Age Prediction in Blogs
A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations

Sara Rosenthal and Kathleen McKeown
Columbia University
ACL 2011
Motivation

• Age prediction is useful for targeted advertising
• Age prediction can be used to improve results in larger tasks such as identifying opinion, persuasion, and power
Research Questions

• What are the differences in how people communicate that most directly reveal their age?

• Which age groups should be chosen to yield the best results?
  • Under and over 18
  • 10s vs. 20s vs. 30s
Approach

- Supervised Machine Learning
- Identify influential features
  - Online Behavior
    - e.g. Number of posts
  - Lexical-Stylistic
    - e.g. Sentence length
  - Lexical-Content
    - e.g. Bag of words
- Perform Classification
  - Using Logistic Regression
Corpus

• 24,500 LiveJournal Blogs
• Each blog is written by a person living in the US.
• Each blog includes the blogger’s age in the profile
• Each blogger has written at least one entry in 2009 or later
Corpus

NUM BLOGGERS

Year of Birth

Older

Younger
Corpus

NUM BLOGGERS

Year of Birth

1981 - 1988

Older
Younger
### Features: Online Behavior

<table>
<thead>
<tr>
<th>Feature</th>
<th>Explanation</th>
<th>Example</th>
<th>Trend as Age Decreases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interests</td>
<td>Top interests provided on their profile page (LiveJournal only)</td>
<td>disney</td>
<td>N/A</td>
</tr>
<tr>
<td># of Friends</td>
<td>Number of friends the blogger has</td>
<td>45</td>
<td>fluctuates</td>
</tr>
<tr>
<td># of Posts</td>
<td>Number of downloadable posts (0-25)</td>
<td>23</td>
<td>decrease</td>
</tr>
<tr>
<td># of Lifetime Posts</td>
<td>Number of posts written in total</td>
<td>821</td>
<td>decrease</td>
</tr>
<tr>
<td>Time</td>
<td>The mode hour (00-23) and day the blogger posts</td>
<td>11, Monday</td>
<td>no change</td>
</tr>
<tr>
<td>Comments</td>
<td>Average number of comments per post</td>
<td>2.64</td>
<td>increase</td>
</tr>
</tbody>
</table>
## Online Behavior Interests

- Extracted the top 200 interests per age group.
- The value refers to the *position* of the interest in its list.
- Keep discriminating interests

<table>
<thead>
<tr>
<th>18-22</th>
<th>28-32</th>
<th>38-42</th>
</tr>
</thead>
<tbody>
<tr>
<td>reading</td>
<td>3</td>
<td>reading</td>
</tr>
<tr>
<td>reading</td>
<td>1</td>
<td>reading</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>drawing</td>
<td>10</td>
<td>love</td>
</tr>
<tr>
<td>fanfiction</td>
<td>11</td>
<td>drawing</td>
</tr>
<tr>
<td>love</td>
<td>15</td>
<td>sci-fi</td>
</tr>
<tr>
<td>disney</td>
<td>39</td>
<td>tori amos</td>
</tr>
<tr>
<td>yaoi</td>
<td>40</td>
<td>hiking</td>
</tr>
<tr>
<td>johnny depp</td>
<td>42</td>
<td>fanfiction</td>
</tr>
<tr>
<td>rent</td>
<td>44</td>
<td>women</td>
</tr>
</tbody>
</table>

...
Online Behavior
Interests

- Extracted the top 200 interests per age group.
- The value refers to the *position* of the interest in its list.
- Keep discriminating interests

