI is still weak to infer the emotions of natural language, especially sarcasm and irony.

- **Requirement of market**
  a) Companies want to make sense of what their customers are saying about them.
  b) To improve the comprehension of Siri and Alex.

- **A challenge**
  Constructing a concrete set of rules to define emotions is hard, different ways to describe one emotion.
Introduction

• DeepMoji

• Why using emojis?

• How does DeepMoji work?

• Further goal of DeepMoji

• The paper try to show
Introduction

• DeepMoji:

DeepMoji is a deep learning model that uses millions of tweets with emojis to learn about emotional concepts in text.

I love mom's cooking

I love how you never reply back..

I love cruising with my homies

I love messing with yo mind!!

I love you and now you're just gone..

This is shit

This is the shit
Introduction

• Why using emojis:

A large amount of information on social media with emojis, which can be just treated as representatives of emotional contents, namely labels of sentences.

Without hearing of expression and body language, emoji is a nice choice. Unified form of emoji avoid ambiguity of emotion's defination
Introduction

• How does DeepMoji work:

From a dataset of 55 billion tweets, we find tweets with emojis and train a deep learning model to predict which emojis was included with what tweet.

The basic idea is that if the model is able to understand which emoji was included with a given sentence, then it has a good understanding of the emotional content of that sentence.
Introduction

• Further goal of DeepMoji:

We can then transfer this knowledge to a target task by doing just a little bit of additional training with the target dataset.
Introduction

• The paper try to show:

  • Extending the distant supervision to a more diverse set of noisy labels enables the models to learn richer representations (DeepMoji) of emotional content.

  • With the help of transfer learning, we can achieve better performance on similar target tasks (like detecting sentiments, emotions and sarcasm) across 5 domains (Tweets, Video Comments, Headlines....)
Method

- Emotional expressions as noisy labels
- Model of DeepMoji
- Transfer Learning
Method

• Emotional expressions as noisy labels
  • Previous research
    a) manually specify which emotional category each emotional expression belong to
    b) time-consuming
    c) prone to misinterpretations
  • The approach in paper
    a) requires no prior knowledge of the corpus, directly obtains “labeled” data from Twitter.
    b) captures diverse usage of 64 types of emojis.
    c) diversity of emoji set is important for transfer learning.
Method

- **Model of DeepMoji**
  - A Long Short-Term Memory (LSTM) model with attention mechanism.
  - Two bidirectional LSTM layers with 1024 hidden units in each (512 in each direction).
  - Attention layer that take all of these layers as input.

The reason why adding this attention mechanism will be specified later on.
Method

- **Transfer learning:**
  - Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task.
  
  - Rapid progress when modeling the second task, save computing power and time.

- *Transfer learning only works in deep learning if the model features learned from the first task are general.*
Method

• **Transfer learning:**
  • Actually the DeepMoji model is constructed to be the pretrained model for different practical target similar tasks.

  • In other words, target tasks take *pretrained DeepMoji Model as initialization of network*.

• Pretraining: avoid a bad local minima
  higher robustness
  small scale of parameter tuning
Method

• **Transfer learning:**
  • Two existing common approach:

  1. Use the pretrained model as an initialization, where the full model is unfrozen.
     
     **Limitation:** small amount of target task's training-set and huge amount of parameters (20 million) will cause overfitting rapidly.

  2. All layers in the pretrained model are frozen during fine-tuning on the target task, except the last layer.
     
     **Limitation:** the last layer could be too complex to target task
Method

• **Transfer learning:**
  - New transfer learning approach: “chain-thaw”, sequentially unfreezes and fine-tunes a single layer at a time

  • Adjust the individual patterns and expand the vocabulary to new domains with a reduced risk of overfitting.

