What the Language You Tweet Says About Your Occupation

Tianran Hu, Haoyuan Xiao, and Jiebo Luo
Department of Computer Science
University of Rochester
Rochester, NY 14623
{thu, hxiao6, jluo}@cs.rochester.edu

Thuy-vy Thi Nguyen
Department of Clinical and
Social Sciences in Psychology
University of Rochester
Rochester, NY 14623
thuy-vy.t.nguyen@rochester.edu

Abstract

Many aspects of people's lives are proven to be deeply connected to their jobs. In this paper, we first investigate the distinct characteristics of major occupation categories based on tweets. From multiple social media platforms, we gather several types of user information. From users' LinkedIn webpages, we learn their proficiencies. To overcome the ambiguity of self-reported information, a soft clustering approach is applied to extract occupations from crowd-sourced data. Eight job categories are extracted, including Marketing, Administrator, Start-up, Editor, Software Engineer, Public Relation, Office Clerk, and Designer. Meanwhile, users' posts on Twitter provide cues for understanding their linguistic styles, interests, and personalities. Our results suggest that people of different jobs have unique tendencies in certain language styles and interests. Our results also clearly reveal distinctive levels in terms of Big Five Traits for different jobs. Finally, a classifier is built to predict job types based on the features extracted from tweets. A high accuracy indicates a strong discrimination power of language features for job prediction task.

Introduction

The recent statistics published by LinkedIn revealed that increasing a number of US employers relies on personality assessments as a way to screen their prospective employees (26% in 2001 to 57% in 2013[1]). It appears that certain jobs are more likely to attract or fit better with certain people. While this hiring practice has been commonly used by organizations, there has not been any research that directly looks at the connections between different job positions and the people that are in those positions. Indeed, many studies on sociology and psychology suggest that many aspects of a person's lives are deeply connected with their jobs (Lindquist, Beilin, and Knuiman 1997; Strully 2009). In this study, we looked at the connections between different job positions and the characteristics of the people that hold those positions, including the language that they use, the topics that they are interested in, and their personality traits. Our findings suggested that jobs are more than just what people do, but also who they are, and how they present themselves to others.

[1]https://www.linkedin.com/pulse

To uncover the connections between jobs and the characteristics of people who hold those jobs, the procedure involves several steps. First, we need to separate people based on their jobs. Categorizing people into job types is particularly challenging because most jobs are less specialized but instead might share some common responsibilities or duties. Therefore, in this step, we rely on the skills that people have to determine the job categories that they are more or less likely to belong to. A soft clustering method is used to assign a normalized weight to each job type for a person based on her skills. In the next step, we also need to learn about each person's characteristics, such as her linguistic patterns, interests, personality traits, and so on. Achieving those two steps allow us to compute the Pearson correlation coefficients between people's characteristics and their weights on different job types. To achieve those two steps, we need to first figure out people's skills, and then learn about their characteristics.

Social media posts have been proven an effective information source in gaining knowledge of people (Silva et al. 2014; Schwartz et al. 2013). Individuals tend to use social media sites as platforms for self-presentation (Schau and Gilly 2003). Therefore, we extract rich information about people's characteristics from what they share online. We select Twitter data to support our study, because Twitter is widely and effectively used in user profiling (Nguyen et al. 2013), and it is relatively easy to collect. For the linguistic features, we apply closed vocabulary and open vocabulary approaches described in (Schwartz et al. 2013) to ones' tweets. The closed vocabulary approach uses a fixed lexicon to analyse text, while the open vocabulary approach does not limit the vocabulary. In the former, we apply Linguistic Inquiry and Word Count (LIWC) to extract 92 linguistic features from tweets. In the latter, two types of linguistic features are learned: representative words and phrases , and topics. Tweets also provide rich information on personalities (Schwartz et al. 2013). We apply the IBM Watson Personality Insights service API to compute the personality traits from tweets. Personality is measured by Big Five Traits, a widely examined theory of the five broad dimensions describing human personality (Goldberg 1993). The five dimensions are, Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

