Dude, srsly?: The *Surprisingly* Formal Nature of Twitter's Language

Yuheng Hu, Kartik Talamadupula, Subbarao Kambhampati @ Arizona State University
Contact: yuhenghu@asu.edu

Motivation

Twitter houses many features that make its language distinct

- It has140 characters limit, so its language is highly compact and brief.
- It encourages discussion on a wide variety of topics and information (e.g., news events).
- Users can build follow other users' tweets and re-post/edit the content of others' tweets too.

A continuing debate on positioning Twitter's linguistics in the spectrum of

- One end: well established "casual communication mediums" like SMS and chat
- Another end: more formal mediums like emails, blogs, magazines and newspapers

Research question in this paper:

- *"Is the language of Twitter closer to informal media such as SMS and IM, or does it share similarities with more curated media like newspapers and magazines?"*

Value of this study:

- importance to various applications in anthropology, communication studies, sociology and many sub-areas within computer science – text mining, computational linguistics, and machine translation

Dataset

- Tweets in Oct 2012
- SMS, collected from 2010 to 2012
- Online chat from 2006
- Email, from Enron dataset
- Blogs, from ICWSM 2011 Spinn3r dataset
- Magazine, from the Open American national corpus (OANC)
- News, from Reuter news

	#Docs	ShortWords (per doc)	Length (by word)	Length (by chars.)
Twitter	46,480,800	7.60	12.21	53.74
SMS	51,654	8.08	10.88	40.65
Chat	10,567	2.56	3.81	18.72
Email	244,626	137.68	255.04	1306.34
Blog	24,004	147.96	269.75	1323.65
Mag.	186,020	382.43	682.09	3274.28
News	10,788	73.83	129.41	619.32

Methods

We conduct our investigation by analyzing a variety of corpora from two aspects:

Quantifying Linguistic Style

The style of a language can be evaluated from two different perspectives: *Orthographic* and *Grammatical* (Wardhaugh 2011):

#feature	Description
Word Frequency (WF)	Measure the difficulty and readability of words, sentences and documents
Lexical Density (LD)	Measure the proportion of the lexical words over the total words. Lexical words are mostly made up of verbs, nouns, adjectives and adverbs
Personal Pronouns (PP)	Measure usage of personal pronouns (first-, second-, third person)
Intensifier (INT)	Measure usage of intensifiers (very, really, …). Used 25 the most commonly used intensifiers
Temporal Reference (TR)	References to the future (going to, gonna, will…) Used to investigate whether a medium is more about the present or the future

Psycholinguistic Analysis

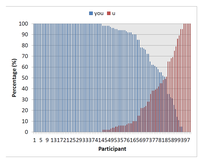
We also investigate whether there are underlying cognitive and affective aspects that differentiate Twitter from the other mediums.

We develop SocLin, a matrix factorization framework for psycholinguistic analysis:

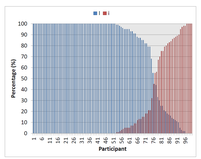
- it factorizes a term-doc matrix (X) into two major factors corresponding to term-aspects (T) and document-aspects (D). We use LIWC as a supervision on T during training.

$$\begin{aligned} \min_{\mathbf{T},\mathbf{G}} \quad & \mathcal{J} = \left\| \mathbf{X} - \mathbf{TSD}^\top \right\|_F^2 \\ & + \alpha Tr\left((\mathbf{T} - \mathbf{T}_0)^\top \mathbf{\Lambda} (\mathbf{T} - \mathbf{T}_0) \right) \\ s.t. \quad & \mathbf{T} \geqslant 0, \mathbf{S} \geqslant 0, \mathbf{D} \geqslant 0 \end{aligned}$$