<table>
<thead>
<tr>
<th></th>
<th>18-22</th>
<th>28-32</th>
<th>38-42</th>
</tr>
</thead>
<tbody>
<tr>
<td>reading</td>
<td>3</td>
<td>reading</td>
<td>1</td>
</tr>
<tr>
<td>drawing</td>
<td>10</td>
<td>love</td>
<td>24</td>
</tr>
<tr>
<td>fanfiction</td>
<td>11</td>
<td>drawing</td>
<td>25</td>
</tr>
<tr>
<td>love</td>
<td>15</td>
<td>sci-fi</td>
<td>37</td>
</tr>
<tr>
<td>disney</td>
<td>39</td>
<td>tori amos</td>
<td>49</td>
</tr>
<tr>
<td>yaoi</td>
<td>40</td>
<td>hiking</td>
<td>55</td>
</tr>
<tr>
<td>johnny depp</td>
<td>42</td>
<td>fanfiction</td>
<td>58</td>
</tr>
<tr>
<td>rent</td>
<td>44</td>
<td>women</td>
<td>61</td>
</tr>
</tbody>
</table>
Online Behavior
Interests

- Extracted the top 200 interests per age group.
- The value refers to the *position* of the interest in its list.
- Keep discriminating interests

<table>
<thead>
<tr>
<th></th>
<th>18-22</th>
<th>28-32</th>
<th>38-42</th>
</tr>
</thead>
<tbody>
<tr>
<td>reading</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>drawing</td>
<td>10</td>
<td>24</td>
<td>21</td>
</tr>
<tr>
<td>fanfiction</td>
<td>11</td>
<td>drawing</td>
<td>25</td>
</tr>
<tr>
<td>love</td>
<td>15</td>
<td>sci-fi</td>
<td>37</td>
</tr>
<tr>
<td>disney</td>
<td>39</td>
<td>tori amos</td>
<td>49</td>
</tr>
<tr>
<td>yaoi</td>
<td>40</td>
<td>hiking</td>
<td>55</td>
</tr>
<tr>
<td>johnny depp</td>
<td>42</td>
<td>fanfiction</td>
<td>58</td>
</tr>
<tr>
<td>rent</td>
<td>44</td>
<td>women</td>
<td>61</td>
</tr>
</tbody>
</table>
Online Behavior

Interests

- Extracted the top 200 interests per age group.
- The value refers to the *position* of the interest in its list.
- Keep discriminating interests

<table>
<thead>
<tr>
<th></th>
<th>18-22</th>
<th></th>
<th>28-32</th>
<th></th>
<th>38-42</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>reading</strong></td>
<td>3</td>
<td><strong>reading</strong></td>
<td>1</td>
<td><strong>reading</strong></td>
<td>1</td>
</tr>
<tr>
<td><strong>drawing</strong></td>
<td>10</td>
<td><strong>love</strong></td>
<td>24</td>
<td><strong>sci-fi</strong></td>
<td>21</td>
</tr>
<tr>
<td><strong>fanfiction</strong></td>
<td>11</td>
<td><strong>drawing</strong></td>
<td>25</td>
<td><strong>love</strong></td>
<td>34</td>
</tr>
<tr>
<td><strong>love</strong></td>
<td>15</td>
<td><strong>sci-fi</strong></td>
<td>37</td>
<td><strong>polyamory</strong></td>
<td>40</td>
</tr>
<tr>
<td><strong>disney</strong></td>
<td>39</td>
<td><strong>tori amos</strong></td>
<td>49</td>
<td><strong>drawing</strong></td>
<td>65</td>
</tr>
<tr>
<td><strong>yaoi</strong></td>
<td>40</td>
<td><strong>hiking</strong></td>
<td>55</td>
<td><strong>sca</strong></td>
<td>67</td>
</tr>
<tr>
<td><strong>johnny depp</strong></td>
<td>42</td>
<td><strong>fanfiction</strong></td>
<td>58</td>
<td><strong>babylon 5</strong></td>
<td>84</td>
</tr>
<tr>
<td><strong>rent</strong></td>
<td>44</td>
<td><strong>women</strong></td>
<td>61</td>
<td><strong>leather</strong></td>
<td>94</td>
</tr>
</tbody>
</table>
Online Behavior
Friends: *no clear trend*

![Graph showing the number of friends over the year of birth, with no clear trend.](image-url)
Online Behavior

Lifetime Posts: older have more posts
Online Behavior

Comments: younger have more comments

![Graph showing the average number of comments in a single entry by year of birth. The graph indicates a trend where younger individuals have more comments over time.]
# Features: Lexical - Stylistic