We take advantage of another social media – LinkedIn – to collect the job information. LinkedIn users share their in-

dustries, career experiences, education status, interests, and so on. A skill list is also displayed on a user's LinkedIn page. The listed skills are endorsed by the users who knows this person. Each skill in the list is associated with the number of votes this skill receives. Although we can learn LinkedIn users' industries from their profiles, this self-reported information is usually imprecise and ambiguous. For example, we manually check the profiles of all the professors in our department. Some of them report the industry as "Computer Science", others fill in "Higher Education" or "Research". To come up with more precise and accurate job categories, we apply a soft clustering method to people's skills endorsed by others, and use a *group* of skills to describe a job. There are three benefits in doing this: 1) Skill is a concept of finer granularity than industry. 2) Skills of a person are voted by the people who know the person. The crowdsourced data is more accurate than the information provided by the person him/herself. 3) Given the complex nature of today's jobs, it makes more sense not to restrict a person to only one occupation. Instead, one individual can perform several responsibilities and possess several skills that might possibly qualify him or her for multiple job categories. For example, the job of a product manger in an IT company could be described with both "programmer" and "manager".

To match LinkedIn and Twitter accounts of the same people, we collect the users from about.me[2]. It is an integration platform, which allows users to link multiple online identities, such as Twitter, LinkedIn, Facebook, and so on in one profile. We randomly sample about.me users, and retain valid users who register both Twitter and LinkedIn accounts.

Interesting and significant differences are uncovered among people with different jobs in this study. The open vocabulary approach provides more dimensions implying the traits of jobs. For example, office clerks talk more about daily life than people of other jobs. According to our results, people with different jobs clearly have different interests. People who are doing public relation and writers show strong interests in politics and social events. The divergences of personalities are also found. For example, managers are the most extroverted, while software engineers are the least. Base on these observations, we build a classifier to predict jobs using the features extracted from tweets. A high accuracy (80%+) indicates that adequate personal trait information is hidden in social media posts.

Our contributions are threefold:

- We propose an approach to categorize proficiencies by softly grouping skills a person has, using skill information gathered from LinkedIn.

- We uncover significant and interesting divergences of linguistic patterns, interests, and personalities among people with different jobs.

- We build up an occupation classifier with high accuracy is built up based on features extracted from Twitter posts.

_____
[2]about.me

Related Work

Studies of sociology and psychology reveal that many aspects of human life, such as health condition, lifestyle, personality, and so on, are deeply connected with their jobs. The effects of work stress on long-term blood pressure is studied in (Lindquist, Beilin, and Knuiman 1997). The paper finds that it is the ways of dealing with work stress rather than the stress itself that are significantly related to blood pressure. In (Strully 2009), the correlation between employment status and health condition is studied. The authors reported that employment status may impact certain health outcomes. In addition, sickness absence, the use of alcohol, and anxiety of reorganization are also proven to be related to working life (Voss, Floderus, and Diderichsen 2004). Jobs also influence lifestyles. Payne et al. discussed the different lifestyles of employees in high-strain jobs and low-strain jobs (Payne, Jones, and Harris 2002). They found that people with high-strain jobs exercise far less than those with low-strain jobs. The relationship between the Big Five traits of personality and job criteria is investigated in (Salgado 1997; Hurtz and Donovan 2000). The findings indicate that Conscientiousness and Neuroticism are valid predictors for job performance. (Judge, Heller, and Mount 2002) investigates job satisfaction and personality. The authors found that, job satisfaction is positively correlated with Extraversion, Agreeableness, and Conscientiousness, while it is negatively correlated with Neuroticism. There is also work focusing on specific job types. Perceived social support, job stress, health, and job satisfaction among nurses are studied in (Bradley and Cartwright 2002). (Reichel and Pizam 1984) compared the U.S. hospitality industry employees' with other industries on work attributes, demographics, and class perceptions. The results show that the people of this industry tend to be less satisfied with the job, and take their work as unimportant elements in their self-accomplishments.

