Social Media Sources for Personality Profiling

David N. Chin and William R. Wright
{chin, wrightwr}@hawaii.edu

University of Hawai'i at Mānoa
Department of Information and Computer Sciences
1680 East-West Road, POST 317, Honolulu, Hawai'i 96822 USA

Abstract. Social media provide a rich source of author-identified text that can be used for personality profiling. However, differences in length and number of entries, syntax, abbreviations, spelling and grammar errors, and topics can affect type and difficulty of preprocessing to extract appropriate text, accuracy of training, time period sampling for training texts, and rate of degradation of accuracy over time. Biases introduced by the topic areas of the social media, author self-selection, and current events affect different social media to varying extents and can bias both the demographics and possibly the personality types of users.

1 Introduction

Many researchers have been able to predict personality from written text [6, 1, 5, 3, 4, 8, 2, 9, 10]. Recent social media outlets provide a rich source for author-identified text that can be used for this purpose. However, the different social media outlets each have different characteristics that will likely affect their effectiveness for personality profiling. For example it is doubtful that any single personality classifier or choice of training features will provide the best results for all social media because of their many differences. This paper will catalog those characteristics of current social media outlets that are expected to affect personality profiling and their implications for personality prediction.

Characteristics of social media likely to affect personality profiling include:

- word length of entries
- number of entries/author
- author identification
- spelling and grammar errors
- topic bias
- time-period bias
- author self-selection bias
- legal access and privacy restrictions
- unusual syntax, usage, abbreviations

The product of the word length of entries and the number of entries per author yields the amount of text available for personality profiling. Relatively small amounts of text can be used to predict personality. For example, [9] found they could achieve more than 81% accuracy with more than 10 email messages. Although [9] do not specify the word length of a typical email message, one study [11] found an average of 3150 characters and a median of 1304 characters for

email messages. So 10 email messages would average about 31K characters. [10] were able to achieve an average accuracy of over 80% with essays averaging 787 words/4002 characters in length. However, to train machine learning classifiers, larger amounts of text are preferable.

Another important characteristic of social media is whether authors are identified or not. Some social media types or outlets allow anonymous postings. For example, reviewers on review sites are often anonymous and Myspace allows anonymous members (Facebook does not). Also, some social media forms allow quoting text from other authors, such as including the original email message in a reply email, reTweeting someone else's Tweet, or quoting a previous post in a forum. Automated personality profiling programs need to be able to identify the author of the text to properly attribute personality traits derived from the text. Author identity is also required for accumulating training data because the authors will need to complete personality assessments. Researchers must also consider both the legality and privacy concerns associated with accessing certain social media. For example, the terms of use for Facebook prohibit scraping content via bots. SMS (Short Message Service) messages are typically unavailable in countries with privacy laws.

The type of social media also affects many qualities of the text. Limits on post length such as on Twitter can lead to unusual grammatical usage, which can affect personality profiling. Also different social media may develop abbreviations and other conventions that may not appear in other forms of text media. The particular social media may affect frequency of spelling errors, which can cause problems for text analysis.[1] The social media outlet may also bias the topic, time-period, and author background. For example, web forums typically have a specific topic and the bias in topic may affect how personality correlates with the written text. Likewise, the time-period may affect the relevant topics of discussion in particular social media. Topic differences will affect the word choice, one of the most important features correlated with personality. Finally, the users of social media self-select to use the media and are not representative of the general population. Usage of social media tends to be higher among women, younger users, and urban users [7]. This may be fine if the researcher is interested in the exact demographic represented by the users of the particular social media, but extreme caution is required when trying to generalize any findings to other populations.

2 Social Media Peculiarities

Current social media that can be mined for personality profiling include:

[1] Counting the misspellings offers a useful feature in itself (extraverts tend to leave their text unedited), yet the misspellings make it difficult, on an automated basis, to correctly identify topics, grammatical structures, and of course the words that are being used.

- Emails
- Twitter
- Facebook Status and Wall
- Myspace Bulletins
- Forums
- LinkedIn
- Review sites
- YouTube
- SMS (Short Message Service)
- Blogs and microblogs (e.g., Tumblr)

Email messages have a specified format and requires email-specific preprocessing to extract author text. [9] describes removing headers, signature blocks, and automatic reply quotes. Emails and forums are conversational, so speech acts can be important features that are not found in monolog social media like blogs and to a lesser extent, Twitter (@replies in Twitter allow some conversational threads). Forums often use common forum software, many of which provide an API (Application Programming Interface) to directly extract posts and other information from the forum software without needing to scrape web pages. For example, phpBB, the current most popular forum software provides the REST API to communicate with a phpBB board. Likewise many common blog software packages provide APIs to access blogs and microblogs such as the Blogger JSON API and the Tumblr API.

Review sites tend to be poor choices for personality profiling because authors can be anonymous and reviewers tend to not write very many reviews. The specific purpose of review sites also biases the text making generalization to other contexts very problematic. YouTube suffers from the same anonymous posters problem, lack of sufficient text per poster, and bias based on topic of the posts. LinkedIn does have identifiable authors, but suffers from very little text per author and a very narrow topic area (the career and expertise description).

