Personality Prediction in Facebook Status Updates Using Multilevel N-gram Profiles (MNP) and Word Features

GEORGIOS MIKROS, VASSILIKI POUKI, EPHTYCHIA TRIANTAFYLLOU
NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS, GREECE
Outline

- Motivation
- Related work
- Our contribution
- Personality (BFM)
- Experiments
- Results
- Future research
Motivation

• Prior research indicated a relationship between author’s personality and the frequency of specific linguistic features that he/she uses on the text he/she writes.

• However, previous research has been focused mostly on “long” texts.
  • Can personality characteristics be detected on Facebook (FB) users via their short text 'Status Updates'?
Why Facebook?

• Huge impact on social relations across the globe
  ◦ Worldwide, there are over 1.71 billion monthly active Facebook users
  ◦ 1.13 billion people log onto Facebook daily active users
  ◦ There are 1.57 billion mobile active users
  ◦ Five new profiles are created every second.
  ◦ Every 60 seconds on Facebook: 293,000 statuses are updated, and 136,000 photos are uploaded.
Why status updates?

Since the inception of Facebook in 2004, status updates have been one of its most preferred features. Status updates allow users to share their thoughts, feelings, and activities with friends, who have the opportunity to “like” and comment in return.

However... it is an under-researched feature (compared to others, e.g. likes, number of friends etc) and focus specifically on the linguistic behavior of the user.
Previous Research

• Personality characteristics correlation with language features in FB, blogs, writings and conversation (not using our feature sets)
Our contribution

Facebook status updates
- LIWC linguistic concepts
- Multilevel N-gram Profiles

Personality
Personality: The big five factor model (FFM)

- Low Score
  - Practical, conventional, prefers routine
  - Impulsive, careless, disorganized
  - Quiet, reserved, withdrawn
  - Critical, uncooperative, suspicious
  - Calm, even-tempered, secure

- Trait
  - Openness (imagination, feelings, actions, ideas)
  - Conscientiousness (competence, self-discipline, thoughtfulness, goal-driven)
  - Extroversion (sociability, assertiveness, emotional expression)
  - Agreeableness (cooperative, trustworthy, good-natured)
  - Neuroticism (tendency toward unstable emotions)

- High Score
  - Curious, wide range of interests, independent
  - Hardworking, dependable, organized
  - Outgoing, warm, seeks adventure
  - Helpful, trusting, empathetic
  - Anxious, unhappy, prone to negative emotions
MyPersonality Project Data
(Bachrach, Kosinski, Graepel, Kohli & Stillwell (2012), Ortigosa, Carro & Quiroga (2014))

FB app 2007-2012.

1. FB users: questionnaire (Big5, political, religious beliefs...)

2. User Data collection.

3. 20+ millions of users

4. Data: Likes, psychometric tests' scores, FB status updates, Big5 scores

---

MyPersonality Project Data
(Bachrach, Kosinski, Graepel, Kohli & Stillwell (2012), Ortigosa, Carro & Quiroga (2014))

FB app 2007-2012.

1. FB users: questionnaire (Big5, political, religious beliefs...)

2. User Data collection.

3. 20+ millions of users

4. Data: Likes, psychometric tests' scores, FB status updates, Big5 scores

---

MyPersonality Project Data
(Bachrach, Kosinski, Graepel, Kohli & Stillwell (2012), Ortigosa, Carro & Quiroga (2014))

FB app 2007-2012.

1. FB users: questionnaire (Big5, political, religious beliefs...)

2. User Data collection.

3. 20+ millions of users

4. Data: Likes, psychometric tests' scores, FB status updates, Big5 scores

---

MyPersonality Project Data
(Bachrach, Kosinski, Graepel, Kohli & Stillwell (2012), Ortigosa, Carro & Quiroga (2014))

FB app 2007-2012.

