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

GEORGIOS MIKROS , VASSILIKI POULI , EPHTYCHIA TRIANTAFYLLOU

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS, GREECE

Outline

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

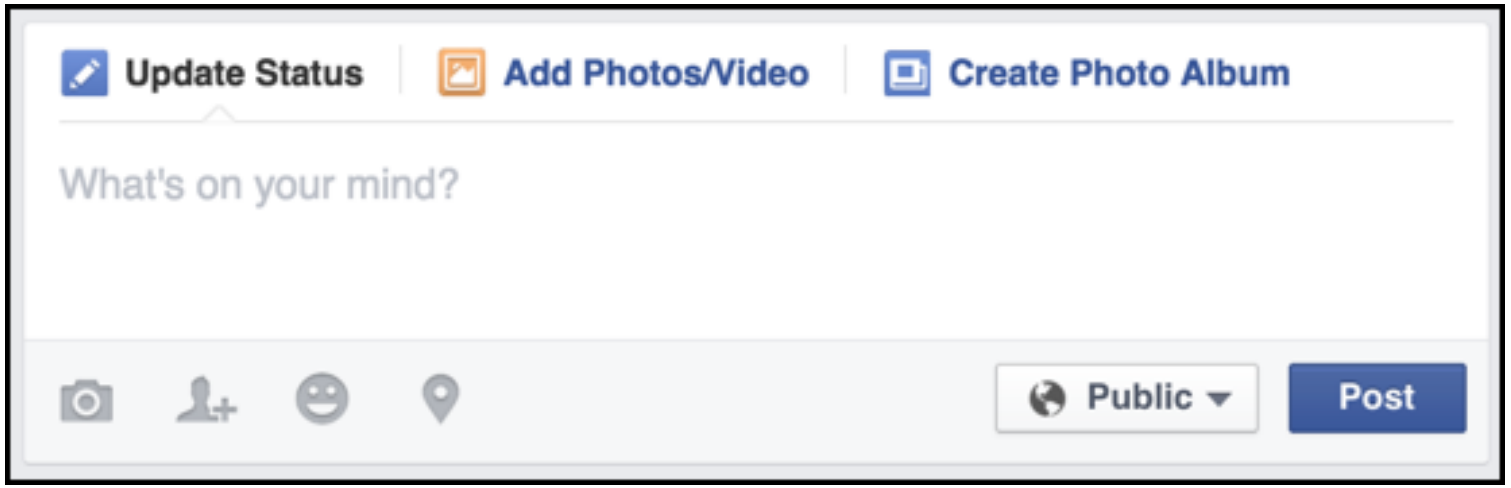
Motivation

- 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.

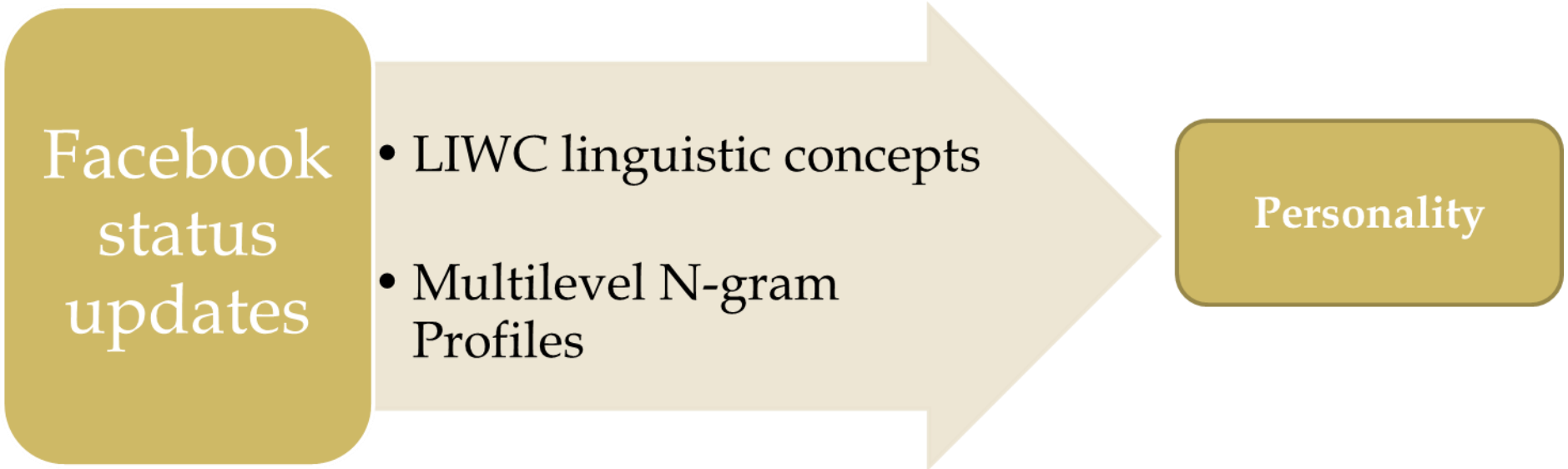
Status updates allow users to share their thoughts, feelings, and activities with friends, who have the opportunity to “like” and comment in return.