For some media and occasions, scarcity of time, limits of technology, costs of transmission, purpose of the medium, and etiquette certainly encourage brevity and directness. For example telegrams and radio-communicated Morse Code messages tended to be brief. So are modern SMS messages and Tweets. Blog posts, though, tend to be wordier and more contemplative. It is in such settings, wherein writers express themselves at length, that we expect grammatical choices to vary more, and to tell us more about the personality of the writer. With no rigorous analysis, we will examine a few examples. First from Twitter.com:

> "Still trying to go scuba diving in Mexico for my birthday this summer. Yet, no one is down for the adventure!" * "Why I went #scuba diving with crocs (video). Actual #underwater croc footage starts at 5min 15sec: http://ow.ly/vfczo" * "Wish I was scuba diving chillin with some dolphins and jellyfish instead of being in the presence of these nerds" * "Would you take a scuba dive in the lake that they say the lochness monster is in for 1000 dollars.?" * "Went scuba diving with my dad in the Great Barrier Reef #rad @bvnwnews pic.twitter.com/ahNiLgGGx6" * "@ZoomTV @Ileana_Official scuba dive! I tried once but failed! But will try again soon"

Next some blog posts:

"The vibration became so intense, I could feel it in my bones, and the sound turned into a deafening roar. I could see waterfalls of sand pouring over the coral, and on the sea floor, a few metres below us, cracks began forming and the sand was sucked down. That's when I realised it was an earthquake. The noise was the sound of the Earth splintering open and grinding against itself." - Jessica Read, The Guardian, 24 January 2014

"Suspended in limbo, 130 feet from the surface and nearly 100 feet to the sandy bottom, I watched the bubbles. They playfully danced around each other expanding, breaking, conjoining, chaotic, but always up. The ever-changing surface glimmered above — where air meets water, where life meets death. ...My consciousness crept out of its silent prison and I looked at my gauges." ... "Where had the time gone? I thought. My dive computer started to flash things I had never seen before." - Kelsey, http://tinyurl.com/kbr84tp

"I was surrounded by more fish than I could count, my eyes unaware of where to look next. I had totally forgotten all about my breathing and equalizing and adjusting my buoyancy level, it all didn't matter anymore. (Well, I guess I did alright since I made it back, thankfully my guide was there for the constant check.) Schools of fish swam through us constantly, and I found myself honestly enjoying where I was, meters beyond meters below sea level. Even though my ears hurt like hell at some points, the fascinating colors of the life underwater kept me from signaling to my instructor that I wanted to go back on land." - http://tinyurl.com/mxakoyf

"When I built valves at AirForce, I tested each by pressurizing them in a fixture and tapping the valve stem with a rubber hammer. I had racks of 100 valves at a time, and I went through and did this to each one in turn. That process seated the valve and created a small ring of contact between the synthetic valve and its seat. Sometimes, the valve needed to be hit several times to seat it properly, but it always worked. And it also worked if a valve had a small piece of dirt anywhere in the seals." - http://tinyurl.com/kqcqgxo

First we immediately note that the Tweets employ very few adjectives. The Tweets contain many errors, even though most of the problems could easily be fixed without exceeding the 140 character limit for a Tweet. Beyond simply determining whether or not the samples are within norms of English usage (does the parser fail?), these errors will complicate attempts to discern individual differences between Twitter users on the basis of grammatical usage. Also the topics and content of Tweets are somewhat constrained by the medium: for example very few Tweets explicitly set forth instructions for users to follow. The broadcast nature of Tweets seems to motivate some users to adopt a style that is maximally accessible to their followers. The wideness of the audience further constrains the expression within short Tweets.

Although they too have a broad audience, blog authors exhibit greater variety in style than the Tweet authors do, perhaps because they have ample space in

which to do so coherently. Naturally there are more anaphoric expressions, and much more time is spent on descriptions of objects and processes rather than simply naming them. The use of subordinate clauses varies amongst the blog posts. All these desirable attributes can be applied to personality profiling.

3 Implications for Personality Profiling

Each type and outlet of social media has idiosyncratic characteristics that require specialized processing to extract appropriate author-attributed text. Even after the author-attributed text has been extracted from the social media, there are still other characteristics of each type and outlet of social media that should be taken into account for personality profiling. For example, social conventions can vary across different social media. Emails include various greetings and execute a variety of speech acts that are less common in other social media. Tweets tend to be much more declarative because of the broadcast nature of Twitter. On the other hand, blogs and microblogs are also one-to-many, but exhibit much less of the declarative nature of Tweets. These differences in social conventions can potentially affect the cross-applicability to other social media types for personality classifiers trained in a different type of social media.

Besides different specialized conventions for different social media types, social media types also differ in how quickly these conventions change over time. For example, SMS messages tend to display greater variety over time than perhaps Tweets do, due to rapid evolving of new "text speak" - expressions, abbreviations, emoticons, etc. Although there has been some encouraging success predicting personality from SMS messages [2], a personality predictor depending upon rapidly changing aspects of language may need to be constantly retrained to avoid degradation due to the changes in conventions over time.

