Differences in Language and Style Between Two Social Media Communities

Cécile Paris, Paul Thomas, and Stephen Wan
CSIRO ICT Centre
PO Box 76, Epping, NSW 1710, Australia
{FirstName.LastName}@csiro.au

Abstract

Microblogs are increasingly used as communication channels for organisations and their related communities. In this work, we are interested in the effect of community on the resulting microblog language use. We analyse content from Twitter, examining tweets relating to two government organisations one conducting scientific research, the other providing social services. We find that the two different communities have significant differences in style and language use, observing marked differences in formality and tone as measured by properties such as pronominal usage, orthographic convention, and use of Twitter features. We posit that these differences arise due to underlying differences in the communication goals of the two user groups. Tools working with Twitter, to extract and represent information, may therefore need different approaches in different domains.

Social Media Monitoring for Government Applications

Social media platforms such as Twitter have rapidly become a major communication tool for a variety of communities, topics, and purposes. As a result, people and organisations turn to social media both to engage with the public and to find out what people say and what is happening. In a government context, agencies are interested in improving their services and communications by listening to and engaging with social media. However, given the volume of data on social media, and limits on government resources, it is desirable to use tools to support tasks where possible.

Different government agencies interact with and give rise to different communities on social media. The result is that each is likely to be discussed amongst different cohorts of interested citizens, and in different ways.

We ask: are these different styles of discussion visible in social media and, if so, how? what does this mean for the design and development of social media monitoring tools?

We approach these questions by comparing features of two data sets: one collects mentions of the Australian national science agency, the other collecting mentions of the agency responsible for the delivery of social services in Australia. We find that the observed frequencies of various linguistic features differ between the two collections. Contrary to early suggestions, we find that Twitter is neither homogenously "conversation-like" nor "written-like" in style. That is, Twitter content can differ in formality depending on the community and underlying communication type.

Related Work

A number of methods have been suggested for natural language processing (NLP) tasks on microblogs, such as sentiment analysis, topic modelling, discussion thread structure analysis, classification to find specific posts to assist crisis management and summarisation.

Our analysis differs in that we wish to explore stylistic differences between communities of social media users within the same type of microblog, specifically Twitter. As such, our work is more closely related to work in URL type classification and web genre identification—e.g., see Kan and Thi (2005) and Lindemann and Littig (2007). Our analysis is based on a number of features that stem from the literature, and we outline here a few examples of work that focuses on specific linguistic features and their underlying linguistic interpretation of communication.

One position in the literature argues that online language, such as email and blogs, can be characterised as being more or less speech-like in nature, and hence more informal, e.g., (Nowson, 2005). This informal language use can include lexical and grammatical differences, humour, misspellings and colloquial language. Our work examines

(a) Sample Tweets mentioning CSIRO, the science agency

Meddling with food genomes is never safe, even when genes are suppressed like in CSIRO's #GMO wheat experiments [link]

CSIRO Researchers develop paint on solar cells [link] #science #climatechange #CSIRO #solar

On air: CSIRO's Leo Joseph talking birds and dinosaurs and evolution with @LouiseVMaher (book "Stray Feathers")

(b) Sample Tweets mentioning Centrelink, the social service agency

centrelink drive me nuts....i just want to record my earnings

Finally opened today's mail. Letter from Centrelink. Carer's bonus is coming!! Yippee!!

F*** you Centrelink. F*** you very much. I do not need to deal with your epic b***** incompetence today

Figure 1. Sample Tweets from our collections.

this further, looking at whether this informal language always occurs.

We follow the work of Herring (2007) which argues against treating online text as being homogenous in terms of data characteristics. Analysis of microblog language in Twitter samples have been conducted to identify categories of content – for example, "conversational", "pass along", etc. (for an overview, see Dann (2010)). Our work differs in that we investigate differences at a community level.

We also suggest that understanding stylistic differences can be beneficial in the development of NLP tools. One example of this is by Foster (2010), who examines the difference between blogs and the formal written style of Wall Street Journal (WSJ) articles.

Gathering Tweets

Our data covers online mentions of two government agencies. Both are nationwide, and often in the public eye; however, they have very different activities and typically interact with different groups of people.

Science—The *science* agency, CSIRO, is active nationwide on a broad range of research projects. It has a high profile in the country's research and government sectors and makes a serious effort to engage with the general community.

Social service—Centrelink, the *social service* agency, administers a wide range of schemes including income support, training and apprenticeships, and sickness benefits for a large number of citizens. The agency is extremely visible, is active on social media, and encourages its communications staff to become involved in online discussions where appropriate.