![Diagram of transfer learning process](image-url)
Experiments

• Description of benchmark datasets

• Performance comparison across benchmark datasets
Experiments

• Description of benchmark datasets:

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Study</th>
<th>Task</th>
<th>Domain</th>
<th>Classes</th>
<th>$N_{train}$</th>
<th>$N_{test}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE0714</td>
<td>(Strapparava and Mihalcea, 2007)</td>
<td>Emotion</td>
<td>Headlines</td>
<td>3</td>
<td>250</td>
<td>1000</td>
</tr>
<tr>
<td>Olympic</td>
<td>(Sintsova et al., 2013)</td>
<td>Emotion</td>
<td>Tweets</td>
<td>4</td>
<td>250</td>
<td>709</td>
</tr>
<tr>
<td>PsychExp</td>
<td>(Wallbott and Scherer, 1986)</td>
<td>Emotion</td>
<td>Experiences</td>
<td>7</td>
<td>1000</td>
<td>6480</td>
</tr>
<tr>
<td>SS-Twitter</td>
<td>(Thelwall et al., 2012)</td>
<td>Sentiment</td>
<td>Tweets</td>
<td>2</td>
<td>1000</td>
<td>1113</td>
</tr>
<tr>
<td>SS-Youtube</td>
<td>(Thelwall et al., 2012)</td>
<td>Sentiment</td>
<td>Video Comments</td>
<td>2</td>
<td>1000</td>
<td>1142</td>
</tr>
<tr>
<td>SE1604</td>
<td>(Nakov et al., 2016)</td>
<td>Sentiment</td>
<td>Tweets</td>
<td>3</td>
<td>7155</td>
<td>31986</td>
</tr>
<tr>
<td>SCv1</td>
<td>(Walker et al., 2012)</td>
<td>Sarcasm</td>
<td>Debate Forums</td>
<td>2</td>
<td>1000</td>
<td>995</td>
</tr>
<tr>
<td>SCv2-GEN</td>
<td>(Oraby et al., 2016)</td>
<td>Sarcasm</td>
<td>Debate Forums</td>
<td>2</td>
<td>1000</td>
<td>2260</td>
</tr>
</tbody>
</table>

To emphasize the importance of the transfer learning ability of the evaluated models.
Experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Measure</th>
<th>State of the art</th>
<th>DeepMoji (new)</th>
<th>DeepMoji (full)</th>
<th>DeepMoji (last)</th>
<th>DeepMoji (chain-thaw)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE0714</td>
<td>F1</td>
<td>.34 [Buechel]</td>
<td>.21</td>
<td>.31</td>
<td>.36</td>
<td>.37</td>
</tr>
<tr>
<td>Olympic</td>
<td>F1</td>
<td>.50 [Buechel]</td>
<td>.43</td>
<td>.50</td>
<td>.61</td>
<td>.61</td>
</tr>
<tr>
<td>PsychExp</td>
<td>F1</td>
<td>.45 [Buechel]</td>
<td>.32</td>
<td>.42</td>
<td>.56</td>
<td>.57</td>
</tr>
<tr>
<td>SS-Twitter</td>
<td>Acc</td>
<td>.82 [Deriu]</td>
<td>.62</td>
<td>.85</td>
<td>.87</td>
<td>.88</td>
</tr>
<tr>
<td>SS-Youtube</td>
<td>Acc</td>
<td>.86 [Deriu]</td>
<td>.75</td>
<td>.88</td>
<td>.92</td>
<td>.93</td>
</tr>
<tr>
<td>SE1604</td>
<td>Acc</td>
<td>.51 [Deriu]³</td>
<td>.51</td>
<td>.54</td>
<td>.58</td>
<td>.58</td>
</tr>
<tr>
<td>SCv1</td>
<td>F1</td>
<td>.63 [Joshi]</td>
<td>.67</td>
<td>.65</td>
<td>.68</td>
<td>.69</td>
</tr>
<tr>
<td>SCv2-GEN</td>
<td>F1</td>
<td>.72 [Joshi]</td>
<td>.71</td>
<td>.71</td>
<td>.74</td>
<td>.75</td>
</tr>
</tbody>
</table>

- DeepMoji model outperforms the state of the art across all benchmark datasets.
- New ‘chain-thaw’ approach consistently yields the highest performance for the transfer learning.
- “Chain-thaw” slightly better or equal to the “last” approach. That means parameters in previous layers didn't change too much. That also reversely prove the feasibility of transfer learning of DeepMoji.
More Analysis to the model