Although much work has been done, previous work on the relationship between human characteristics and jobs focuses on either employment and work status, or a single job type. It is still not clear what divergences exist across various jobs. Due to the inherent limits of transitional data collection methods, previous work of sociology and psychology usually suffers from the problem of small sample size. In our work, a large population is collectied from social media platforms, and we compare multiple human characteristics across jobs.

The psychological meaning of words is well studied in Computer Science and Linguistics. Linguistic Inquiry and Word Count (LIWC), as a computerized text method, is introduced in (Tausczik and Pennebaker 2010). (Pennebaker and King 1999) investigates the difference of individual linguistic styles. This work reports the significant difference across their language patterns, and proves the effectiveness of LIWC. Based on the linguistic features, (Mairesse et al. 2007) introduce an approach to recognizing personalities from conversation. The boom of social media attracts a lot of research that are based on this new data source. It is shown that personalities can be recognized using people's social media network structures (Staiano et al. 2012), pro-

files (Quercia et al. 2011), and contents of posts (Qiu et al. 2012; Golbeck et al. 2011). (Vinciarelli and Mohammadi 2014) provide a nice survey on computing personality from social media data. In this paper, we follow the approaches described in (Schwartz et al. 2013) to extract linguistic patterns. In their work, Schwartz et al. propose two approaches to learning people's language styles from social media texts. They report significant differences in language styles across several features including genders, ages, and personalities.

Data Collection and Preprocessing

User Collection

We use about.me search API[3] to collect users. The input of this API is a name, and it returns the information of at most 100 users who has the same or a similar name. We gather the 1,000 popular male and female first names in the U.S, and feed these 2,000 names to the API. The API returns 150K user profiles. We first select the users who have both LinkedIn and Twitter accounts. Following their Twitter accounts, we download these users' 3,000 most recent tweets using Twitter search API[4]. We remove the users who do not have enough English tweets (less than 2,000), to guarantee the significance of our results. We then collect the job information through LinkedIn links. We also remove the people who does not have a skill list on the LinkedIn page, because we cannot categorize their jobs for them. Eventually, we end up with 9,800 users in total.

Tweets Cleaning and Phrase Selection

Twokenizer[5] is an NLP tool designed for tweets especially. It detects abbreviations or slang (*b4* for *before*, *fb* for *Facebook*), misspellings or spelling variants (*fir* for preposition *for*), and emoticons (*:)*, *<3*) in tweets. However, in practice Twokenizer usually fails in extracting Twitter official emojis[6], because these emojis concatenate with other words in many cases. Therefore, we apply Twokenizer to separate words from tweets, and we also use a fixed emoji list to detect all the official emojis. Terms that are too popular (used by more than 95% users) and too unpopular (used by less than 10% users) are removed.

We apply the same thresholds to phrases (2-grams and 3-grams). To avoid extracting simple word combinations instead of meaningful phrases, we apply Point-wise Mutual Information (PMI) to distinguish these two (Schwartz et al. 2013). PMI measures how more informative it is to take a phrase as a whole compared with taking it as separate words. It is formally defined as follows:

$$pmi(phrase) = \log \frac{P(phrase)}{\Pi_{t \in phrase} P(t)} \qquad (1)$$

where t indicates the terms in the phrase, and $P(phrase)$ is the probability of observing the phrase. We filter out all phrases with PMI value lower than 2* length, where length

is the number of words in this phrase, and we consider these phrases as uninformative.