We gathered tweets for July, August, and September 2011 from three sources: Twitter's streaming API, SocialMention, and Google Alerts. The latter two were used to ensure relevant tweets had not been missed. Over the three months we collected 15,471 unique tweets, which we believe is a complete collection of those tweets explicitly mentioning the two agencies. Examples of tweets from our two collections are shown in Figure 1.

General Characteristics

Table 1 summarises the general characteristics of our collections. Over the three months, we have 6810 tweets in the science collection and 8661 in the social service collection, which represents on average about 74 and 94 per day. Users are very nearly completely partitioned. Of the 7761 users publishing tweets across the two domains, only 207, or 2.7%, are represented in both sets. This clearly indicates that the two agencies interact with different communities in the public arena.

Users in the science collection make much greater use of Twitter hashtags to label their posts. The frequency of @mentions, a Twitter feature for identifying other users, also differs: 71% of tweets in the science corpus *vs* 56% of tweets in the social service corpus. We note that many more messages in the science collection include a link.

The timing of tweets also differs in the two collections. The volume of tweets in the science collection depends on the time of day, with peaks during the day and during the working week, but this is much less evident in the social service collection. One explanation for this is that the science tweets are work-related, while those about the social service agency relate to personal experiences.

Language Use

The two collections comprise tweets by different users, online at different times, and talking about different agencies. We thus expect them to have different communicative goals, suggesting that there should be differences in the language used. We consider linguistic features in two classes: variations of English, including spelling, and differences in emotive and personal language, including pronouns, interrogatives, and exclamations. Across the board, we find more non-standard English,

	Science	Social service
Tweets (92 days)	6810	8661
Mean tweets/day	74	94
Unique users	3603	4365
Mean len (chars)	121 ***	95
Mean no. words	10.5	12.4
#hashtags	51% ***	17%
@mention	71% ***	56%
Web link	61% ***	12%

Table 1. General characteristics of tweets for our two agencies, over the three months June September 2011. "***" indicates differences significant at $p<0.005$ ($\chi 2$ test, except "mean len" Mann Whitney U test).

	Science	Social service
Contractions	23%	32% ***
…without apostrophe	0.5%	0.1% ***
"u" as a word	0.8%	2% ***
"r" as a word	0.6% *	0.4%
"k" as a word	0.1%	<0.1%
"y" as a word	0.1%	0.3% ***
"b" as a word	0.5%	0.3%

Table 2. Difference in lexical features in the two data sets. "*" indicates differences significant at $p<0.05$, "***" at $p<0.005$ (χ^2 test).

more emotive language, and more personal language in the social service collection. We also find more evidence of curatorial practice in the science collection. We look at these features in detail below.

Variation in Lexical Conventions Tweets in the social service corpus are 40% more likely to contain contractions (such as "can't"), contractions with missing apostrophes (such as "cant" and "didnt") or abbreviations (e.g., "u" for "you" or "k" for "ok") (Table 2)[1].

Emotive and Personal Language The two collections also show striking differences in tone, and the degree to which posts describe personal experiences or opinions. Table 3 summarises these differences.

The social service collection has more instances of exclamations and questions, and more non-standard strings of exclamation and question marks. The abundance of exclamations suggests the messages are more likely to be strongly emotive. Emotions of one kind are also suggested by the number of messages containing any of several dozen swear words, and those all in upper case.

Pronouns too are much more prominent in tweets mentioning the social service agency: 49% of these include a first-person pronoun, and 71% include a pronoun of any

[1] The abundance of "r" as a word in the science corpus is partly due to the phrase "R&D". This is because we treated "&" as a word boundary.

	Science	Social service
Ends with "!"	3%	11% ***
Ends with "?"	4%	9% ***
Repeated "!", "?"	2%	7% ***
ALL CAPS	<0.1%	0.4% ***
all lowercase	0.4%	9.3% ***
Swearing	0.8%	8% ***
Demonstratives	12%	17% ***
First person	29%	49% ***
Second person	10%	23% ***
Third person	16%	29% ***
Any pronoun	45%	71% ***

Table 3. Features demonstrating emotive and personal language in the two data sets. "***" indicates differences significant at $p<0.005$ (χ^2 test).

kind. An informal inspection of these tweets indicates that mentions of the social service agency are likely to be in the context of a personal experience or a personal opinion, while posts in the science collections tend to pass on facts and information. In the social service collection, we also observe a higher proportion of tweets with a second person pronoun: we suggest this means that users mentioning the social service agency are more likely to be engaged in discussions with each other. Similarly, we found that the use of demonstratives differs in the two collections.

Differences in Style The conversations in our two collections vary in three aspects: formality, intent, and curatorial techniques.