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

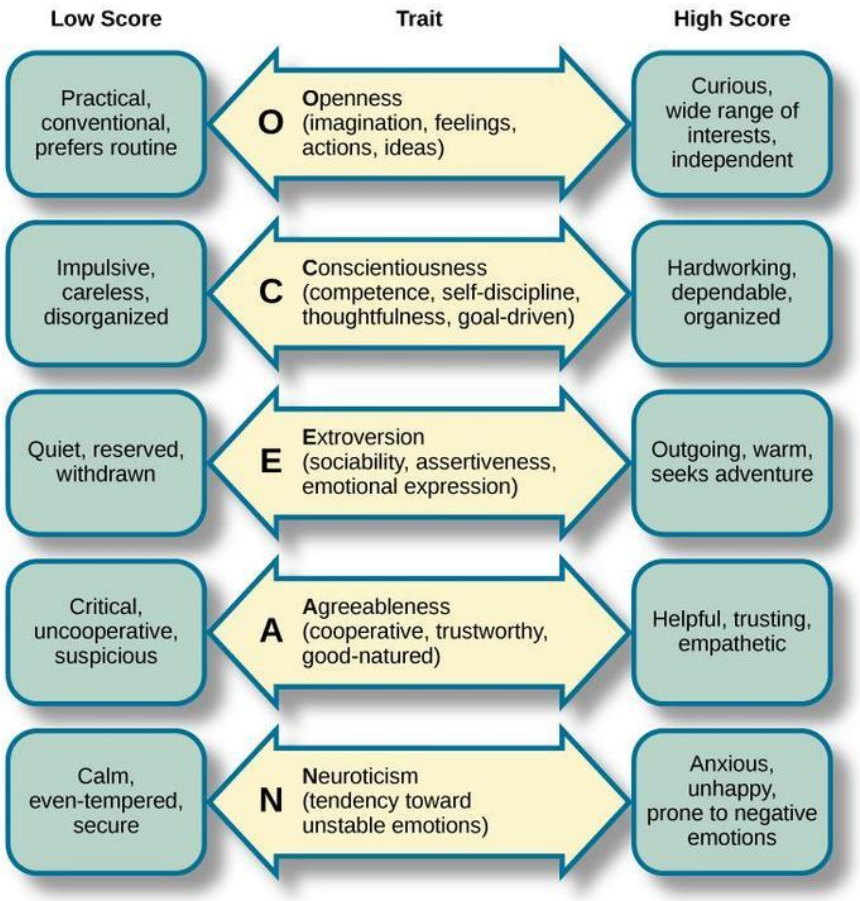
Previous Research

- Personality characteristics correlation with language features in FB, blogs, writings and conversation (not using our feature sets)
  - Bachrach, Y. et al. (2012). Personality and Patterns of Facebook Usage. WEBSCI'12 – ACM WEB SCIENCE CONFERENCE 2012, 22 June - 24 June 2012, Evanston, IL.
  - Noecker, J., Ryan, M., & Juola, P. (2013). Psychological profiling through textual analysis. *Literary and Linguistic Computing*. doi: 10.1093/llc/fqs070
  - Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE, 8(9), e73791. http://doi.org/10.1371/journal.pone.0073791
  - Solinger, C., Hirshfield, L., Hirshfield, S., Friendman, R., & Leper, C. (2014). Beyond Facebook Personality Prediction. In G. Meiselwitz (Ed.), Social Computing and Social Media: 6th International Conference, SCSM 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014. Proceedings (pp. 486-493). Cham: Springer International Publishing.
  - Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036-1040. doi: 10.1073/pnas.1418680112