However, it is infeasible to compute PMI for every phrase, due to the large size of candidate set. We exploit the Apriori property of phrases (Agrawal et al. 1996) to reduce the candidate set. The apriori property states that if a set is not frequent, then all the super sets are not frequent neither. In our case, for example, if "love you" is used by less than 10% of the users, then "I love you" has to be used by less than 10% of the users. In practice, we first obtain all the 1-gram terms using Twokenizer and emoji list, and remove the 1-gram terms that are used by less than 10% of the users. We then combine 1-gram terms pairwise to generate 2-gram phrases. The same steps are applied in generating 3-gram candidate set from 2-gram phrases. At last, we remove words and phrases that are used by more than 95% of the users from the candidate sets, and apply PMI to remove the uninformative phrases. The preprocessing leaves us 18,082 words, 7,852 2-gram, and 10,903 3-gram phrases.

Job Categorization

Most LinkedIn users report their industries in the profiles, such as "Marketing", "Public Relation", and "Web Design". However, these industry tags are not precise enough; and some of them are even ambiguous. We observed many cases where people with different industry tags actually work in the same position. For example, a computer science professor could report his/her industry as "Computer Science", "Higher education" or "Research". To learn the real job a person is doing, we utilize the skills a person has. A person's skills are endorsed by people who connect with this person, and are listed in one's linkedIn page. Besides a list of endorsed skills, the number of endorsements of each skill is also available.

We vectorize a person's skill list into a fixed order vector. Each dimension represents a skill, and the values of elements in a vector are the votes this person receives on the corresponding skills. We feed the skill vectors to Latent Dirichlet Allocation (LDA) model to generate a soft clustering of people. Two matrices are learned using LDA. The topic matrix describes the distribution of skills in each cluster. In other words, clusters are linear combinations of skills. These clusters are used as definitions of job in this study. The document matrix describes the people's weights on each job. Therefore, each person is not limited to one job. These assigned weights are used to calculate the correlations between jobs and linguistic patterns, as well as personality traits. There are three benefits from categorizing jobs in this way. First, comparing with a broad industry, it is more precise to describe proficiency with a set of skills. Second, as skills of a person are voted by others, the crowdsourced data is more accurate than the self-reported information. Third, soft clustering solves the ambiguity in defining one's job. Taking the example we used in introduction, now a product manger in an IT company can be defined as 30% "programmer" and 70% "manager".

The number of clusters (topics) k is a crucial parameter in LDA setting. We use perplexity to determine how many topics are needed. Perplexity measures how good a language

---

Job Name	First Skill	Second Skill	Third Skill	Fourth Skill	Fifth Skill
1. Marketing	Digital MKTG	Social Media MKTG	Online MKTG	Digital Strategy	Advertising
2. Administrator	Public Speaking	Leadership	Fundraising	Event Planning	Coaching
3. Start-up	Start-ups	Entrepreneurship	Strategy	Business DEV	Management
4. Editer&Writer	Blogging	Editing	Journalism	Copy Editing	Storytelling
5. Software Engr	MySQL	CSS	JavaScript	PHP	jQuery
6. Office Clerk	Microsoft Office	Microsoft Excel	PowerPoint	Microsoft Word	Customer Service
7. Public Relation	Public Relations	Media Relations	Press Releases	Strategic COMM	Corporate COMM
8. Designer	Graphic Design	Web Design	Photography	Illustrator	Photoshop

Table 1: Extracted job categories using LDA. We list 5 most weighted skills of each job, and manually label them according to the skills. "Engr", "COMM", "MKTG", and "DEV" are short for Engineer, Communications, Marketing, and Development, respectively.

model predicts unseen documents, and decreases monotonically as the number of topics increases (Blei, Ng, and Jordan 2003). We use 10-cross validation to test k from 2 to 100. Average perplexity drops fast as k increases from 2 to 20, and keeps almost unchanged as k goes above 20. Therefore, we select number of cluster as 20. To guarantee that each job covers a relatively large population, we remove the jobs that contains less than 300 users. We end up with 8 jobs after these processes. In Table 1, we list the 5 most heavily-weighted skills for each job category and also manually assign a label for each job category. The results are consistent with common sense. Take the fifth job "Software Engineer" as an example, the skills associated with this job are mostly programming languages (*JavaScript*), and tools (*MySQL*). The remaining 7 jobs are Marketing, Administrator, Star-up, Editor&Writer, Office Clerk, Public Relation, and Designer.

