User Profile Mining through Linguistic Feature
Specialty of Social Media Language

Miaomiao Yu[1,*]

[1]*Whitman College, Walla Walla, WA, 99362, United States*
**Corresponding author. Email: yum@whitman.edu*

ABSTRACT
Internet and social media usage have reached previously unheard-of levels in recent years. People express their opinions, share feelings, and reflect on news and current affairs on social media. The abundant conversation and expression that took place in those platforms contain valuable linguistic feature which helps researchers study people's social behavior and mining user's profile. This study summarizes previous studies and approaches to understand and analyze linguistic feature in social media. The study also highlights some valuable, meaningful, and accessible linguistic features that researchers should extract from messy information. The paper finds that social media language is more causal than the speaking and writing language, and contain more feature such as emoji, hashtags and URL which can contribute to user profile mining.

Keywords: *Social media, Language, Data Mining, Personality*

1. INTRODUCTION

Since social media has become one of the leading platforms for modern people to share their lifestyles and thoughts, social media users have dramatically increased in the last decade. From 2006, Twitter's first launch, 199 billion active users and 500 million tweets were shared every day until 2020[1]. Facebook has 2.85 billion monthly active users in the first quarter year of 2021[2]. The social interaction that took place in those platforms can help researchers to study people's social behavior without geographic restriction.

Those social media posts become the most common way people express their opinions, share feelings, and reflect on news and current affairs. Social media posts are usually composed of two parts, images, and text. The text in social media posts translates their internal cognition and emotions into a public and understandable form. The link between people's linguistic features such as word choice or language used in conversation and people's cognition process, including thought, perception, and emotion, is well established. Many studies have shown that either people's writing style or informal language can predict their personality and reflect their geographic information [3]. Therefore, analyzing text in user profiles, status updates, comments, and posts that usually contain abundant linguistic data can help researchers to understand users' psychological processes. There is a huge potential for social media research in social psychology, personality psychology, and mental health.

This paper will review the previous studies about language analysis that have been used on or have the potential to apply to social media language analysis. Social media contains a large amount of data that the traditional coding procedure is not efficient and functional well. At the same time, there are many interferences in massive information, so the author also wants to highlight some valuable, meaningful, and accessible linguistic features that researchers should extract from messy information. Other than that, the author will summarize previous analysis of approaches and applications tools to help researchers pull linguistic features from social media, including Natural Language Processing (NLP) and deep learning tools. It hopes that the paper would help researchers understand and process the language data in social media.

2. DEMOGRAPHIC INFORMATION EXTRACTION

The most commonly used social media linguistic features can be extracted from the user's personal post content, personal profile, public interaction with other users, and conversation. Using Facebook is an example the figure one is a sample Facebook page.

User's usually will write their social identities and most representative tags in their profile, for example, their gender, sexual orientation, graduate school and year, etc. That information in their profile can help people to identify and connect with them. Users also use hashtag # to tag contents or trends to represent their identity and demographic information in their profile (e.g. #pridenyc).

3. CLOSE-VOCABULARY VS. OPEN-VOCABULARY APPROACHES OF LANGUAGE ANALYSIS

Social medal language analysis has two different approaches: the open-vocabulary language approach and the close-vocabulary language approach. The close-vocabulary approach is understanding the linguistic feature at the lexical level by using dictionaries and word counting. The open-vocabulary approach does not rely on a single lexicon but extracts the features that relate to semantic and understanding the topic of the sentence. Both methods have advantages and disadvantages in analyzing social media language. There are some key features that can be extracted from text, and researchers can analyze those features through both the close-vocabulary and open-vocabulary approaches.

3.1 Close-Vocabulary Approach Analysis

A sentence is composed of sequence words and commas. The principle behind understanding a sentence is conceptualizing the words in a hierarchical structure. Linguistic interpret a sentence from phonetic, which is the pronunciation of words, to the pragmatic level, which is the context behind a sentence [4]. Then the language will understand from the semantic level for the meaning of words and sentences, to the syntactic level which deals with the grammar, and to the lexical level which is the meaning of the word.