Our contribution



Personality: The big five factor model (FFM)

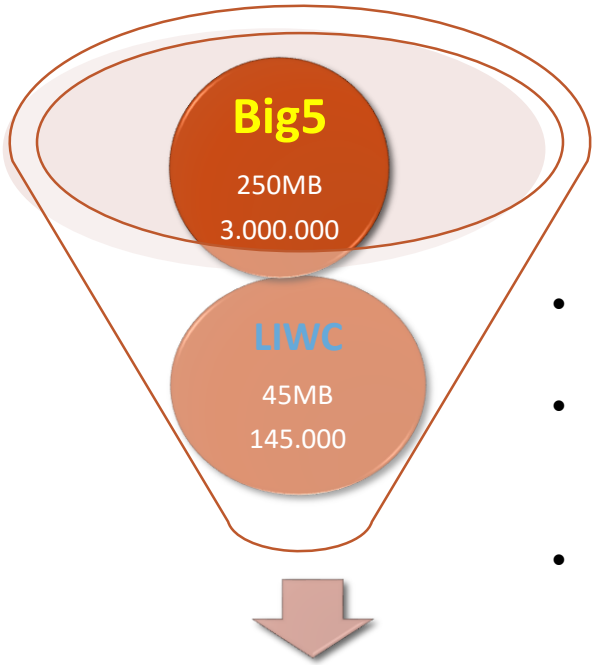


MyPersonality Project Data

(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**



7000 users

- **64 GB RAM**
- **14 days for final sample**
- **1-2 days needed for each experiment**

Stylometric features

1. LIWC

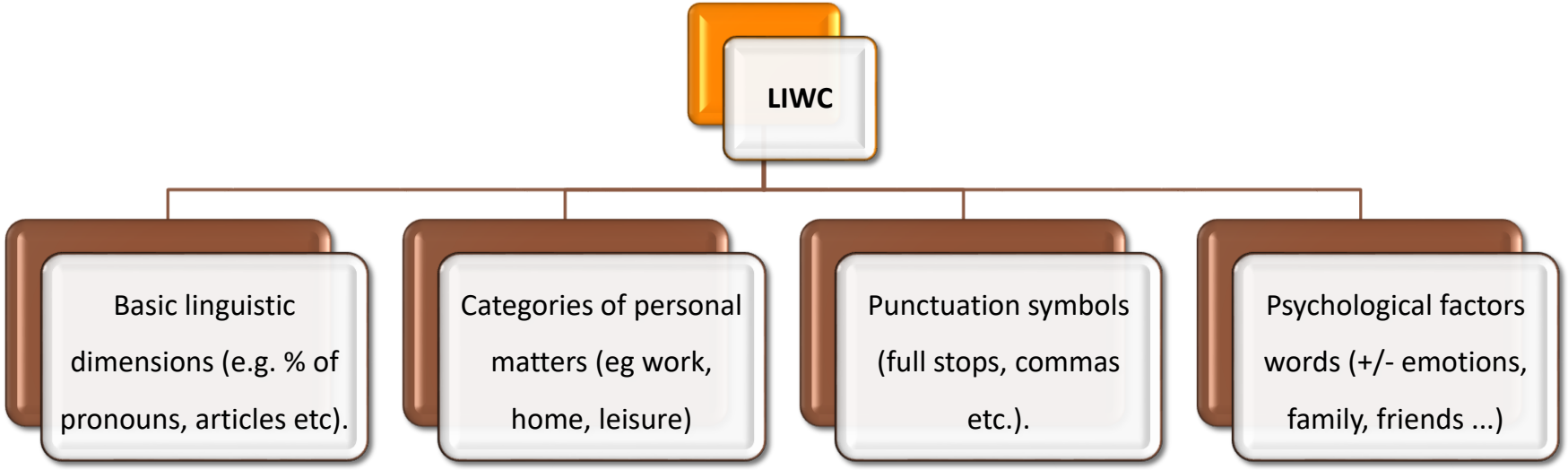
- Dictionary containing words that belong to 64 language and psychological processes
- Language-dependent, semantics, no Greek