Please note that we do not aim at finding out all the job types, but focus on the divergences among people with different jobs. Therefore, we did not try to extract an exhaustive or complete list of jobs in this study.

Closed Vocabulary Approach and Results

We first used a fixed lexicon (LIWC) to analyse the linguistic patterns of different jobs. Since the vocabulary is fixed in the study, this approach is called closed vocabulary approach in (Schwartz et al. 2013). Linguistic Inquiry and Word Count (LIWC) is a computer program for language analysis (Pennebaker and King 1999). LIWC2015 computes a 92-Dimensional vector for a document based on a given dictionary. The dimensions include: Sixltr (percentage of words that longer than 6 letters), Anger (percentage of words that indicates anger), and so on.

We combine all the tweets of a person as a single document, and then calculate the LIWC features. Pearson Correlation Coefficients are computed to uncover the unique language traits of different jobs. Due to the length limit, in Figure 1 we list 25 LIWC features, and their correlations with each job. The p-value of these correlations are all smaller than 10^{-4}. Significant and interesting correlations are observed. We summarize them as follows (the terms in brackets are the feature names of LIWC outputs):

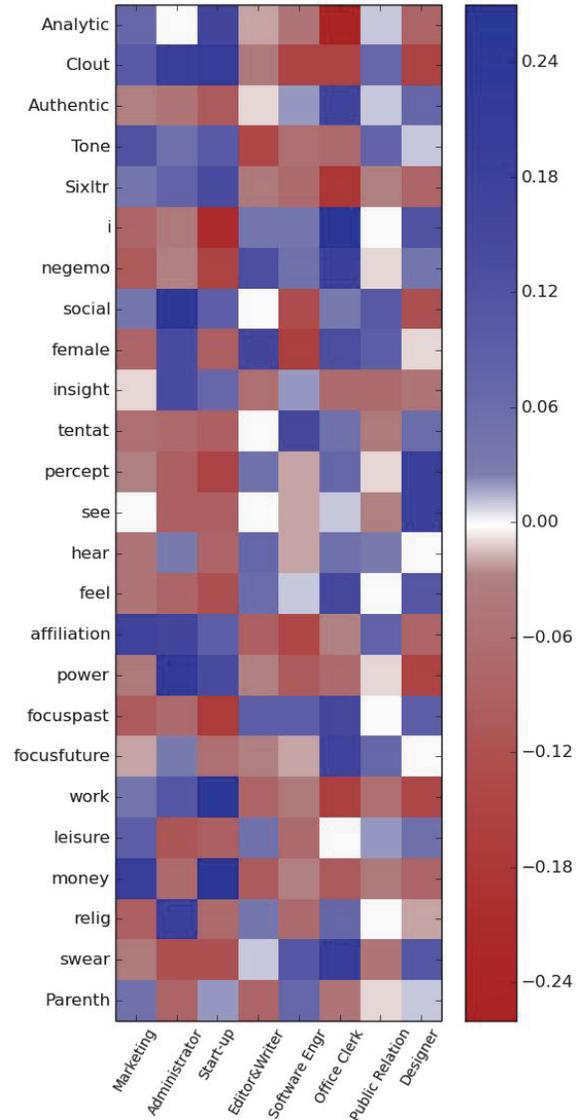- Marketing people talk a lot about money and affiliation

Figure 1: 25 LIWC features, and their correlations with each occupation.

(*money*, *affiliation*), and they mention more about leisure (*leisure*). Their emotional tone scale (*tone*) is high. A higher emotional tone scale indicates more lively than other people. Meanwhile, they talk less about religions (*relig*), and show fewer negative emotions (*negemo*).

- For the administrators, they prefer to use first-person plural, and mention more about power, insight and society. On the other hand, this group of people use fewer swear words (*swear*) and parentheses (*Parenth*) comparing with people of other jobs.