The first step to analyzing a text is to tokenize, break sentences into lexicons, and understand the text from the basic components of the sentence. Through deriving dictionaries from the text, researchers can understand the semantic meaning of words and further use them for content analysis. Linguistic Inquiry and Word Count Program (LIWC2015) is the most commonly used dictionary [5] and a dictionary for social media language will introduce following.

3.1.1 Linguistic Inquiry and Word Count Program (LIWC)

LIWC searches for more than 100,000 words and divides them into more than 80 categories, based on the word's linguistic meaning (e.g., phase, particles, and pronouns), and psychological dimensions (e.g., emotion words).

Words can be generally categorized as content words which are the nouns, verbs, and adjectives, and adverbs that convey the content of the communication; and style words or function words, include pronouns, prepositions, articles, conjunctions, auxiliary verbs. But the only 0.05% style words composed 55% of all the words in speaking and writing English [6].

LIWC compares each word in a given text to the words in an existing dictionary. For example, when researchers need to identify the emotional expression of language, they will define a word as a positive emotion word that (e.g., love, happy, sweet) shares positive feelings or negative emotion words that are used in writing about negative events.

3.1.2 Dictionary of Social media Language and Abbreviation

Unlike writing or speaking language, social media language usually contains a large number and variety of abbreviations, contractions, slang words. Therefore, current dictionary such as LIWC is not enough for social media language analysis. Gómez-Adorno et al.[7] designed a lexical resource to process casual social media language.

This dictionary includes English shortened expressions that are commonly used in social media, such as "xoxo" (kisses and hugs), "BFF" (best friend forever), "LOL" (laughing out loud), and etc, and they grouped slang words have different versions of the expression. They also include abbreviations that consist of the initial letters or parts of words used in both formal and informal English, such as "NY" (New York) and "cm" (centimeters), and constructions for the English language, such as "let's" (Let us) and isn't (Is not). Additionally, the dictionary contains many emoticons which represent facial expressions for emotional expressions aims. Such as ":-)" (smiley face), ":-/" (doubtful face), but these emoticons play different roles and are displayed in different ways across the culture.

3.2 Open-Vocabulary Approches

3.2.1 Pronouns and Verb Tense

Pronouns and verb tenses are both content word categories that can explicitly reveal individuals' attention and focus [6]. Pronouns can be analyzed in a close-vocabulary approach or word counts. For example, older people prefer to use first-person[7]pronouns and traumatized people use more first-person singular pronouns[8]. In the open-vocabulary approach, the pronouns and verb tenses emphasize the speaker's attention. Attention implicitly reveals people's cognition process of encoding their memory and perception. For example, when describing a revealed incident, people used more past tense, whereas discussing a hidden event,

they used more present tense. Comparing the pronoun usage of people and their partners, the relationship is more stable if their pronouns usage is similar.

3.2.2 Emoji and Emoticons

Emoji is the pictographic form of emoticons. Emoji and emoticons function similarly to actual words, acting as a substitute for non-verbal signals and contributing to the overall meaning of social media discourse [9]. Similar to the procedure of analyzing other words, counts, and comparing the emojis and emoticons to the existing dictionaries.

3.2.3 Context and Topic

Words that tend to co-occur with each other would be clustered, and co-occurrence word information is commonly utilized in defining the context and topic of the text. There are four relationships between co-occurring words: a hierarchical relationship in which a concept hierarchy includes the sub-concepts words (e.g. “clothes” and “dresses”); case relation which emphasizes a combination of verb and noun phrase in the meaning restricted cases (e.g. run, dog, toy); compound word relation (e.g. United States of America, American); Synonym relation that several words are synonym (e.g. fear, afraid) [10]. One of the widely used methods is the latent Dirichlet allocation (LDA) method to identify a relevant topic without mentioning the topic[7].

4. LANGUAGE USAGE AND USER PROFILE MINING

4.1 Personality

Personality is a collection of characteristics, attitudes, feelings, and actions that characterize unique individuals. In the last century, psychologists found that linguistic patterns that people could reflect their social needs, power, and achievement [11]. Previous research is mostly based on analysis from “mypersonality”, “Personalitwit" and other personality tests datasets. The personality tests have slight differences on their scale but describe people’s personalities in five aspects: extraversion, agreeableness, openness, conscientiousness, and neuroticism.