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



Sample LIWC Features

LIWC (Linguistic Inquiry and Word Count)

Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX

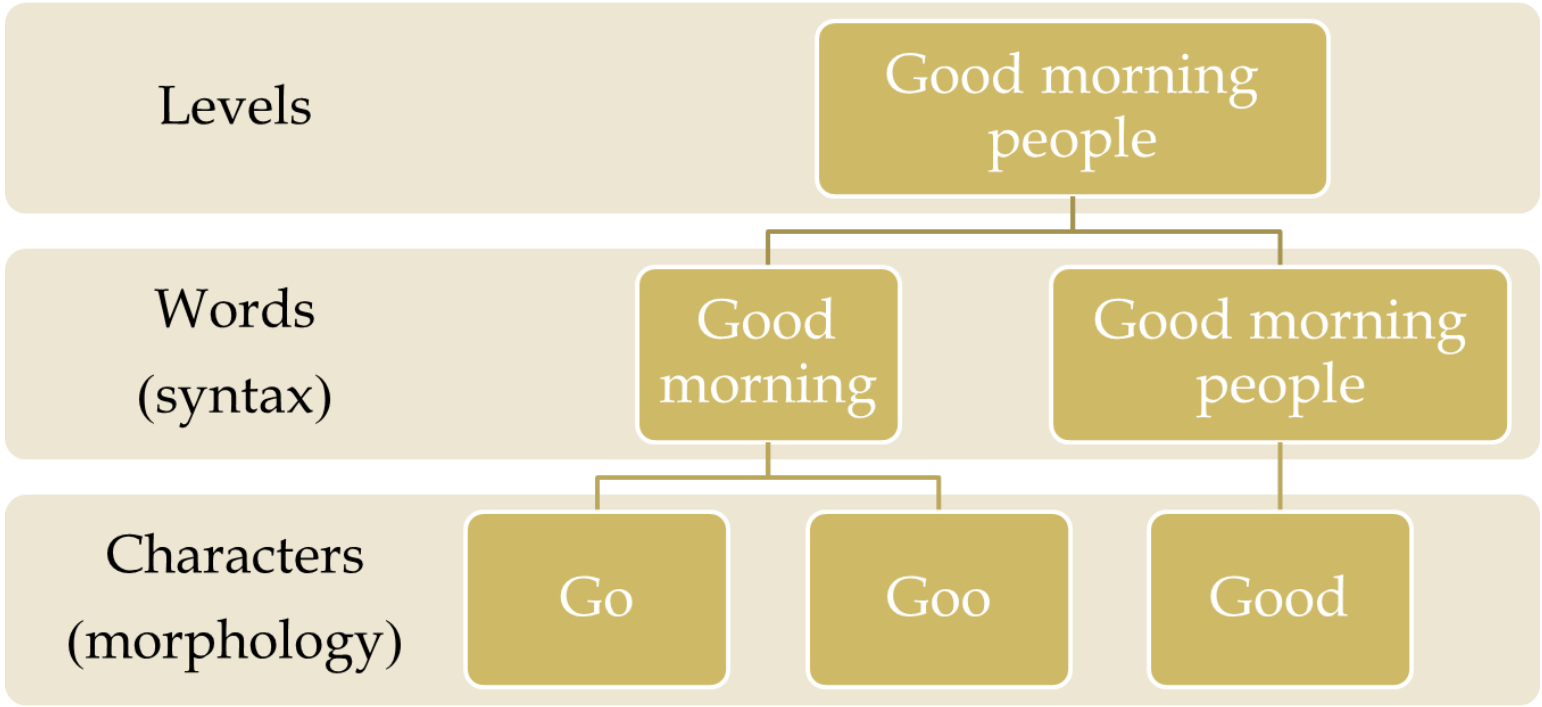
Feature	Type	Example
Anger words	LIWC	hate, kill, pissed
Metaphysical issues	LIWC	God, heaven, coffin
Physical state/function	LIWC	ache, breast, sleep
Inclusive words	LIWC	with, and, include
Social processes	LIWC	talk, us, friend
Family members	LIWC	mom, brother, cousin
Past tense verbs	LIWC	walked,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