• Additional importance of emoji diversity

• Importance of model architecture
More Analysis to the model

- **Additional importance of emoji diversity**
  - We also trained DeepMoji model to predict just positive or negative.
  - A new model called DeepMoji-PosNeg using training set with only P and N emoji.
  - Showing that the diversity of our emoji types encourage the model to learn a richer representation of emotional content in text that is more useful for transfer learning.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pos/Neg emojis</th>
<th>Standard LSTM</th>
<th>DeepMoji</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE0714</td>
<td>.32</td>
<td>.35</td>
<td>.36</td>
</tr>
<tr>
<td>Olympic</td>
<td>.55</td>
<td>.57</td>
<td>.61</td>
</tr>
<tr>
<td>PsychExp</td>
<td>.40</td>
<td>.49</td>
<td>.56</td>
</tr>
<tr>
<td>SS-Twitter</td>
<td>.86</td>
<td>.86</td>
<td>.87</td>
</tr>
<tr>
<td>SS-Youtube</td>
<td>.90</td>
<td>.91</td>
<td>.92</td>
</tr>
<tr>
<td>SE1604</td>
<td>.56</td>
<td>.57</td>
<td>.58</td>
</tr>
<tr>
<td>SCv1</td>
<td>.66</td>
<td>.66</td>
<td>.68</td>
</tr>
<tr>
<td>SCv2-GEN</td>
<td>.72</td>
<td>.73</td>
<td>.74</td>
</tr>
</tbody>
</table>
More Analysis to the model

- **Importance of model architecture**
  - We compare DeepMoji model (with attention mechanism) and standard 2-layer LSTM
  - Two architectures performed equally on the pretraining task, while DeepMoji model indeed better on transfer learning.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pos/Neg emojis</th>
<th>Standard LSTM</th>
<th>DeepMoji</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE0714</td>
<td>.32</td>
<td>.35</td>
<td>.36</td>
</tr>
<tr>
<td>Olympic</td>
<td>.55</td>
<td>.57</td>
<td>.61</td>
</tr>
<tr>
<td>PsychExp</td>
<td>.40</td>
<td>.49</td>
<td>.56</td>
</tr>
<tr>
<td>SS-Twitter</td>
<td>.86</td>
<td>.86</td>
<td>.87</td>
</tr>
<tr>
<td>SS-Youtube</td>
<td>.90</td>
<td>.91</td>
<td>.92</td>
</tr>
<tr>
<td>SE1604</td>
<td>.56</td>
<td>.57</td>
<td>.58</td>
</tr>
<tr>
<td>SCv1</td>
<td>.66</td>
<td>.66</td>
<td>.68</td>
</tr>
<tr>
<td>SCv2-GEN</td>
<td>.72</td>
<td>.73</td>
<td>.74</td>
</tr>
</tbody>
</table>
More Analysis to the model

• Importance of model architecture
  • Two factors (conjecture):
    a) The attention mechanism with skip-connections provides easy access to learn low-level information for any time step, if it is needed for a new task.

    b) Because of skip-connections, the improved gradient-flow is efficient when adjusting parameters in early layers as part of transfer learning to small datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pos/Neg emojis</th>
<th>Standard LSTM</th>
<th>DeepMoji</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE0714</td>
<td>.32</td>
<td>.35</td>
<td>.36</td>
</tr>
<tr>
<td>Olympic</td>
<td>.55</td>
<td>.57</td>
<td>.61</td>
</tr>
<tr>
<td>PsychExp</td>
<td>.40</td>
<td>.49</td>
<td>.56</td>
</tr>
<tr>
<td>SS-Twitter</td>
<td>.86</td>
<td>.86</td>
<td>.87</td>
</tr>
<tr>
<td>SS-Youtube</td>
<td>.90</td>
<td>.91</td>
<td>.92</td>
</tr>
<tr>
<td>SE1604</td>
<td>.56</td>
<td>.57</td>
<td>.58</td>
</tr>
<tr>
<td>SCv1</td>
<td>.66</td>
<td>.66</td>
<td>.68</td>
</tr>
<tr>
<td>SCv2-GEN</td>
<td>.72</td>
<td>.73</td>
<td>.74</td>
</tr>
</tbody>
</table>
Comparing with human-level agreement

• Comparison of agreement

<table>
<thead>
<tr>
<th>Method</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>50.1%</td>
</tr>
<tr>
<td>fastText</td>
<td>71.0%</td>
</tr>
<tr>
<td>MTurk</td>
<td>76.1%</td>
</tr>
<tr>
<td>DeepMoji</td>
<td>82.4%</td>
</tr>
</tbody>
</table>

DeepMoji performs better than human's judgement
Conclusion

• We find that the diversity of our emoji set is important for the performance of our method.