1. FB users: questionnaire (Big5, political, religious beliefs...)

2. User Data collection.

3. 20+ millions of users

4. Data: Likes, psychometric tests' scores, FB status updates, Big5 scores

---
Stylometric features

1. LIWC

- Dictionary containing words that belong to 64 language and psychological processes
- Language-dependent, semantics, no Greek

2. Multilevel Ngrams Profiles

- Language Independent features measured in FB status updates (~3-15 words)
- n={2,3}, word/character layers, total 2000 most frequent ngrams
1. LIWC
Francis & Pennebaker (1993)/ LIWC2007: revised and improved version

- Efficient method for personality study
- Word count text analysis

LIWC

- Basic linguistic dimensions (e.g. % of pronouns, articles etc.)
- Categories of personal matters (eg work, home, leisure)
- Punctuation symbols (full stops, commas etc.).
- Psychological factors words (+/- emotions, family, friends ...)
Sample LIWC Features

LIWC (Linguistic Inquiry and Word Count)


<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger words</td>
<td>LIWC</td>
<td>hate, kill, pissed</td>
</tr>
<tr>
<td>Metaphysical issues</td>
<td>LIWC</td>
<td>God, heaven, coffin</td>
</tr>
<tr>
<td>Physical state/function</td>
<td>LIWC</td>
<td>ache, breast, sleep</td>
</tr>
<tr>
<td>Inclusive words</td>
<td>LIWC</td>
<td>with, and, include</td>
</tr>
<tr>
<td>Social processes</td>
<td>LIWC</td>
<td>talk, us, friend</td>
</tr>
<tr>
<td>Family members</td>
<td>LIWC</td>
<td>mom, brother, cousin</td>
</tr>
<tr>
<td>Past tense verbs</td>
<td>LIWC</td>
<td>walked, were, had</td>
</tr>
<tr>
<td>References to friends</td>
<td>LIWC</td>
<td>pal, buddy, coworker</td>
</tr>
</tbody>
</table>
Sample LIWC output