Sample LIWC output

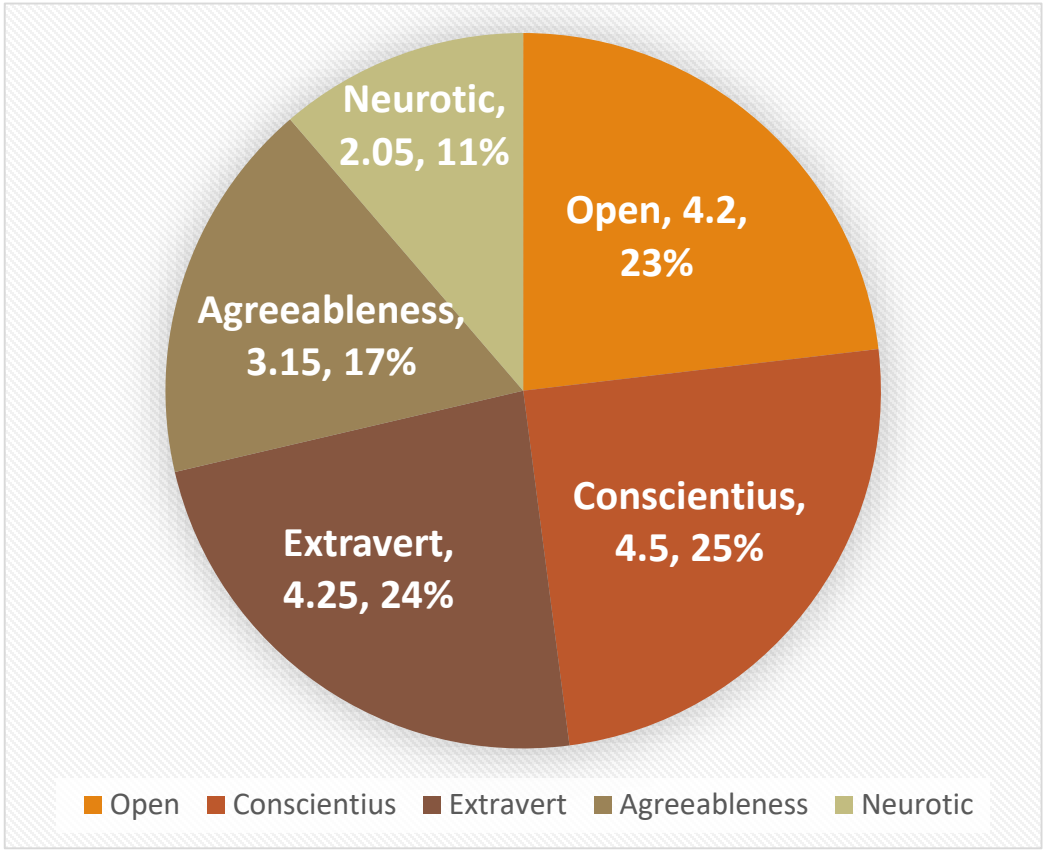
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Filename	Seg	WC	WPS	Sixltr	Dic	funct	pronoun	ppron	i	we	you	shehe	they	ipron	article	verb	auxverl
2	LIWC60Test_001.txt	1	19	19	21.05	94.74	68.4	10.53	0	0	0	0	0	0	10.53	10.53	15.79	10.53
3	LIWC60Test_002.txt	1	36	36	30.56	97.22	63.9	5.56	2.78	0	0	0	2.78	0	2.78	2.78	13.89	13.89
4	LIWC60Test_003.txt	1	34	34	23.53	88.24	61.8	5.88	2.94	0	0	0	0	2.94	2.94	5.88	11.76	8.82
5	LIWC60Test_004.txt	1	91	11.38	23.08	81.32	63.7	13.19	8.79	1.1	0	7.69	0	0	4.4	6.59	10.99	8.79
6	LIWC60Test_005.txt	1	118	14.75	11.02	87.29	65.3	22.03	8.47	2.54	4.24	0.85	0	0.85	13.56	5.93	16.95	16.1
7	LIWC60Test_006.txt	1	21	21	28.57	90.48	52.4	9.52	9.52	9.52	0	0	0	0	0	0	23.81	19.05
8	LIWC60Test_007.txt	1	53	26.5	26.42	90.57	41.5	7.55	3.77	1.89	0	1.89	0	0	3.77	5.66	18.87	5.66
9	LIWC60Test_008.txt	1	100	16.67	26	81	45	20	12	2	9	0	0	1	8	2	9	3
10	LIWC60Test_009.txt	1	8	8	12.5	75	37.5	25	25	12.5	0	12.5	0	0	0	0	12.5	0
11	LIWC60Test_010.txt	1	59	29.5	18.64	79.66	55.9	6.78	3.39	0	1.69	0	0	1.69	3.39	16.95	10.17	6.78
12	LIWC60Test_011.txt	1	121	30.25	21.49	90.91	51.2	4.96	1.65	0	0	1.65	0	0	3.31	9.09	13.22	8.26
13	LIWC60Test_012.txt	1	46	15.33	30.43	76.09	45.7	13.04	0	0	0	0	0	0	13.04	2.17	10.87	6.52
14	LIWC60Test_013.txt	1	166	13.83	25.9	74.1	47	11.45	7.83	4.82	1.2	0	1.2	0.6	3.61	4.82	10.24	6.02
