

WHAT CAN OSS MAILING LISTS TELL US? A PRELIMINARY PSYCHOMETRIC TEXT ANALYSIS OF THE APACHE DEVELOPER MAILING LIST

Peter C. Rigby, Ahmed E. Hassan

PRESENTER: NUR ATHIRAH

AIM

- To evaluate the appropriateness of LIWC tool for information extraction from mailing lists

RESEARCH QUESTIONS

- *RQ1: What is the personality type of OSS developers?*
- *RQ2: Does the language and attitude of a developer change as he or she moves from being a new, to a current, to a departing developer?*
- *RQ3: Does word usage change around the time of a release?*

WHAT IS LIWC?

- Linguistic Inquiry and Word Count (LIWC)
- Based on counting function words or particles
- Cannot catch sarcasm, irony, or a given contextual meaning
- Dictionary that divides the different counts into meaningful dimensions
- Uses a psychometrically-based dictionary

WHAT IS LIWC?

3 top level categories used:

1. *Language composition*
2. *Psychological processes*
3. *Relativity*

Dimension	Example Words	Explanation
Pronouns	'our', 'I'	Subject or object in a sentence.
Negation of other words	'no', 'not'	Words that negate other words.
Insight	'think', 'know'	Words that indicate understanding about a topic.
Inhibition	'block', 'constrain'	Words that show the individual is restrained.
Tentative	'maybe', 'guess'	Words that indicate that the speaker is uncertain.
Causation	'because', 'effect'	Words that indicate why something happened
Social Processes	'talk', 'us'	Words that are related to social interaction.
Positive and Negative Emotion	'good', 'hate'	Words related to the emotion of the speaker.
Optimism and energy	'pride', 'win'	Subdimension of positive emotion.
Past, Present, and Future	'walked', 'walk', 'will'	The tense of the verb.
Reference to Time	'hour', 'day'	Words related to time.
Inclusive and Exclusive	'with', 'except'	Words that include or exclude others and their ideas.

Table 1. Selected LIWC dimensions

METHOD

- Need mailing list data
- The Apache httpd server project
 - A large, successful, mature OSS project
- Extracted the body of all email messages (1995 ~2005)
 - Approx. 104,000 emails
- Select four developers that had the most commits for a given two year period

EXPERIMENTS

- 1) PERSONALITY
- 2) NEW, CURRENT AND DEPARTING DEVELOPERS
- 3) RELEASES

1) PERSONALITY

What is the personality type of OSS developers?

- Certain LIWC dimensions were correlated to the big five personality
- The big five personality factors:
 - Neuroticism
 - Express negative emotions such as anger
 - Extroversion
 - Seek the company of others
 - Openness
 - Understand and respect unusual ideas
 - Agreeableness
 - Be cooperative
 - Conscientiousness
 - Be self-disciplined and diligent

1) PERSONALITY

- Hypothesize:
 - Top committers will have a similar personality type to each other
 - The personality of top committers will differ from that of the entire mailing list
- Top committers will be:
 - Less neurotic
 - More extroverted
 - More open
 - More agreeable
 - More conscientious
- E.g. successful leaders will be more open to others' ideas

1) PERSONALITY

Subject	Neuro	Extro	Open	Agree	Consc
Num. of Clusters	1	3	3	2	3
A	8.10 ¹	2.15 ^{1,2}	22.47 ¹	-9.20 ²	-4.07 ^{2,3}
B	8.79 ¹	2.61 ²	19.30 ¹	-10.46 ¹	-4.24 ²
C	7.35 ¹	5.95 ³	32.84 ²	-10.58 ¹	-4.34 ²
D	7.86 ¹	0.53 ¹	44.18 ³	-11.12 ¹	-5.36 ¹
≤ 30	7.07 ¹	6.41 ³	32.87 ²	-11.15 ¹	-3.42 ³
> 30	7.33 ¹	7.61 ³	32.85 ²	-10.87 ¹	-4.24 ²

less
extroverted,
less open, but
equally
conscientious

Table 2. Composite measure of Personality.
Values are relative, not absolute. Post Hoc
Tukey's HSD Test ($\alpha = 0.05$). Superscript rep-
resents Tukey cluster number.

2) NEW, CURRENT AND DEPARTING DEVELOPERS

Does the language and attitude of a developer change as he or she moves from being a new, to a current, to a departing developer?