<table>
<thead>
<tr>
<th>Feature</th>
<th>Explanation</th>
<th>Example</th>
<th>Trend as Age Decreases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emoticons</td>
<td>Number of emoticons(^1)</td>
<td>:)</td>
<td>increase</td>
</tr>
<tr>
<td>Acronyms</td>
<td>Number of internet acronyms(^1)</td>
<td>lol</td>
<td>increase</td>
</tr>
<tr>
<td>Slang</td>
<td>Number of words that are not found in the dictionary(^1)</td>
<td>wazzup</td>
<td>increase</td>
</tr>
<tr>
<td>Punctuation</td>
<td>Number of stand-alone punctuation(^1)</td>
<td>!</td>
<td>increase</td>
</tr>
<tr>
<td>Capitalization</td>
<td>Number of words (with length of &gt;1) that are all CAPS(^1)</td>
<td>YOU</td>
<td>increase</td>
</tr>
<tr>
<td>Links/Images</td>
<td>Number of URL and image links(^1)</td>
<td><a href="http://www.site.com">www.site.com</a></td>
<td>fluctuates</td>
</tr>
<tr>
<td>Sentence Length</td>
<td>average sentence length</td>
<td>40</td>
<td>decrease</td>
</tr>
</tbody>
</table>

\(^1\) Normalized per sentence per entry
Lexical - Stylistic
Emoticons: younger use more emoticons
Lexical - Stylistic
Capitalizations: younger use more capital words

CAPITALIZATIONS

Per Sentence, Per Entry vs. Year of Birth

Older   |   Younger
Lexical - Stylistic

Sentence Length: *older have longer sentences*
Lexical - Stylistic
Links/Images: no clear trend
## Features: Lexical - Content

<table>
<thead>
<tr>
<th>Feature</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collocations</td>
<td>Top collocations in the age group</td>
<td>in [] relationship</td>
</tr>
<tr>
<td>Syntax Collocations</td>
<td>Top syntax collocations in the age group&lt;sup&gt;1&lt;/sup&gt;</td>
<td>have [] clue</td>
</tr>
<tr>
<td>POS Collocations</td>
<td>Top Part-of-Speech (POS) collocations in the age group</td>
<td>wouldn’t VB</td>
</tr>
<tr>
<td>Words</td>
<td>Top words in the age group</td>
<td>his</td>
</tr>
</tbody>
</table>

---

### Words

- Extracted the top 200 words per age group using post frequency
- The value refers to the *position* of the word in its list.
- Keep discriminating words

<table>
<thead>
<tr>
<th></th>
<th>18-22</th>
<th>28-32</th>
<th>38-42</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ldqout (')</td>
<td>great</td>
<td>may</td>
</tr>
<tr>
<td>2</td>
<td>t</td>
<td>find</td>
<td>old</td>
</tr>
<tr>
<td>3</td>
<td>school</td>
<td>many</td>
<td>house</td>
</tr>
<tr>
<td>4</td>
<td>x</td>
<td>years</td>
<td>world</td>
</tr>
<tr>
<td>5</td>
<td>anything</td>
<td>week</td>
<td>please</td>
</tr>
</tbody>
</table>
Related Work

• Age and geographic inferences of the LiveJournal social network [Mackinnon and Warren 2006]
  • Use mean age of a bloggers social network
  • +/- 5 years: 98% accuracy

• An exploration of observable features related to blogger age [Burger and Henderson 2006]
  • Identify style and online behavior features
  • Under/Over 18: unsuccessful

• Effects of Age and Gender in Blogging [Schler et al 2006]
  • Use style and content features to predict whether a blogger is in their 10s, 20s, or 30s

Experiments

• Data: Portions of the LiveJournal corpus
• Classification: Logistic Regression and 10-fold cross-validation in Weka [Hall et al]
• Statistical significance: t-test