15	LIWC60Test_014.txt	1	79	26.33	37.97	82.28	45.6	5.06	3.8	2.53	0	1.27	0	0	1.27	6.33	5.06	5.06
16	LIWC60Test_015.txt	1	35	17.5	20	91.43	48.6	14.29	11.43	5.71	2.86	2.86	0	0	2.86	2.86	14.29	2.86
17	LIWC60Test_016.txt	1	78	19.5	30.77	83.33	46.2	5.13	3.85	2.56	1.28	0	0	0	1.28	7.69	7.69	6.41
18	LIWC60Test_017.txt	1	40	20	37.5	82.5	45	5	0	0	0	0	0	0	5	10	7.5	7.5
19	LIWC60Test_018.txt	1	64	21.33	20.31	87.5	54.7	9.38	4.69	1.56	1.56	0	0	1.56	4.69	7.81	14.06	9.38
20	LIWC60Test_019.txt	1	22	11	27.27	86.36	59.1	9.09	0	0	0	0	0	0	9.09	9.09	18.18	18.18
21	LIWC60Test_020.txt	1	52	17.33	15.38	94.23	69.2	15.38	7.69	0	5.77	1.92	0	0	7.69	3.85	23.08	17.31
22	LIWC60Test_021.txt	1	76	12.67	26.32	89.47	60.5	14.47	9.21	1.32	7.89	0	0	0	5.26	3.95	17.11	13.16
23	LIWC60Test_022.txt	1	142	20.29	23.94	85.92	57.8	14.08	7.75	1.41	5.63	0.7	0	0	6.34	4.93	15.49	11.97
24	LIWC60Test_023.txt	1	82	27.33	15.85	82.93	61	10.98	7.32	0	7.32	0	0	0	3.66	12.2	7.32	7.32
25	LIWC60Test_024.txt	1	48	16	14.58	95.83	50	12.5	4.17	2.08	0	2.08	0	0	8.33	10.42	10.42	6.25
26	LIWC60Test_025.txt	1	117	23.4	18.8	78.63	51.3	9.4	3.42	0.85	0.85	0	1.71	0	5.98	11.11	4.27	2.56
27	LIWC60Test_026.txt	1	7	7	28.57	71.43	57.1	14.29	14.29	0	14.29	0	0	0	0	14.29	28.57	14.29
28	LIWC60Test_027.txt	1	32	16	25	71.88	43.8	12.5	6.25	3.12	3.12	0	0	0	6.25	3.12	15.62	15.62