- People in start-ups post a lot of contents about money and work (*work*) in their tweets, while they do not use first-person singular much, and they do not express anger or negative emotion much.

- As to the editors and writers, they use many words about genders (*female*, *male*). Different from other jobs, people with this job mention more about negative emotions.

- Software Engineers' posts are more tentative (*tentat*), and the usage of swear words is positively correlated to the job. Not surprisingly, programmers mention less about female and social. Likewise, Programmers' power-awareness (*Clout*, *power*) is relatively low.

- Office clerks mention a lot about themselves (*i*) in their tweets, and the usages of negations (*negat*), negative emotions are higher. People doing this job post fewer analytic tweets (*Analytic*) and long words (*Sixltr*), and they do not talk much about work.

- People who are doing public relation tend to be more future-focused (*focusfuture*). In other words, they like to use words like *will*, *may*, and *soon* in their tweets. Same as marketing people, this group of people's emotional tone scale is relatively high.

- Designers like to describe what they see, hear, and feel (*seen*, *percept*) in tweets. Similar to programmers, they have low power-awareness (*Clout*, *power*).

Open Vocabulary Approach

In addition to using a fixed lexicon to analyse the linguistic patterns of different jobs, in this section we discuss an approach leveraging all the words in ones' posts. This approach is called the Open Vocabulary Approach in (Schwartz et al. 2013). We first count the words and phrases in a person's tweet, then use TF-IDF to weight all the terms based on their frequencies. Pearson Correlation Coefficients are computed between the people's TF-IDF weights of each term and their weights on each job. The positively correlated terms to a job are used more by the people with the job. In other words, these terms are positively distinguishing to this job. Oppositely, negatively correlated terms to a job are used less often by the people with the job, they are negatively distinguishing to this job. After this, we aggregate each person's tweets into a single document, and feed the documents to LDA to extract 2,000 topics. We then calculate the correlation between people's weights on the 2,000 topics and their weights on the 8 jobs. By doing

this, we find the positively and negatively distinguishing topics to a job. Because topics extracted from tweets are meaningful combinations of words, we can learn more about the interests and focusings across people with different jobs.

Results of Open Vocabulary Approach

We find interesting differences across jobs. Each job has its unique preference to words, phrase, and topics, suggesting the divergences on working content, interests, and characteristics of people with different jobs. Some topics are highly positively correlated to one job, while highly negatively correlated to others. We plot these distinctive terms and topics in the form of word cloud (Figure 2), due to the length limit, we only plot two jobs. We summarize our results as follows.

Marketing The positively correlated terms and topics indicate that the most distinguishing contents to marketing people are about the methods, platforms, and strategies of marketing. Meanwhile, this group of people have less interests in politics, policy, and education.

The most positively correlated word is *marketers*, indicating the occupation of this group of people. Since social media is growing into a more and more important marketing platform, the correlation of *social media* is quite high. Others words and phrases, such as *your brand*, *targeting*, *campaigns*, *marketing strategy*, and *lead generation*, are also extracted, implying the unique tweet contents of marketing people. To our surprise, many most negatively correlated words and phrases to this job are about education, such as *paper*, *on campus*, *professor*, and *semester*. Meanwhile, some legal words, such as *justice*, *evidence*, *civil*, *federal*, and *violence*, are also highly negatively correlated to the people who have this job.

The most positively correlated topic is about mobile and mobile advertising to this job (*adroid*, *google*, *adwords*, *brands*, and *ad*). The following two topics are about data analysis (*analytics*, *infographics*, *data*), and social media platforms (*twitter*, *pintrest*, *facebook*, *instagram*), respectively. The most negatively correlated topic is formed by words about school and politics (*inequality*, *racial*, *racism*, *college*, *students*), which aligns with the negatively correlated words and phrases. The second most negatively related topic is about politics (*hillary*, *demdebate*[7], *#sotu*[8]) and social events (*ebola*, *feminist*).