Experiment I
Age Groups

NUM BLOGGERS

0 200 400 600 800 1000 1200 1400 1600
Older Age Younger
Experiment I
Age Groups

NUM BLOGGERS

0 200 400 600 800 1000 1200 1400 1600

Older Age Younger

22-29
Experiment I
Age Groups

NUM BLOGGERS

Older  Age  Younger

30s  20s

10s
Experiment I
Age Groups – Schler et al
Experiment I
Age Groups

NUM BLOGGERS

Age

Older
30s
38-42

20s
28-32

Younger
10s
18-22
Experiment I
Age Groups

- Schler et al’s corpus is much larger

<table>
<thead>
<tr>
<th></th>
<th>Blogger (Schler et al)</th>
<th>LiveJournal</th>
</tr>
</thead>
<tbody>
<tr>
<td># Blogs</td>
<td>19,320</td>
<td>11,521</td>
</tr>
<tr>
<td># Posts</td>
<td>1.4 million</td>
<td>256,000</td>
</tr>
<tr>
<td># of Words</td>
<td>295 million</td>
<td>50 million</td>
</tr>
</tbody>
</table>
Experiment I
Age Groups

- 10s vs. 30s is classified the best
- 10s vs. 20s are classified well
- 20s vs. 30s are classified the worst
- Adding online behavior features increased accuracy by 10%

<table>
<thead>
<tr>
<th></th>
<th>Blogger (Schler et al)</th>
<th>LiveJournal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline</td>
<td>44% (13-17)</td>
<td>48% (28-32)</td>
</tr>
<tr>
<td>Style and Content</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Online-Behavior</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>10s vs. 30s</td>
<td>96%</td>
<td>92%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95%</td>
</tr>
<tr>
<td>10s vs. 20s</td>
<td>87%</td>
<td>72%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>81%</td>
</tr>
<tr>
<td>20s vs. 30s</td>
<td>77%</td>
<td>68%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>72%</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>76%</td>
<td>57%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>67%</td>
</tr>
</tbody>
</table>
Experiment I
Age Groups

- 10s vs. 30s is classified the best
- 10s vs. 20s are classified well
- 20s vs. 30s are classified the worst
- Adding online behavior features increased accuracy by 10%

<table>
<thead>
<tr>
<th></th>
<th>Blogger (Schler et al)</th>
<th>LiveJournal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline</td>
<td>44% (13-17)</td>
<td>48% (28-32)</td>
</tr>
<tr>
<td>Style and Content</td>
<td>✓</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>Online-Behavior</td>
<td></td>
<td>✓ ✓</td>
</tr>
<tr>
<td>10s vs. 30s</td>
<td>96%</td>
<td>92%</td>
</tr>
<tr>
<td>10s vs. 20s</td>
<td>87%</td>
<td>72%</td>
</tr>
<tr>
<td>20s vs. 30s</td>
<td>77%</td>
<td>68%</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>76%</td>
<td>57%</td>
</tr>
</tbody>
</table>
## Experiment I
### Age Groups

- 10s vs. 30s is classified the best
- 10s vs. 20s are classified well
- 20s vs. 30s are classified the worst
- Adding online behavior features increased accuracy by 10%

<table>
<thead>
<tr>
<th></th>
<th>Blogger (Schler et al)</th>
<th>LiveJournal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline</td>
<td>44% (13-17)</td>
<td>48% (28-32)</td>
</tr>
<tr>
<td>Style and Content</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Online-Behavior</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>10s vs. 30s</td>
<td>96%</td>
<td>92%</td>
</tr>
<tr>
<td>10s vs. 20s</td>
<td>87%</td>
<td>72%</td>
</tr>
<tr>
<td>20s vs. 30s</td>
<td>77%</td>
<td>68%</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>76%</td>
<td>57%</td>
</tr>
</tbody>
</table>
Experiment I
Age Groups

- **10s vs. 30s** is classified the best
- **10s vs. 20s** are classified well
- **20s vs. 30s** are classified the worst
- Adding online behavior features increased accuracy by 10%