- Examine how developers change throughout their time with the project
- Two top committers who have left the project
 - compare the word usage of first and last 50 emails

2) NEW, CURRENT AND DEPARTING DEVELOPERS

Dimension	Subdimension	Dev A	Dev B
Standard WC Pronouns	Total Pronouns	-3.4	-1.9
	We	-1.0	-1.7
	Self	*	-2.6
	You	*	1.8
	Other people	-2.3	*
Standard WC	Negation	1.1	2.3
Emotion	Positive	*	-2.2
Cognitive	Cause	*	1.0
	Insight	*	-1.4
	Inhibit	*	-0.4
Social Processes		-3.23	*
Relative	Time	*	-2.4
	Present	*	-2.0
	Past	3.4	*
	Future	*	1.0
	Inclusive	-3.0	*
	Exclusive	*	1.9

Table 3. Mean differences for the first and last 50 email messages. (* $p > 0.05$, otherwise $p \leq 0.05$)

3) RELEASES

Does word usage change around the time of a release?

- A consistent change in word usage before and after a release
- Examine the Apache 1.3 and 2.0 releases

3) RELEASES

Dimension	-1.3	-2.0
Optimism	-0.37	*
Tentative	-1.3	*
References to Time	1.1	*
Future tense verbs	-0.7	*
Social Processes	*	0.74
Inclusive	*	-0.64

Table 4. Mean differences for Apache 1.3 and 2.0 releases. (* $p > 0.05$, otherwise $p \leq 0.05$)

LIMITATIONS

- 1) The LIWC was created as tool to measure various psychological phenomenon
- 2) The LIWC may be biased against individuals who's first language is not English
- 3) Apache httpd server may not be representative of OSS development in general
- 4) The three experiments were preliminary
 - Further experimentation and validation are needed

CONCLUSIONS

- First attempt at using a psychometrically-based word count tool to understand OSS mailing lists
- Used the LIWC dictionary
- Conducted three preliminary experiments on the Apache httpd server's developer mailing list
 - Proposed directions for future research

CONCLUSIONS

1) Personality

- Created a baseline personality score for the entire mailing list
 - Compared the top four committers to the baseline and to each other
- The two developers that were responsible for two major Apache releases
 - Had similar personalities
 - Had different personalities from the baseline and other developers
- Future work:
 - Run further comparisons with other projects
 - Examine the effect of a developer's role on his or her personality

CONCLUSIONS

2) New, Current, and Departing Developers.

- Two top committers who have left the Apache project
 - Extracted the LIWC dimensions for their first and last 50 emails
- The statistically significant LIWC dimensions are consistent
- Dividing developers' emails into:
 1. Equal groups
 - Span multiple events
 - Produce statistically nonsignificant results
 2. Event based periods
 - Produce more interesting results
- Future work:
 - Characteristics of past successful and unsuccessful developers
 - Predict whether a newcomer will become a successful member of the development team

CONCLUSIONS

3) Releases

- Calculated one month before and one month after the releases of Apache 1.3 and 2.0
- 1.3 release
 - Had higher pre-release levels of optimism
- 2.0 release
 - Had higher post-release levels of social words
- No common statistically significant LIWC dimensions
- Did not take into account the success of each release
- Future work:
 - Assess the “attitude” of developers around successful and failed releases

CONCLUSIONS

- Gained some insight:
 - Type of people who participate
 - Discussions that occur on the Apache mailing list
- Attempted to:
 - Determine the personality types of four top developers
 - Understand why developers join and leave a project
 - Examine the general attitude of developers before and after a release
- Show promise for future work

EXTRA

LIWC

1. I sent these patches in a couple weeks ago. They may be of interest, maybe not.
2. This patch fixes the important half of the bug.

- First quotation, the author is uncertain
- Second quotation, the author seems confident
- Words like
 - 'maybe' → tentativeness
 - 'important' → certainty
- only counts words that are contained in its dictionary
- grouped into basic linguistic and psychometric dimensions
 - each word belonging to one or more dimension

1) PERSONALITY

Subject	Neuro	Extro	Open	Agree	Consc
Num. of Clusters	1	3	3	2	3
A	8.10 ¹	2.15 ^{1,2}	22.47 ¹	-9.20 ²	-4.07 ^{2,3}
B	8.79 ¹	2.61 ²	19.30 ¹	-10.46 ¹	-4.24 ²
C	7.35 ¹	5.95 ³	32.84 ²	-10.58 ¹	-4.34 ²
D	7.86 ¹	0.53 ¹	44.18 ³	-11.12 ¹	-5.36 ¹
≤ 30	7.07 ¹	6.41 ³	32.87 ²	-11.15 ¹	-3.42 ³
> 30	7.33 ¹	7.61 ³	32.85 ²	-10.87 ¹	-4.24 ²

Table 2. Composite measure of Personality. Values are relative, not absolute. Post Hoc Tukey's HSD Test ($\alpha = 0.05$). Superscript represents Tukey cluster number.