<table>
<thead>
<tr>
<th>Filename</th>
<th>WC</th>
<th>WPS</th>
<th>Sixtr</th>
<th>Dic</th>
<th>funct</th>
<th>pronoun</th>
<th>pppron</th>
<th>i</th>
<th>we</th>
<th>you</th>
<th>she</th>
<th>they</th>
<th>ipron</th>
<th>article</th>
<th>verb</th>
<th>auxver</th>
</tr>
</thead>
<tbody>
<tr>
<td>LWCG50Test_001.txt</td>
<td>1</td>
<td>19</td>
<td>21.05</td>
<td>94.74</td>
<td>68.4</td>
<td>10.53</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10.53</td>
<td>10.53</td>
<td>15.79</td>
<td>10.53</td>
</tr>
<tr>
<td>LWCG50Test_002.txt</td>
<td>1</td>
<td>36</td>
<td>30.56</td>
<td>97.22</td>
<td>63.9</td>
<td>5.56</td>
<td>2.78</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.78</td>
<td>2.78</td>
<td>2.78</td>
<td>11.89</td>
<td>2.78</td>
</tr>
<tr>
<td>LWCG50Test_003.txt</td>
<td>1</td>
<td>34</td>
<td>23.53</td>
<td>88.24</td>
<td>61.8</td>
<td>5.88</td>
<td>2.94</td>
<td>0</td>
<td>0</td>
<td>2.94</td>
<td>0</td>
<td>2.94</td>
<td>1.79</td>
<td>5.88</td>
<td>11.76</td>
<td>8.82</td>
</tr>
<tr>
<td>LWCG50Test_004.txt</td>
<td>1</td>
<td>91</td>
<td>23.08</td>
<td>81.32</td>
<td>63.7</td>
<td>13.19</td>
<td>8.79</td>
<td>1.1</td>
<td>1</td>
<td>0</td>
<td>7.69</td>
<td>0</td>
<td>4.4</td>
<td>6.59</td>
<td>10.99</td>
<td>8.79</td>
</tr>
<tr>
<td>LWCG50Test_005.txt</td>
<td>1</td>
<td>118</td>
<td>14.75</td>
<td>11.02</td>
<td>87.29</td>
<td>65.3</td>
<td>22.03</td>
<td>8.47</td>
<td>2.54</td>
<td>4.24</td>
<td>0.85</td>
<td>0.85</td>
<td>13.56</td>
<td>5.93</td>
<td>16.95</td>
<td>16.1</td>
</tr>
<tr>
<td>LWCG50Test_006.txt</td>
<td>1</td>
<td>21</td>
<td>28.57</td>
<td>90.48</td>
<td>52.4</td>
<td>9.52</td>
<td>9.52</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>23.81</td>
<td>19.05</td>
<td></td>
</tr>
<tr>
<td>LWCG50Test_007.txt</td>
<td>1</td>
<td>53</td>
<td>26.42</td>
<td>90.57</td>
<td>41.5</td>
<td>7.55</td>
<td>3.77</td>
<td>1.89</td>
<td>0</td>
<td>1.89</td>
<td>0</td>
<td>3.77</td>
<td>5.66</td>
<td>18.87</td>
<td>5.66</td>
<td></td>
</tr>
<tr>
<td>LWCG50Test_008.txt</td>
<td>1</td>
<td>100</td>
<td>16.67</td>
<td>26</td>
<td>81</td>
<td>45</td>
<td>20</td>
<td>12</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>LWCG50Test_009.txt</td>
<td>1</td>
<td>8</td>
<td>8</td>
<td>12.5</td>
<td>75</td>
<td>37.5</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>12.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12.5</td>
<td>0</td>
</tr>
<tr>
<td>LWCG50Test_010.txt</td>
<td>1</td>
<td>59</td>
<td>29.5</td>
<td>18.64</td>
<td>79.66</td>
<td>55.9</td>
<td>6.78</td>
<td>3.39</td>
<td>0</td>
<td>1.69</td>
<td>0</td>
<td>1.69</td>
<td>3.39</td>
<td>16.95</td>
<td>10.17</td>
<td>6.78</td>
</tr>
<tr>
<td>LWCG50Test_011.txt</td>
<td>1</td>
<td>121</td>
<td>30.25</td>
<td>21.49</td>
<td>90.91</td>
<td>51.2</td>
<td>4.96</td>
<td>1.65</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.65</td>
<td>0</td>
<td>3.31</td>
<td>9.09</td>
<td>13.27</td>
</tr>
<tr>
<td>LWCG50Test_012.txt</td>
<td>1</td>
<td>46</td>
<td>15.33</td>
<td>30.43</td>
<td>76.09</td>
<td>45.7</td>
<td>13.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13.04</td>
<td>2.17</td>
<td>10.87</td>
<td>6.52</td>
</tr>
<tr>
<td>LWCG50Test_013.txt</td>
<td>1</td>
<td>166</td>
<td>13.83</td>
<td>25.9</td>
<td>74.1</td>
<td>47</td>
<td>11.45</td>
<td>7.83</td>
<td>4.82</td>
<td>1.2</td>
<td>0</td>
<td>1.2</td>
<td>3.61</td>
<td>4.82</td>
<td>10.24</td>
<td>6.02</td>
</tr>
<tr>
<td>LWCG50Test_014.txt</td>