<table>
<thead>
<tr>
<th></th>
<th>Blogger (Schler et al)</th>
<th>LiveJournal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline</td>
<td>44% (13-17)</td>
<td>48% (28-32)</td>
</tr>
<tr>
<td>Style and Content</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Online-Behavior</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10s vs. 30s</td>
<td>96%</td>
<td>92%</td>
</tr>
<tr>
<td>10s vs. 20s</td>
<td>87%</td>
<td>72%</td>
</tr>
<tr>
<td>20s vs. 30s</td>
<td>77%</td>
<td>68%</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>76%</td>
<td>57%</td>
</tr>
</tbody>
</table>
## Experiment I

### Age Groups

- 10s vs. 30s is classified the best
- **10s vs. 20s are classified well**
- 20s vs. 30s are classified the worst
- Adding online behavior features increased accuracy by 10%

<table>
<thead>
<tr>
<th></th>
<th>Blogger (Schler et al)</th>
<th>LiveJournal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline</td>
<td>44% (13-17)</td>
<td>48% (28-32)</td>
</tr>
<tr>
<td>Style and Content</td>
<td>✓</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>Online-Behavior</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>10s vs. 30s</td>
<td>96%</td>
<td>92%</td>
</tr>
<tr>
<td><strong>10s vs. 20s</strong></td>
<td><strong>87%</strong></td>
<td>72%</td>
</tr>
<tr>
<td>20s vs. 30s</td>
<td>77%</td>
<td>68%</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>76%</td>
<td>57%</td>
</tr>
</tbody>
</table>
Experiment I
Age Groups

- 10s vs. 30s is classified the best
- 10s vs. 20s are classified well
- **20s vs. 30s are classified the worst**
- Adding online behavior features increased accuracy by 10%

<table>
<thead>
<tr>
<th></th>
<th>Blogger (Schler et al)</th>
<th>LiveJournal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline</td>
<td>44% (13-17)</td>
<td>48% (28-32)</td>
</tr>
<tr>
<td>Style and Content</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Online-Behavior</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>10s vs. 30s</td>
<td>96%</td>
<td>92%</td>
</tr>
<tr>
<td>10s vs. 20s</td>
<td>87%</td>
<td>72%</td>
</tr>
<tr>
<td><strong>20s vs. 30s</strong></td>
<td><strong>77%</strong></td>
<td><strong>68%</strong></td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>76%</td>
<td>57%</td>
</tr>
</tbody>
</table>
Experiment I
Age Groups

- 10s vs. 30s is classified the best
- 10s vs. 20s are classified well
- 20s vs. 30s are classified the worst
- **Adding online behavior features increased accuracy by 10%**

<table>
<thead>
<tr>
<th></th>
<th>Blogger (Schler et al)</th>
<th>LiveJournal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline</td>
<td>44% (13-17)</td>
<td>48% (28-32)</td>
</tr>
<tr>
<td>Style and Content</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Online-Behavior</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10s vs. 30s</td>
<td>96%</td>
<td>92%</td>
</tr>
<tr>
<td>10s vs. 20s</td>
<td>87%</td>
<td>72%</td>
</tr>
<tr>
<td>20s vs. 30s</td>
<td>77%</td>
<td>68%</td>
</tr>
<tr>
<td><strong>Overall Accuracy</strong></td>
<td><strong>76%</strong></td>
<td><strong>57%</strong></td>
</tr>
</tbody>
</table>
Experiment I
Age Groups – Schler et al

NUM BLOGGERS

- Older
- Age
- Younger

NUM BLOGGERS

- 30s (33-37)
- 20s (23-27)
- 10s (13-17)

Age Groups:

- Younger
- Older

NUM BLOGGERS
Experiment I
Age Groups

NUM BLOGGERS

0 200 400 600 800 1000 1200 1400 1600

Older 30s 20s 10s
38-42 28-32 18-22
Younger
Experiment II
Social Media and Generation Y

• Generation Y uses social networking, blogs, and instant messaging more than their elders [Zickuhr 2010]