• Pretrained DeepMoji model with the hope that other researchers will find good use of them for various emotion-related NLP tasks.
Discussion
Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

Xue Yuan
yx9485@gmail.com
Outline

• Background

• Existing benchmarks

• Intersectional benchmark

• Pilot Parliaments Benchmark

• Commercial gender classifications on PPB

• Summary

• Conclusion
Background

• Facial Analysis Related Works

• Current situation

• Some existing benchmarks

• Serious implication
Background

• Facial Analysis Related Works

Baseline of Automated Facial Analysis
→ face detection, recognition and classification

Some furtherly gone works:
→ Identifying emotion from image of people's face
→ Determine an individual's characteristics like IQ, terrorism
→ Law enforcement to supervise and prevent potential crime
Background

- Current situation

  In widely used facial analysis benchmarks, lighter-skinned individuals overwhelm the datasets

  It means the famous “highly accurate” facial recognition tools are based on “inaccurate” benchmarks.
Background

• Some existing benchmarks

  • LFW dataset for face recognition with 77.5% male and 85.5% white
    $\rightarrow$ significant demographic bias

  • IJB-A from IARPA in 2015 for face recognition
    $\rightarrow$ geographically diverse

• Adience, released in 2014

• one common: underrepresentation of darker individuals
Background

• Serious implication

Many face recognition based on machine learning could have serious implications if the benchmarks are unbalanced.

Example: At least 117 million Americans are included in law enforcement face recognition networks and African-American individuals are more likely to be stopped by law enforcement.

Right of all citizen is under threat.
Intersectional Benchmark

- Dataset
- Intersectional evaluation
- Two Innovations
- Existing Benchmark Labeling
- Intersectional Benchmark Labeling
Intersectional Benchmark:

- **Dataset**
  
  New dataset with more balanced skin type and gender representations. Two labels: gender, skin type

- **Intersectional evaluation**
  
  Intersectional evaluation further requires a validation in subgroups by gender with a range of phenotypes.
Intersectional Benchmark:

• **Two Innovations**
  1) The frist benchmark composed of 1270 unique individuals that is more phenotypically balanced on the basis of skin type labeled.

  2) The first intersectional demographic and phenotypic evaluation. → examined on 4 intersectional subgroups: DF,DM,LF,LM
Intersectional Benchmark

• Existing Benchmark Labeling: *Race Labeling*

*Two Limitations of Race labeling:*
1) *within a racial category, phenotypic labels can vary widely*
2) *racial and ethnic categories are not consistent across geographies*
Intersectional Benchmark

- Intersectional Benchmark Labeling: *Phenotypic Labeling*

  Since race and ethnic labels are unstable

  Compared with Race Labeling, Phenotypic Labeling is a more *visually precise label* to measure dataset diversity

  Settled numbers of skin type
Pilot Parliaments Benchmark

- Creation of Pilot Parliaments Benchmark
- Why images of parliamentarians
- Image characteristics
- PPB Labeling
Pilot Parliaments Benchmark

• **Creation of Pilot Parliaments Benchmark**
  → better intersectional representation on the basis of gender and skin type
  → 1270 individuals from three African countries and European countries with gender parity.

• **Why images of parliamentarians**
  → public figures with known identities
  → profile photos taken under conditions with similar pose and illumination and expressions
Pilot Parliaments Benchmark

Intersectional representation based on gender and skin type
Pilot Parliaments Benchmark

- **Image characteristics**

<table>
<thead>
<tr>
<th>Property</th>
<th>PPB</th>
<th>IJB-A</th>
<th>Adience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Release Year</td>
<td>2017</td>
<td>2015</td>
<td>2014</td>
</tr>
<tr>
<td>#Subjects</td>
<td>1270</td>
<td>500</td>
<td>2284</td>
</tr>
<tr>
<td>Avg. IPD</td>
<td>63 pixels</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BBox Size</td>
<td>141 (avg)</td>
<td>≥36</td>
<td>-</td>
</tr>
<tr>
<td>IM Width</td>
<td>160-590</td>
<td>-</td>
<td>816</td>
</tr>
<tr>
<td>IM Height</td>
<td>213-886</td>
<td>-</td>
<td>816</td>
</tr>
</tbody>
</table>
Pilot Parliaments Benchmark