<td>1</td>
<td>79</td>
<td>26.33</td>
<td>37.97</td>
<td>82.28</td>
<td>45.6</td>
<td>5.06</td>
<td>3.8</td>
<td>2.53</td>
<td>0</td>
<td>1.27</td>
<td>0</td>
<td>1.27</td>
<td>6.33</td>
<td>5.06</td>
<td></td>
</tr>
<tr>
<td>LWCG50Test_015.txt</td>
<td>1</td>
<td>35</td>
<td>17.5</td>
<td>20</td>
<td>91.43</td>
<td>48.6</td>
<td>14.29</td>
<td>11.43</td>
<td>5.71</td>
<td>2.86</td>
<td>0</td>
<td>2.86</td>
<td>2.86</td>
<td>14.29</td>
<td>2.86</td>
<td></td>
</tr>
<tr>
<td>LWCG50Test_016.txt</td>
<td>1</td>
<td>78</td>
<td>19.5</td>
<td>30.77</td>
<td>83.33</td>
<td>46.2</td>
<td>5.13</td>
<td>3.85</td>
<td>2.56</td>
<td>1.28</td>
<td>0</td>
<td>1.28</td>
<td>7.69</td>
<td>7.69</td>
<td>6.41</td>
<td></td>
</tr>
<tr>
<td>LWCG50Test_017.txt</td>
<td>1</td>
<td>40</td>
<td>20</td>
<td>37.5</td>
<td>82.5</td>
<td>45</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td>LWCG50Test_018.txt</td>
<td>1</td>
<td>64</td>
<td>21.33</td>
<td>20.31</td>
<td>87.5</td>
<td>54.7</td>
<td>9.38</td>
<td>4.69</td>
<td>1.56</td>
<td>1.56</td>
<td>0</td>
<td>1.56</td>
<td>4.69</td>
<td>7.81</td>
<td>14.06</td>
<td>9.38</td>
</tr>
<tr>
<td>LWCG50Test_019.txt</td>
<td>1</td>
<td>22</td>
<td>11</td>
<td>27.27</td>
<td>86.36</td>
<td>59.1</td>
<td>9.09</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9.09</td>
<td>18.18</td>
<td>18.18</td>
<td></td>
</tr>
<tr>
<td>LWCG50Test_020.txt</td>
<td>1</td>
<td>52</td>
<td>17.33</td>
<td>15.38</td>
<td>94.23</td>
<td>69.2</td>
<td>15.38</td>
<td>7.69</td>
<td>0</td>
<td>5.77</td>
<td>1.92</td>
<td>0</td>
<td>3.85</td>
<td>23.08</td>
<td>17.31</td>
<td></td>
</tr>
<tr>
<td>LWCG50Test_021.txt</td>
<td>1</td>
<td>76</td>
<td>12.67</td>
<td>26.32</td>
<td>89.47</td>
<td>60.5</td>
<td>14.47</td>
<td>9.21</td>
<td>1.32</td>
<td>7.89</td>
<td>0</td>
<td>0</td>
<td>5.26</td>
<td>3.95</td>
<td>17.11</td>
<td>13.16</td>
</tr>
<tr>
<td>LWCG50Test_022.txt</td>
<td>1</td>
<td>142</td>
<td>20.29</td>
<td>23.94</td>
<td>85.92</td>
<td>57.8</td>
<td>14.08</td>
<td>7.75</td>
<td>1.41</td>
<td>5.63</td>
<td>0.7</td>
<td>0</td>
<td>6.34</td>
<td>4.93</td>
<td>15.49</td>
<td>11.97</td>
</tr>
<tr>
<td>LWCG50Test_023.txt</td>
<td>1</td>
<td>82</td>
<td>27.33</td>
<td>15.85</td>
<td>82.93</td>
<td>61</td>
<td>10.98</td>
<td>7.32</td>
<td>0</td>
<td>7.32</td>
<td>0</td>
<td>0</td>
<td>3.66</td>
<td>12.2</td>
<td>7.32</td>
<td></td>
</tr>
<tr>
<td>LWCG50Test_024.txt</td>
<td>1</td>
<td>48</td>
<td>16</td>
<td>14.58</td>
<td>95.83</td>
<td>50</td>
<td>12.5</td>
<td>4.17</td>
<td>2.08</td>
<td>0</td>
<td>8.33</td>
<td>0</td>
<td>8.33</td>
<td>10.42</td>
<td>6.25</td>
<td></td>
</tr>
<tr>
<td>LWCG50Test_025.txt</td>
<td>1</td>
<td>117</td>
<td>23.4</td>
<td>18.8</td>
<td>78.63</td>
<td>51.3</td>
<td>9.4</td>
<td>3.42</td>
<td>0.85</td>
<td>0.85</td>
<td>1.71</td>
<td>0</td>
<td>5.98</td>
<td>11.11</td>
<td>4.27</td>
<td>2.56</td>
</tr>
<tr>
<td>LWCG50Test_026.txt</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>28.57</td>
<td>71.43</td>
<td>57.1</td>
<td>14.29</td>
<td>14.29</td>
<td>0</td>
<td>14.29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14.29</td>
<td>28.57</td>
<td>14.29</td>
</tr>
<tr>
<td>LWCG50Test_027.txt</td>
<td>1</td>
<td>32</td>
<td>16</td>
<td>25</td>
<td>71.88</td>
<td>43.8</td>
<td>12.5</td>
<td>6.25</td>
<td>3.12</td>
<td>3.12</td>
<td>0</td>
<td>0</td>
<td>6.25</td>
<td>3.12</td>
<td>15.62</td>
<td>15.62</td>
</tr>
</tbody>
</table>
2. Multilevel Ngram Profiles (MNP)