• For each birth year $X = 1975-1988$
  • get 1500 blogs ($\sim 33,000$ posts) balanced across years BEFORE $X$
  • get 1500 blogs ($\sim 33,000$ posts) balanced across years IN/AFTER $X$
  • Perform binary classification between blogs BEFORE $X$ and IN/AFTER $X$

Experiment II
Social Media and Generation Y

Style vs. Content

Online-Behavior + Lexical Stylistic

BOW

Accuracy (%)

Older Year of Birth Younger
Experiment II
Social Media and Generation Y

BOW + Online-Behavior + Lexical-Stylistic

Accuracy (%)  

Older  Year of Birth  Younger
Experiment II
Year of Birth

BOW + Online-Behavior + Lexical-Stylistic

Accuracy (%)
## Experiment II
Social Media and Generation Y

**College Age:** 22 21 20 19 18

<table>
<thead>
<tr>
<th>Year or birth</th>
<th>Year became popular</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Forums (1995)</td>
<td>18 18 22 22 21 22 21 20 19 20 22 22 22</td>
</tr>
<tr>
<td>AIM (1997)</td>
<td>20 18 20 18 21 21 20 19 20 22 22 22 22</td>
</tr>
<tr>
<td>Weblogs (1999)</td>
<td>22 20 21 18 21 20 19 20 22 22 22 22 22</td>
</tr>
<tr>
<td>SMS Messaging (2000)</td>
<td>22 20 21 18 21 20 19 20 22 22 22 22 22</td>
</tr>
<tr>
<td>MySpace (2003)</td>
<td>22 20 21 18 21 20 19 20 22 22 22 22 22</td>
</tr>
<tr>
<td>Facebook (2004)</td>
<td>22 20 21 18 21 20 19 20 22 22 22 22 22</td>
</tr>
<tr>
<td>Twitter (2007)</td>
<td>22 20 21 18 21 20 19 20 22 22 22 22 22</td>
</tr>
</tbody>
</table>
# Experiment III

## A closer look

<table>
<thead>
<tr>
<th>Features</th>
<th>1979</th>
<th>1984</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baselines</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online-Behavior</td>
<td>59.7</td>
<td>61.6</td>
</tr>
<tr>
<td>Interests</td>
<td>70.2</td>
<td>74.6</td>
</tr>
<tr>
<td><strong>Excluding Interests</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical-Stylistic</td>
<td>65.4</td>
<td>67.3</td>
</tr>
<tr>
<td>Slang + Emoticons + Acronyms</td>
<td>60.6</td>
<td>62.1</td>
</tr>
<tr>
<td>Online-Behavior + Lexical-Stylistic</td>
<td>67.2</td>
<td>71.3</td>
</tr>
<tr>
<td>BOW</td>
<td>75.3</td>
<td>77.8</td>
</tr>
<tr>
<td>BOW + Online-Behavior</td>
<td>76.4</td>
<td>79.2</td>
</tr>
<tr>
<td>BOW + Online-Behavior + Lexical-Stylistic</td>
<td>77.5</td>
<td>80.9</td>
</tr>
<tr>
<td>BOW + Online-Behavior + Lexical-Stylistic + Syntax Collocations</td>
<td>74.8</td>
<td>80.4</td>
</tr>
<tr>
<td><strong>Including Interests</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOW + Online-Behavior + Interests + Lexical-Stylistic</td>
<td>80</td>
<td>81.6</td>
</tr>
<tr>
<td><strong>All Features</strong></td>
<td>71.3</td>
<td>74.1</td>
</tr>
</tbody>
</table>
### Experiment III
A closer look - *baselines*