Many studies have shown that user’s personalities are highly correlated with their language usage [9], [11], [12], so more researchers try to use a regression model to predict user’s personalities based on vulnerable language data on social media. Most databases are used Big-Five Trait as a personality scale, but the accuracy is relatively low. However, with the development of the NLP algorithm, sentiment analysis, and lexicon database, the accuracy of the personality prediction has dramatically increased in recent years. The table 1 shown the recent research which test the different learning model to predict user personality.

Table 1. Deep Learning Model of Personality Mining

Reference	Dataset	Personality Trait	Learning Model	Average Test Accuracy
[13]	Twitter Data from Commercial source	Sentiments	1. Unigram model (our baseline) 2. Tree kernel model 3. 100 Senti-features model 4. Kernel plus Senti-features 5. Unigram plus Senti-features	60.83%
[14]	Twitter API	Dark Triad; Big Five	1. Support Vector Machine (SVM) 2. Random Forest 3. J48 4. Naïve Bayes (NB) classifier 5. Kaggle models	12.33%
[18]	Twitter Recruitment	Big-Five Persoanlity	1.Linear Regression (LIN), 2. Ridge Regression (RID), 3. Support Vector Machines (linear SVM), and 4. Logistic Regression (LOG).	N/a
[15]	PersonalityCafe	MBTI	Binary logistic regression	38%
[23]	Mypersonality	Big-Five Persoanlity	1. BERT 2.RoBERTa 3.XLNet 4. Combined Model	77.34%

4.2 Others

There are many previous studies that have shown that political ideology is connected to language use[16]. People's liberal-conservative (or left-right) orientation is associated with some psychological traits, such as dogmatism and mental rigidity[17]. Additionally, based on terror management theory that ideological extremity, no matter supporting left or right-wing, promotes self-defense language use[18].

Meanwhile, users' unhealthy mental states and abnormal behavior can reflect their linguistic features. [19] have found that there are language usage differences between ADHD, Anxiety Disorder, Bipolar Depression, Borderline Personality Disorder, Major Depression, Eating Disorder, OCD, PTSD, Schizophrenia, Seasonal Affective Depression group, and control group. Additionally, people who have suicide attempts have similar language signals when they are at high risk [20]. Through language usage and profile analysis, the machine learning model can distinguish users who have Narcissistic personalities, Machiavellianism, and psychopathy [21].

5. CONCLUSION

This paper presented and reviewed several approaches to extract, process, and analyze language data from social media. The paper finds that the language data in social media is highly accessible, various, and trackable. Unlike the formal writing language or verbal language, social media language is more causal. Social media languages include more abbreviation which can be inteprate and analyze through updated abbreviation dictionary. The emoji and hashtags usage as a significant part of social media language, they are also meaningful to study. Thus, instead of delating hashtags and URL in data cleaning, researchers can focus on more about the topic mentioned by hashtags and URL which can provide more insight of analyzing context of language. Emoji may provide mood and attitudes. It needs to be noticed that different culture and generation have different interpretation of emoji, and currently there are not too much studies on this area. All of those features in language correlate to user's profile and can help on mining user's implicit information.

The author believes that the application of social media language usage analysis has huge and unprecedented potential. For example, the analysis of depression user profiles on social can contribute to building alarms to prevent suicide automatically [23]. Some research provides a reference to playing a role in the regulation of public speech and stopping the spread of rumors and fake news. The detector of dangerous speech from bomb threats and gun shooters can efficiently prevent terrorist attack and reduce damage.

However, the reliability and validity of current language analysis and personality prediction model are realtive low. Although previous analysis and model based on personality dataset which contain large amounts of population, there are still many low socioeconomic and unrepresentative populations being ignored during the research process. Moreover, some research has shown that people prefer to act differently online [24]

Future researchers can take the advantage of this abundant social-linguistic data in social media and analyze the interaction and conversation between users. Meanwhile, by the development of NLP and semantic analysis model, researchers can better process ambiguous language and more precisely predict the user's profile. In conclusion, there is a huge potential for using social media linguistic data to study human behavior and mining user's profile.

REFERENCES

[1] "Twitter by the numbers: Stats, Demographics & Fun Facts." [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/.

[2] "Facebook revenue and Usage Statistics (2022)," Business of Apps, 19-Jan-2022. [Online]. Available: https://www.businessofapps.com/data/facebook-statistics/.