• PPB Labeling
  • Skin type labels
    → The six-point Fitzpatrick classification system labels skin types Type1-Type6, skewed from darker skin towards lighter skin
  • Gender labels
    → Gender identity (biological sex), in PPB using female and male to indicate subjects perceived as woman or man
    (Name, Title or appearance of photos)
Pilot Parliaments Benchmark

- PPB Labeling

![Bar chart showing the percentage of subjects in benchmarks](chart.png)

The table below shows the decomposition of PPB for different sets:

<table>
<thead>
<tr>
<th>Set</th>
<th>n</th>
<th>F</th>
<th>M</th>
<th>Darker</th>
<th>Lighter</th>
<th>DF</th>
<th>DM</th>
<th>LF</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Subjects</td>
<td>1270</td>
<td>44.6%</td>
<td>55.4%</td>
<td>46.4%</td>
<td>53.6%</td>
<td>21.3%</td>
<td>25.0%</td>
<td>23.3%</td>
<td>30.3%</td>
</tr>
<tr>
<td>Africa</td>
<td>661</td>
<td>43.9%</td>
<td>56.1%</td>
<td>86.2%</td>
<td>13.8%</td>
<td>39.8%</td>
<td>46.4%</td>
<td>4.1%</td>
<td>9.7%</td>
</tr>
<tr>
<td>South Africa</td>
<td>437</td>
<td>41.4%</td>
<td>58.6%</td>
<td>79.2%</td>
<td>20.8%</td>
<td>35.2%</td>
<td>43.9%</td>
<td>6.2%</td>
<td>14.6%</td>
</tr>
<tr>
<td>Senegal</td>
<td>149</td>
<td>43.0%</td>
<td>57.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>43.0%</td>
<td>57.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Rwanda</td>
<td>75</td>
<td>60.0%</td>
<td>40.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>60.0%</td>
<td>40.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Europe</td>
<td>609</td>
<td>45.5%</td>
<td>54.5%</td>
<td>3.1%</td>
<td>96.9%</td>
<td>1.3%</td>
<td>1.8%</td>
<td>44.2%</td>
<td>52.7%</td>
</tr>
<tr>
<td>Sweden</td>
<td>349</td>
<td>46.7%</td>
<td>53.3%</td>
<td>4.9%</td>
<td>95.1%</td>
<td>2.0%</td>
<td>2.9%</td>
<td>44.7%</td>
<td>50.4%</td>
</tr>
<tr>
<td>Finland</td>
<td>197</td>
<td>42.6%</td>
<td>57.4%</td>
<td>1.0%</td>
<td>99.0%</td>
<td>0.5%</td>
<td>0.5%</td>
<td>42.1%</td>
<td>56.9%</td>
</tr>
<tr>
<td>Iceland</td>
<td>63</td>
<td>47.6%</td>
<td>52.4%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>47.6%</td>
<td>52.4%</td>
</tr>
</tbody>
</table>
Commercial gender classifications on PPB

• Existing commercial gender classifications

• Performance on PPB

• Worst performance on darker female
Commercial gender classifications on PPB

- Existing commercial gender classifications
  - Microsoft, IBM, Face++
  - Male subjects were more accurately classified than female subjects
  - Lighter subjects were more accurately classified than darker subjects
  - All classifiers perform worst on darker female subjects
Commercial gender classifications on PPB