<table>
<thead>
<tr>
<th></th>
<th>Features</th>
<th>1979</th>
<th>1984</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baselines</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online-Behavior</td>
<td>59.7</td>
<td>61.6</td>
<td></td>
</tr>
<tr>
<td>Interests</td>
<td>70.2</td>
<td>74.6</td>
<td></td>
</tr>
<tr>
<td><strong>Excluding Interests</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical-Stylistic</td>
<td>65.4</td>
<td>67.3</td>
<td></td>
</tr>
<tr>
<td>Slang + Emoticons + Acronyms</td>
<td>60.6</td>
<td>62.1</td>
<td></td>
</tr>
<tr>
<td>Online-Behavior + Lexical-Stylistic</td>
<td>67.2</td>
<td>71.3</td>
<td></td>
</tr>
<tr>
<td>BOW</td>
<td>75.3</td>
<td>77.8</td>
<td></td>
</tr>
<tr>
<td>BOW + Online-Behavior</td>
<td>76.4</td>
<td>79.2</td>
<td></td>
</tr>
<tr>
<td>BOW + Online-Behavior + Lexical-Stylistic</td>
<td>77.5</td>
<td>80.9</td>
<td></td>
</tr>
<tr>
<td>BOW + Online-Behavior + Lexical-Stylistic + Syntax Collocations</td>
<td>74.8</td>
<td>80.4</td>
<td></td>
</tr>
<tr>
<td><strong>Including Interests</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOW + Online-Behavior + Interests + Lexical-Stylistic</td>
<td>80</td>
<td>81.6</td>
<td></td>
</tr>
<tr>
<td>All Features</td>
<td>71.3</td>
<td>74.1</td>
<td></td>
</tr>
</tbody>
</table>
## Experiment III

**A closer look - Lexical-Stylistic**

<table>
<thead>
<tr>
<th>Features</th>
<th>1979</th>
<th>1984</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baselines</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online-Behavior</td>
<td>59.7</td>
<td>61.6</td>
</tr>
<tr>
<td>Interests</td>
<td>70.2</td>
<td>74.6</td>
</tr>
<tr>
<td><strong>Excluding Interests</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online-Behavior</td>
<td>65.4</td>
<td>67.3</td>
</tr>
<tr>
<td>Slang + Emoticons + Acronyms</td>
<td>60.6</td>
<td>62.1</td>
</tr>
<tr>
<td>Online-Behavior + Lexical-Stylistic</td>
<td>67.2</td>
<td>71.3</td>
</tr>
<tr>
<td>BOW</td>
<td>75.3</td>
<td>77.8</td>
</tr>
<tr>
<td>BOW + Online-Behavior</td>
<td>76.4</td>
<td>79.2</td>
</tr>
<tr>
<td>BOW + Online-Behavior + Lexical-Stylistic</td>
<td>77.5</td>
<td>80.9</td>
</tr>
<tr>
<td>BOW + Online-Behavior + Lexical-Stylistic + Syntax Collocations</td>
<td>74.8</td>
<td>80.4</td>
</tr>
<tr>
<td><strong>Including Interests</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOW + Online-Behavior + Interests + Lexical-Stylistic</td>
<td>80</td>
<td>81.6</td>
</tr>
<tr>
<td>All Features</td>
<td>71.3</td>
<td>74.1</td>
</tr>
</tbody>
</table>
## Experiment III

### A closer look - *Lexical-Content*

<table>
<thead>
<tr>
<th>Features</th>
<th>1979</th>
<th>1984</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baselines</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online-Behavior</td>
<td>59.7</td>
<td>61.6</td>
</tr>
<tr>
<td>Interests</td>
<td>70.2</td>
<td>74.6</td>
</tr>
<tr>
<td><strong>Excluding Interests</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical-Stylistic</td>
<td>65.4</td>
<td>67.3</td>
</tr>
<tr>
<td>Slang + Emoticons + Acronyms</td>
<td>60.6</td>
<td>62.1</td>
</tr>
<tr>
<td>Online-Behavior + Lexical-Stylistic</td>
<td>67.2</td>
<td>71.3</td>
</tr>
<tr>
<td>BOW</td>
<td>75.3</td>
<td>77.8</td>
</tr>
<tr>
<td>BOW + Online-Behavior</td>
<td>76.4</td>
<td>79.2</td>
</tr>
<tr>
<td>BOW + Online-Behavior + Lexical-Stylistic</td>
<td><strong>77.5</strong></td>
<td><strong>80.9</strong></td>
</tr>
<tr>
<td>BOW + Online-Behavior + Lexical-Stylistic + Syntax Collocations</td>
<td>74.8</td>
<td>80.4</td>
</tr>
<tr>
<td><strong>Including Interests</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOW + Online-Behavior + Interests + Lexical-Stylistic</td>
<td><strong>80</strong></td>
<td><strong>81.6</strong></td>
</tr>
<tr>
<td>All Features</td>
<td>71.3</td>
<td>74.1</td>
</tr>
</tbody>
</table>
Conclusion & Future Work