Different types of social media are produced under different circumstances and to different audiences. These differences lead to differences in the number of spelling and grammatical errors in the text. SMS messages, and Tweets tend to have the most errors and blogs tend to have the least with other social media types falling in between these extremes. Such differences affect the effort that might be needed to correct errors when processing text for personality profiling. Independent of whether errors are corrected, error rates (before correction) might provide a valuable feature for predicting personality and its value may vary among social media types.

Different kinds of information are conveyed across different social media. Some social media like Twitter are much more topical, dealing more with current events and trends, than other media like blogs and forums, which tend to be on specific topic areas that are less influenced by current events. The specific topic of a blog or forum can influence topicality. For example, political blogs and forums will be much more topical than blogs/forums about topics like parenting that are less influenced by current events. This means that training on text from some social media may need a wider sampling of time periods to avoid over-fitting on topical peculiarities that appear rarely in other time periods.

Different social media differ in the amount of text per author that are typically available. This has implications for the potential accuracy of training personality classifiers. Tweets, SMS text messages, Myspace bulletins and Facebook updates/wall posts tend to have shorter posts. Of these, SMS messages tend to be the most prolific, so many SMS messages over time can easily add up to enough text for accurate training. LinkedIn review sites, YouTube and email tend to have medium sized entries. However LinkedIn entries tend to be single entries, so will not have enough text per person for good training. On the other hand, email is sent constantly over time, so accumulating these can easily provide enough text for training. Review sites may have multiple entries per person, so may provide enough text for training depending on the productivity of the reviewers. However most review sites tend to have very few prolific reviewers, so review sites likely will not have enough authors with enough text for training. Likewise, there may be enough accumulated text accompanying prolific YouTube posters for good training, but more investigation is needed to determine whether there are enough prolific YouTube posters for training. Blogs, Forums, and microblogging present the best sources for large amounts of text by many authors for training purposes.

Users of the different social media types self-select to use that particular form of social media. This will certainly skew the demographics of the studied users. Even more problematic, there are no studies about whether particular personality types are more or less likely to use particular social media. Thus not only are the demographics skewed, which might be corrected with appropriate sampling techniques, but also there might be as yet undocumented biases in personality types introduced because of the selection effect. Likewise, the topic area of a forum, blog, microblog, or YouTube video may bias not only the demographics, but also the personality types of authors.

4 Conclusion

The variety of linguistic expression seen in the blog entries encourages personality profiling applications, whereas inflexible conventions forced by brevity of Tweets tend to narrow the range of linguistic choices. Researchers are then motivated to adapt their methods by focusing on the particular aspects of a given medium that are most useful for personality prediction. Also, some social media, specifically SMS text, exhibit rapid changes in the specialized language employed by users. A classifier trained on unstable aspects of language will quickly degrade in usefulness. This phenomenon enhances the need to identify and exploit those aspects of language usage that change slowly within a given social media context.

References

[1] S. Argamon et al. “Lexical predictors of personality type”. In: *Proceedings, Interface and the Classification Society of North America*. 2005.

[2] T. Holtgraves. “Text messaging, personality, and the social context”. In: *Journal of Research in Personality* 45.1 (2011), pp. 92–99.

[3] F. Mairesse et al. “Using linguistic cues for the automatic recognition of personality in conversation and text”. In: *Journal of Artificial Intelligence Research* 30.1 (2007), pp. 457–500.

[4] S. Nowson. “Identifying more bloggers: Towards large scale personality classification of personal weblogs”. In: *In Proceedings of the International Conference on Weblogs and Social*. Citeseer. 2007.

[5] J. Oberlander and A.J. Gill. “Language with character: A stratified corpus comparison of individual differences in e-mail communication”. In: *Discourse Processes* 42.3 (2006), pp. 239–270.

[6] J.W. Pennebaker and L.A. King. “Linguistic styles: language use as an individual difference.” In: *Journal of personality and social psychology* 77.6 (1999), p. 1296.

[7] Online source PewResearch. *Retrieved*. Apr. 2014. URL: `http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/`.

[8] A. Roshchina et al. “A comparative evaluation of personality estimation algorithms for the twin recommender system”. In: *Proceedings, Workshop on Search and mining user-generated contents*. ACM. 2011, pp. 11–18.

[9] Jianqiang Shen et al. “Understanding Email Writers: Personality Prediction from Email Messages”. In: *User Modeling, Adaptation, and Personalization*. Springer, 2013, pp. 318–330.

[10] William R. Wright and David N. Chin. “Personality profiling from text: Introducing Part-of-speech N-grams”. In: *User Modeling, Adaptation, and Personalization*. Springer, 2014.

[11] Online source www.activityowner.com. *Retrieved*. Apr. 2014. URL: `http://www.activityowner.com/2009/03/14/how-many-characters-are-there-in-a-typical-email-message/`.