The posts exhibit a marked difference in formality: Tweets in the science collection are more formal than those in the social service collection. This is borne out through a number of features: post length; lower use of contractions or of informal lexical variants for pronouns or verbs; the rarity of posts ending with question marks or exclamation marks, or of posts with repeated punctuation; the low occurrence of swear words; and the more conventional typographical features. Posts in the science collection employ more conventional language than the posts in the social service collection, making them both more formal and less speech- or conversational-like.

With respect to intent, the posts in the science collection do not often use personal pronouns. The lower number of occurrences of first person pronouns suggests that users in this collection do not use Twitter to explicitly state opinion as often as the users of the social service. The infrequent use of the second person pronoun indicates people are not as involved in discussions.

Users in this collection also make much greater use of Twitter hashtags to label their posts. We believe this points to a more careful use of the posts and a curatorial intent, where hashtags serve to direct a tweet to the right audience when author and readers do not know each other.

Finally, the posts in the science collection often include a link. This suggests that messages in the science corpus are more likely attempts to pass on information, as opposed to asking questions or discussing personal experiences.

In contrast, tweets in the social service collection use a more personal and emotive language, talk about experiences, and ask more questions.

To summarise, not all social media language is the same. In our collection, talk in the science arena is more formal, with more care taken to address an audience, and more concern with passing on non-personal information, while the tweets in the social service domain are more on a personal note, expressing emotions, describing experiences and asking more questions. This mirrors what could be the equivalent genres in other media: conversation on the one hand, scientific writing on the other (e.g. Biber, 1991; Conrad and Biber, 2001).

Implications for Monitoring Tools

We have described a number of differences between the posts of our two collections. These suggest that monitoring tools may be able to obtain different things from tweets in different communities. It also suggests the tools themselves may need to be built differently, to be adapted to the information available from, and the linguistic conventions of, each community.

Social media is generally considered as being difficult to process because of its use of non-conventional language—see, e.g., (Nowson, 2005). Our analysis reveals that not all uses of a single medium are equal, and it pays to study the language of the community under consideration to develop the appropriate tool. Tools can then be adapted to fit the genre under consideration. This is the focus of our on-going work with the science and the social services agencies

Conclusions

Different agencies deal with different programmes and interact with different communities. While social media has been characterised as being more or less speech-like in nature, and hence more informal, the language in social media in different communities includes significant differences in language use, as is the case in other media.

We collected and analysed posts from two communities: one concerned with science, one concerned in social services. There several significant differences. The communicative goals differ: on the one hand providing non-personal information, on the other engaging in discussion and sharing personal experiences and opinions. Language features also differ: in particular, the science set is more formal. Implied audiences also seem to differ, with much more use of second-person pronouns in the social science collection. Understanding these differences in goal and register helps us identify the types of tools that can be built for each community and the technical problems in each case.

In the present work, we have compared two communities. It is possible that these communities have their respective idiosyncracies; gathering data for similar communities (another technical agency, for example, or another agency with wide public exposure) would help confirm or refute the patterns we see here. We are also hoping to validate some of our findings by building a classifier to determine to which collection a post belongs.

Acknowledgements

This research has been partially funded under the CSIRO-Centrelink Human Services Delivery Research Alliance (HSDRA). We would like to thank James McHugh for his help in gathering and analysing the data.

References

Biber, D. 1991. *Variation across Speech and Writing*. Cambridge, UK: Cambridge University Press.

Conrad, S. and Biber, D. (eds.) 2001. *Variation in English: Multi Dimensional Studies.* Essex, UK: Pearson Education Limited.

Dann, S. 2010. Twitter Content Classification. *First Monday* 15(12).

Foster, J. 2010. "cba to check the spelling": Investigating Parser Performance on Discussion Forum Posts. In Proceedings of Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the ACL, 381 384. Los Angeles, Calif: Association for Computational Linguistics (ACL).

Herring, S. 2007. A Faceted Classification Scheme for Computer Mediated Discourse. *Language @ Internet*, article 761 Retrieved from http://www.languageatinternet.org/articles/2007/761

Kan, M. Y. and Thi, H. 2005. Fast webpage classification using URL features. In Proceedings of the 2005 ACM International Conference on Information and Knowledge Management (CIKM), 325 326. Bremen, Germany: Association for Computing Machinery (ACM).

Lindemann, C. and Littig, L. 2007. "Classifying web sites". In Proceedings of the International Conference on World Wide Web, 1143 1144. Banff, Canada: Association for Computing Machinery (ACM).

Nowson, S.; Oberlander, J. and Gill, A. J. 2005. Weblogs, Genres and Individual Differences. In Proceedings of the 27th Annual Conference of the Cognitive Science Society. 1666 1671. Stresa, Italy: Cognitive Science Society.