• Style, Content, and Online Behavior are useful in age prediction

• Significant changes in writing style coincide with the popularity of social media technologies

• In the future, we want to experiment with using ranking, regression, and/or clustering for age prediction
Effects of Age and Gender on Blogging

- Experiment on 3 age groups consisting of 19,320 bloggers downloaded from blogger.com in 2004
- 1,405,209 blog entries and 295,526,889 words
- Features
  - Style-based:
    - select part of speech (pronouns, determiners)
    - function words (negation, assent)
    - blog-specific (hyperlinks, post length, blog words)
  - Content-based:
    - content words – with the highest information gain
    - LIWC categories – job, money

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>13-17</td>
<td>8240</td>
</tr>
<tr>
<td>23-27</td>
<td>8086</td>
</tr>
<tr>
<td>33-37</td>
<td>2994</td>
</tr>
</tbody>
</table>

Effects of Age and Gender on Blogging

- Use the Multi-Class Real Winnow learning algorithm
- Majority Baseline (13-17): 43.8%
- Results
  - Find content to be slightly more useful than style, but the combination is most useful.
  - 10s are distinguishable from 30’s with accuracy above 96%
  - 10s are distinguishable from 20’s with accuracy of 87.3%
  - Many 30s are misclassified as 20’s, yielding overall accuracy of 76.2%

<table>
<thead>
<tr>
<th></th>
<th>10s</th>
<th>20s</th>
<th>30s</th>
</tr>
</thead>
<tbody>
<tr>
<td>10s</td>
<td>7036</td>
<td>1027</td>
<td>177</td>
</tr>
<tr>
<td>20s</td>
<td>916</td>
<td>6326</td>
<td>844</td>
</tr>
<tr>
<td>30s</td>
<td>178</td>
<td>1465</td>
<td>1351</td>
</tr>
</tbody>
</table>

- But... age changes. Will the writing style of these people change drastically in a few years?

LiveJournal

- Livejournal provides a bot policy page providing data formats to be used instead of scraping the website.
  - XML Profile format
  - XML Entries format
  - Friends Data
  - Interests Data
- Modifications
  - Added type of friend connection to profile information
    - MutualFriend, Friend, FriendOf
  - Added interest count for entire LiveJournal network
  - Added comments to entries
- Java API
  - Can be used to download and read blogs
- Downloaded Data
  - 10000 blogs all containing age (Downloaded in 2009)
  - 30000 blogs containing age, originating from US, UK, AU, CA, and updated within the last two years (Downloaded in 2010)
Livejournal Blog

• Profiles
  • profile name
  • full name
  • age
  • e-mail
  • bio
  • interests
  • friends list
  • Country and city/state
  • etc...

• Blog Entries
  • entries
    • date written
    • profile name
  • comments
    • profile name of person that left the comment
    • date written
    • thread structure
Xtract

Retrieving Collocations from Text

• Stage 1: Extract significant collocations from all words within a +/- 5 window. Filter out words that do not have significant strength and spread. Part of speech can be used as part of the collocation as well.
  • strength: \( w_1 \) occurs with \( w_0 \) more than average
  • spread:

• Stage 2: Use bigrams from stage 1 to create n-grams
  • Precision: 40%

• Stage 3: Filter bigrams further to keep the ones that have good syntax relationships such as VO,SV,NN, NJ. Use those collocations to create new n-grams.
  • Precision: 80%, Recall: 90%*

* Figure taken from paper