Administrator Administrators talk more about leading, education, and religion in their tweets, while they do not express negative emotions or use negations much. Comparing with other jobs, these group of people are more careless about techniques.

Administrators' tweets are more invigorated. They use many didactic words and phrases like *make a difference*, *courage*, and *honor*. Vocabularies related to leading are also frequently used. For example, words and phrase, such as *leaders*, *we must*, and *leadership* have high positive correlation scores. Meanwhile, administrators prefer to use phrases related to unity such as *we need to*, *we must*, *we are*, *how*

---

[7] meaning: Democratic Party Presidential Debates
[8] meaning: State of the Union

(a) Word cloud for Editors&Writers

(b) Word cloud for Marketing

Figure 2: A word cloud illustration of highly correlated words, phrase, and topics of two occupations: Marketing and Editors&Writers. Positively correlated ones are plotted in blue (left), and the negative ones are in red (right). The big circle in the middle contains the highly correlated words and phrases. Small surrounding circles represent topics, with each circle being a topic. We select top 200 positively and negatively scores words and phrase, and for each topic we select the top 10 prevalent words.

do we. This agrees with our finding using closed vocabulary approach about their preference for useing first-person plurals. There are many religious vocabularies highly correlated to them too, such as *blessing*, *pray*, and *god's*. Different from marketing people, education vocabularies (*students*, *youth*, *college*) are positively correlated to this group of people. A large percentage of most negatively correlated words to managers is related to techniques, such as *app*, *version*, *iphone*, *ui*, and *interface*.

The most positively correlated topic to this job is about education and politics (*inequality*, *racial*, *racism*, *college*, *students*). Interestingly, this topic is among the most negatively correlated topics to marketing people. A topic about religion (*worship*, *jesus*, *bible*, *lord*) also has high positive correlation with this job. This is consistent with the fact that religious vocabularies are positively correlated. The third most positively correlated topic is about education (*#edchat*, *#edtech*, *#education*, *tearchers*, *#edtechchat*). The most prevalent terms in this topic are usually attached with hash-tag, indicating their interests on related discussion on Twitter. The most negatively correlated topic is formed by negation terms (*dont*, *cant*, *:d*), and negative emotion words, such as *jealous* and *bloody*. The second topic is about popular mobile device (*ios*, *apple's*, *iphone*), and sports (*nfl*, *nike*, *sports*). The insulativity of managers from programming skills is revealed by the third most negatively correlated topic, consisting mainly of programming words (*php*, *js*, *api*, *github*, *jquery*).

Start-up Start-ups people share some likes and dislikes with marketing people and mangers. Avoidance of using self-denial expression is one of their unique features.

Almost all the highly positively correlated words and phrases, such as *founders*, *investors*, *growth*, *valuation*, and *companies*, are about running company, investment, and business. Not surprisingly, *silicon* is also one of the most positively correlated words. Like administrators, start-up

people dislike to use negating words (*can't*, *dont*), but they especially dislike self-denial. *i can't* is the most negatively correlated phrase to this group of people. This dislike of self-denial is also reflected by the high negative correlation score of *i don't know* and *i didn't*. *she* and *her* also do not appear much in their tweets.

The most positively scored topic contains words like *bitcoin*, *tesla*, *crowdfunding*, *startups*, *inverstors*, and so on. The following two topics are also among the top positively correlated topics of marketing people. These two topics are about data analysis (*analytics*, *infographics*, *data*), and mobile advertising (*adroid*, *google*, *adwords*, *brands*, and *ad*). The overlapping indicates the similar concerns of people with these two jobs. Same as administrators, the most negatively scored topic of start-ups people is the one formed by neglecting terms (*dont*, *cant*, *:d*), and negative emotion words(*jealous*, *bloody*). The following two negative topics mainly consist of emoticons, such as *;)*, *;-)*, and *:(*.