[3] M. C. M. Jalbuena, "Linguistic Features of English in Twitter," p. 13.

[4] P. M. McCarthy and C. Boonthum-Denecke, Eds., Applied Natural Language Processing: Identification, Investigation and Resolution. IGI Global, 2012. doi: 10.4018/978-1-60960-741-8.

[5] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," J. Lang. Soc. Psychol., vol. 29, no. 1, pp. 24–54, Mar. 2010, doi: 10.1177/0261927X09351676.

[6] H. A. Schwartz and L. H. Ungar, "Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods," Ann. Am. Acad. Pol. Soc. Sci., vol. 659, no. 1, pp. 78–94, May 2015, doi: 10.1177/0002716215569197.

[7] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," Cogn. Emot., vol. 18, no. 8, pp. 1121–1133, Dec. 2004, doi: 10.1080/02699930441000030.

[8] D. Marengo, F. Giannotta, and M. Settanni, "Assessing personality using emoji: An exploratory study," Personal. Individ. Differ., vol. 112, pp. 74–78, Jul. 2017, doi: 10.1016/j.paid.2017.02.037.

[9] K. Morita, E.-S. Atlam, M. Fuketra, K. Tsuda, M. Oono, and J. Aoe, "Word classification and hierarchy using co-occurrence word information," Inf. Process. Manag., vol. 40, no. 6, pp. 957–972, Nov. 2004, doi: 10.1016/j.ipm.2003.08.009.

[10] H. A. Schwartz et al., "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," PLoS ONE, vol. 8, no. 9, p. e73791, Sep. 2013, doi: 10.1371/journal.pone.0073791.

[11] L. Qiu, H. Lin, J. Ramsay, and F. Yang, "You are what you tweet: Personality expression and perception on Twitter," J. Res. Personal., vol. 46, no. 6, pp. 710–718, Dec. 2012, doi: 10.1016/j.jrp.2012.08.008.

[12] Celli, F.; Rossi, L. The Role of Emotional Stability in Twitter Conversations. In Proceedings of the Workshop on Semantic Analysis in Social Media, Avignon, France, 12 April 2012; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 10–17

[13] E. Tighe and C. Cheng, "Modeling personality traits of Filipino twitter users," Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, 2018.

[14] Cui, B., and Qi, C. Survey analysis of machine learning methods for natural language processing for MBTI personality type prediction 2017.

[15] J. Sterling, J. T. Jost, and R. Bonneau, "Political psycholinguistics: A comprehensive analysis of the language habits of liberal and conservative social media users.," J. Pers. Soc. Psychol., vol. 118, no. 4, pp. 805–834, Apr. 2020, doi: 10.1037/pspp0000275.

[16] J.-W. van Prooijen and A. P. M. Krouwel, "Psychological Features of Extreme Political Ideologies," Curr. Dir. Psychol. Sci., vol. 28, no. 2, pp. 159–163, Apr. 2019, doi: 10.1177/0963721418817755.

[17] J. Greenberg and E. Jonas, "Psychological motives and political orientation--The left, the right, and the rigid: Comment on Jost et al. (2003).," Psychol. Bull., vol. 129, no. 3, pp. 376–382, 2003, doi: 10.1037/0033-2909.129.3.376.

[18] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses," in Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, Colorado, 2015, pp. 1–10. doi: 10.3115/v1/W15-1201.

[19] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses," Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2015.

[20] C. Sumner, A. Byers, R. Boochever, and G. J. Park, "Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets," in 2012 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, Dec. 2012, pp. 386–393. doi: 10.1109/ICMLA.2012.218.

[21] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," J. Big Data, vol. 8, no. 1, p. 68, Dec. 2021, doi: 10.1186/s40537-021-00459-1.

[22] S. T. Rabani, Q. R. Khan, and A. M. Khanday, "Quantifying suicidal ideation on social media using Machine Learning: A Critical Review," Iraqi Journal of Science, pp. 4092–4100, 2021.

[23] Vince, Gaia. "Evolution Explains Why We Act Differently Online." BBC Future, 2018, www.bbc.com/future/article/20180403-why-do-people-become-trolls-online-and-in-social-media.