Results



(a) usage of *you* vs. *u*



(b) usage of *I* vs. *i*

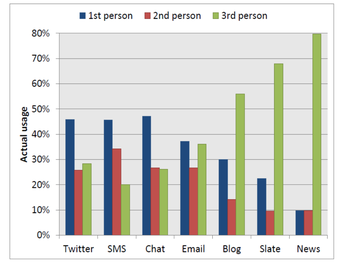
Figure 1: *Usage of Singular Pronouns*

	Tw.	SMS	Chat	Email	Blog	Slate	News
WF	4.64	5.52	5.98	4.54	4.33	2.87	2.63
LD	0.47	0.42	0.40	0.44	0.54	0.48	0.58

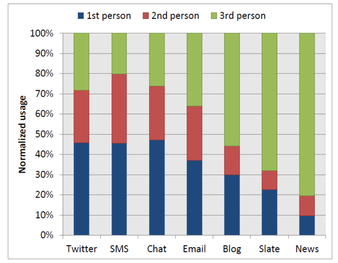
Analysis	Results
WF	Similar to email and blog language; more sophisticated than SMS and chat.
LD	Close to blogs and news; tweets are used primarily to convey information, but are restricted by length.
PP	Mostly 1st and 3rd person, very distinct from other mediums; Tweets are about *self* as well as information sharing (e.g., breaking news). Much less conversational than SMS and chat.
INT	More usage of *really*, indicating a younger population of users than traditional mediums like email and news where *very* is mostly used. Higher net intensifier usage than chat.
TR	Highest number of references to the present; most content related to current events (real-time platform).
AA	Contains significantly more positive emotion than negative. Displays a much lesser variation of affect when compared to email, blogs, magazines and news.
CA	Contains less cognitive words than email and news. Contains strong opinions (more words on *certainty*) and lesser *tentativeness* than SMS and chat, meaning more information sharing than discourse.

Table 4: *Summary of Results for Twitter; WF: Word Frequency; LD: Lexical Density; PP: Personal Pronouns; INT: Intensifiers; TR: Temporal References; AA: Affective Aspects; CA: Cognitive Aspects*

Results

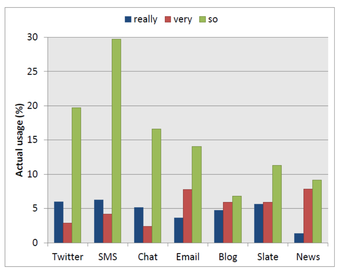


(a) Personal Pronoun Usage

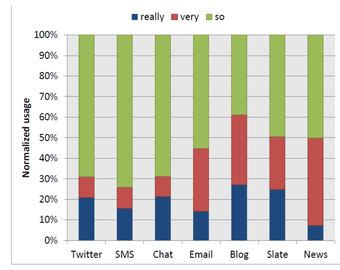


(b) Normalized Usage

Figure 2: *Usage of Personal Pronouns*

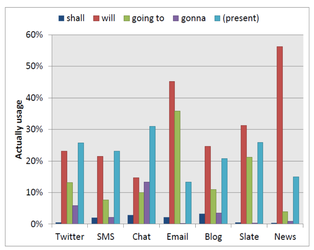


(a) Intensifier Usage

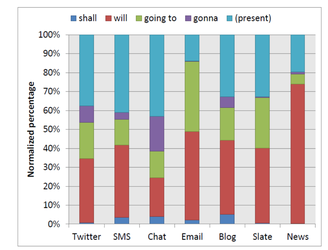


(b) Normalized Usage

Figure 3: *Usage of Intensifiers*



(a) Temporal Reference Usage



(b) Normalized Usage

Figure 4: *Usage of Temporal References*

Conclusions

- We proposed a two-part computational framework to offer insights into linguistic styles on Twitter, and other popular mediums.
- We concluded that the language of Twitter is highly dynamic, and that depending on the measure that is used, it shows similarities to different media. We believe that this proves – more than anything else – the fact that Twitter is a rich, evolving medium whose language is a projection of the language of more formal media like news and blogs into a space restricted by size, leading to adaptations that endow Twitter with characteristics that are similar to short media like SMS and chat as well.