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)

In accordance with previous research

- use many verbs of past and present.
- use cognitive verbs *(know, ought*)
- words with negative emotional content(hurt, ugly)

In contrast with previous research

- Frequent self-references (use 1st person pronouns)

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 – 2nd 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

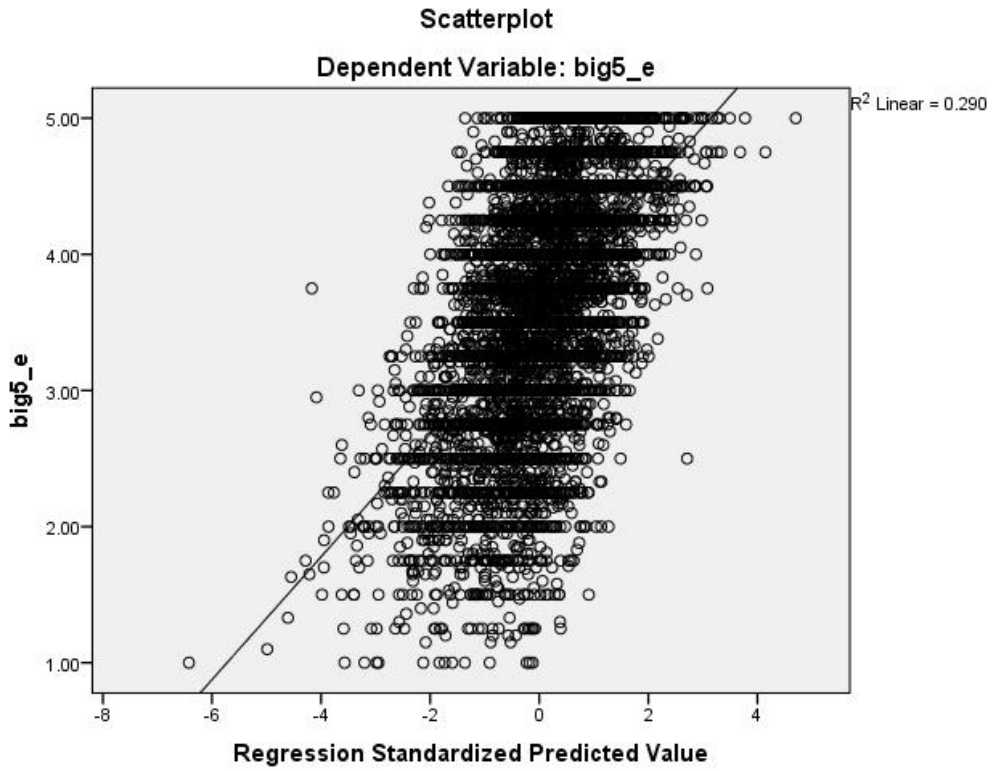
Top-2 ngrams

tm

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

TH

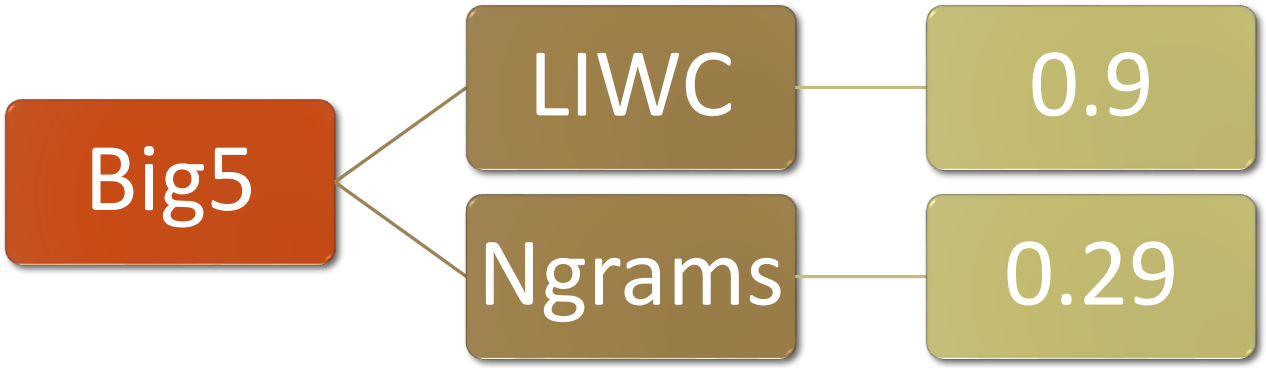
- HAPPY BIRTHDAY JOHN LENNON
- I LOVE THE RAIN!!
- CONGRATULATIONS TO THE WAUTOMA FOOTBALL TEAM



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!