- Performance on PPB

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Metric</th>
<th>All</th>
<th>F</th>
<th>M</th>
<th>Darker</th>
<th>Lighter</th>
<th>DF</th>
<th>DM</th>
<th>LF</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSFT</td>
<td>PPV(%)</td>
<td>93.7</td>
<td>89.3</td>
<td>97.4</td>
<td>87.1</td>
<td>99.3</td>
<td>79.2</td>
<td>94.0</td>
<td>98.3</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>6.3</td>
<td>10.7</td>
<td>2.6</td>
<td>12.9</td>
<td>0.7</td>
<td>20.8</td>
<td>6.0</td>
<td>1.7</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>93.7</td>
<td>96.5</td>
<td>91.7</td>
<td>87.1</td>
<td>99.3</td>
<td>92.1</td>
<td>83.7</td>
<td>100</td>
<td>98.7</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>6.3</td>
<td>8.3</td>
<td>3.5</td>
<td>12.9</td>
<td>0.7</td>
<td>16.3</td>
<td>7.9</td>
<td>1.3</td>
<td>0.0</td>
</tr>
<tr>
<td>Face++</td>
<td>PPV(%)</td>
<td>90.0</td>
<td>78.7</td>
<td>99.3</td>
<td>83.5</td>
<td>95.3</td>
<td>65.5</td>
<td><strong>99.3</strong></td>
<td>94.0</td>
<td>99.2</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>10.0</td>
<td>21.3</td>
<td>0.7</td>
<td>16.5</td>
<td>4.7</td>
<td><strong>34.5</strong></td>
<td>0.7</td>
<td>6.0</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>90.0</td>
<td>98.9</td>
<td>85.1</td>
<td>83.5</td>
<td>95.3</td>
<td>98.8</td>
<td>76.6</td>
<td><strong>98.9</strong></td>
<td>92.9</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>10.0</td>
<td>14.9</td>
<td>1.1</td>
<td>16.5</td>
<td>4.7</td>
<td><strong>23.4</strong></td>
<td>1.2</td>
<td>7.1</td>
<td>1.1</td>
</tr>
<tr>
<td>IBM</td>
<td>PPV(%)</td>
<td>87.9</td>
<td>79.7</td>
<td>94.4</td>
<td>77.6</td>
<td>96.8</td>
<td>65.3</td>
<td>88.0</td>
<td>92.9</td>
<td><strong>99.7</strong></td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>12.1</td>
<td>20.3</td>
<td>5.6</td>
<td>22.4</td>
<td>3.2</td>
<td><strong>34.7</strong></td>
<td>12.0</td>
<td>7.1</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>87.9</td>
<td>92.1</td>
<td>85.2</td>
<td>77.6</td>
<td>96.8</td>
<td>82.3</td>
<td>74.8</td>
<td><strong>99.6</strong></td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>12.1</td>
<td>14.8</td>
<td>7.9</td>
<td>22.4</td>
<td>3.2</td>
<td><strong>25.2</strong></td>
<td>17.7</td>
<td>5.20</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Gender classification performance on PPB
Commercial gender classifications on PPB

- Performance on South African subset of PPB

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Metric</th>
<th>DF</th>
<th>DM</th>
<th>LF</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PPV(%)</td>
<td>76.2</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>MSFT</td>
<td>Error Rate(%)</td>
<td>23.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>TPR(%)</td>
<td>100</td>
<td>84.2</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>FPR(%)</td>
<td>15.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Face++</td>
<td>PPV(%)</td>
<td>64.0</td>
<td>99.5</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>36.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>TPR(%)</td>
<td>99.0</td>
<td>77.8</td>
<td>100</td>
<td>96.9</td>
</tr>
<tr>
<td></td>
<td>FPR(%)</td>
<td>22.2</td>
<td>1.03</td>
<td>3.08</td>
<td>0.0</td>
</tr>
<tr>
<td>IBM</td>
<td>PPV(%)</td>
<td>66.9</td>
<td>94.3</td>
<td>100</td>
<td>98.4</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>33.1</td>
<td>5.7</td>
<td>0.0</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>TPR(%)</td>
<td>90.4</td>
<td>78.0</td>
<td>96.4</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>FPR(%)</td>
<td>22.0</td>
<td>9.7</td>
<td>0.0</td>
<td>3.6</td>
</tr>
</tbody>
</table>
Commercial gender classifications on PPB

• Worst performance on darker female

• Intersectional demographic and phenotypic performance: darker females account for the largest proportion of misclassified subjects.

• Darker skin may be highly correlated with facial geometries or gender display norms that were less represented in the training data of the evaluated classifiers.
Summary

• We annotated the dataset of benchmark with the Fitzpatrick skin classification system and validated gender classification performance on 4 subgroups: DM, DF, LM, LF

• We measured the accuracy of 3 commercial gender classifiers on the new Pilot Parliaments Benchmark which is balanced by gender and skin type.

• We found that all classifiers perform better for lighter and male individuals overall. The classifiers perform worst for darker females.
Conclusion

- This work aimed to show why we need rigorous reporting on the performance metrics

- The work focuses on increasing phenotypic and demographic representation in face datasets and evaluation.

- Increase transparency and accountability in artificial intelligence.
Discussion
Thank you!