- Efficient feature space model for capturing a wide spectrum of the author’s linguistic production.
- Robust performance even in very small texts (e.g. tweets, blog posts and emails, Mikros & Perifanos 2012, 2013, 2015).
Big 5 characteristics overlapping

Overlapping: Each user can belong to one or more categories
Statistical analysis

We used a multivariate linear regression analysis:
- **Dependent variables:** Each of the personality traits scores
- **Independent variables:** LIWC variables and MNP features
- **Goodness of model fit:** $R^2$
Prediction – 1st experiment

Big5 with LIWC
1. Openness ($R^2=0.056$)

<table>
<thead>
<tr>
<th>In accordance with previous research</th>
<th>In contrast with previous research</th>
</tr>
</thead>
<tbody>
<tr>
<td>• use many verbs of past and present.</td>
<td>• Frequent self-references (use 1st person pronouns)</td>
</tr>
<tr>
<td>• use cognitive verbs (<em>know, ought</em>)</td>
<td></td>
</tr>
<tr>
<td>• words with negative emotional content (<em>hurt, ugly</em>)</td>
<td></td>
</tr>
</tbody>
</table>
2. Conscientiousness ($R^2=0.088$)

In accordance with previous research

- Words with positive emotional content (love, sweet)
- Words related to work (job)
- Conjunctions
- Personal pronouns in 1st and 2nd person
- Verbs in present and future tense
4. Agreeableness $(R^2=0.055)$

In contrast with previous research

- They don't choose words with positive emotional content (*love, sweet*)
- No self-references

Extra findings

- Verbs in past tense, knowledge verbs, praise verbs (*think, consider*), modals (*should, would, could*).
- Intercalary types and sounds or combinations of sounds of void content (*hm, umm, er*)
5. Neuroticism ($R^2=0.049$)

In accordance with previous research

- Many self-references
- Use of words related to anxiety and concern.

In contrast with previous research

- Words with positive emotional content (love, sweet)
- Personal pronouns of 1st, 2nd, 3rd person in singular.
Prediction – 2\textsuperscript{nd} experiment

Big5 with MNP
1. Conscientiousness ($R^2=0.295$)

Top-3 ngrams

*ready for the*
  - ready for the dance!!! (motivation)

*I can t*
  - I can t do the job (reliability)

*. My*
  - installing new lighting...My lower back already hurts (discipline, ambition, hard-working)
2. Extroversion ($R^2=0.29$)

Strong interpersonal, social and active people

Top-2 ngrams

*tm*
- got you something shinny for Christmas
- merry Christmas to everyone!!
- Im secretly batman
- nightmare revisited best cd EVA!!!

*TH*
- HAPPY BIRTHDAY JOHN LENNON
- I LOVE THE RAIN!!
- CONGRATULATIONS TO THE WAUTOMA FOOTBALL TEAM
MNP & Big5

For the categories: **Openness, Neuroticism, Agreeableness** the model did not fit well (n.s. for p<0.05).
Conclusion

Satisfactory $R^2$ fit for very small texts as are the status updates (5-10 words on average)
Future research

- Merging the above mentioned feature sets
- Exploring other features (sentiment analysis vocabularies, lexical “richness” indices etc)
- Apply to other textual genres or other social networks.
- Implementation of alternative learning methods and comparison of results.
Thank you!