Editor & Writer Editors and writers show more interest in politics, as well as in social events. They also talk more about reading and books. Comparing with the jobs we mention above, this people with this job use more emoticons.

The four most positively correlated words to editors are *editors*, *journalist*, *writer*, and *reporter*, clearly implying the occupation of these people. In addtition to, the words closely related to this job, such as *headline*, *pulitzer*, and *newspaper*. We also observed many words related to social events like *murder*, *investigation*, and *police*. The two most negatively correlated type of words are: techniques words (*interface*, *setup*, *framework*), and businesses words (*management*, *customer*, *business*).

The topic that attracts the most attention from this group is the one about politics (*hillary*, *demdebate*, *#sotu*) and social events (*ebola*, *feminist*). Please note that this topic is also the one that attracts the least attention from marketing people. Editors also would like to talk about reading.

186

The second most positively scored topic is about *#romance*, *#thriller*, *#books*, and *#kindle*, followed by the third emoticon topics (*;), ;-), - -*). The most unpopular topic to editors is about analytics and techniques, words like *data*, *analytics*, and *mobile* are among the representatives of this topic.

Software Engineer   Terms and topics indicate that software engineers use way more technique terms than others, while they tend to talk less about females and social life.

Software engineers mention a lot about techniques and coding (*web*, *ui*, *code*, *plugin*). In fact, we could not find a word or phrase that is not about techniques among the top 200 most positively correlated. The most negatively correlated word to software engineers is *summer*, followed by *girl* and *her*. *relationship* is also a less often used word by them. They also express less interest on excitements and celebrations. For example, *love this!*, *so excited*, *sunday*, and *gift* are all highly negatively scored. Moreover, their expressions of compliment (*thank you*, *cutest*, *so proud of*) appear less than people with other jobs.

The two most positively correlated topics, with no surprise, are both about programming techniques. The representative words of them are mainly about programming languages (*php*, *java*, *python*), and tools (*github*, *photoshop*, *api*). The last topic among the top three positive topic is about apple products (*ios*, *apple*, *iphone*) and sports (*nike*, *sports*, *nfl*). The most negatively correlated topic to this job is mixed by three types of words: marketing vocabularies (*marketing*, *advertising*, *#marketing*), family vocabularies (*husband*, *dinner*, *family*), and vocabularies of praising (*fabulous*, *angeles*). The following two topics are formed by Twitter official emojis (please refer to the supplementary material for the illustration) and regular emoticons (*<3, ;)*), respectively.

Office Clerk   The most positively distinguishing words and phrases of people with other jobs are related to their job contents, such as vocabulary of programming languages to software engineers, and vocabulary of reading and writing to editors. However, office clerks' most positively correlated words have nothing to do with their job. They emphasize life and family in their posts instead, and express strong self-awareness. They tend to use more emojis and emoticons, and their negative emotions are stronger than people with other jobs. Negatively scored topics indicate that office clerks show less interest on business, and data analysis.

The most positively scored term to them is *my life*. Other phrases related to daily life such as *woke up*, *fall asleep*, and *my hair* are also highly scored. Words and phrases of self-expression (*i just want*, *i wish i*, *i hate*) appear a lot in office clerk's tweets. It is interesting that a lot of self-expression portrayed among office clerks are emotional. They mainly express lack of motivation, negative emotions, unfulfilled wants and wishes. Words such as *semester*, *studying*, and *homework* get high positive scores, implying this group of people's strong interest in education. They have a mixed negatively correlated vocabulary. The less usage of *the future of* aligns with the low future-focused score we calculated using the closed vocabulary approach. Technique vocabularies (*web*, *online*, *app*) also appear less in their tweets.

Moreover, their posts are less motivated. The terms like *interesting*, *creating*, *great* all have high negative scores.

Contrary to software engineers, this group of people is most positively related to the topic consisting of emojis. Moreover, unlike administrators and star-ups people, the topic formed by neglecting terms (*dont*, *cant*, *:d*), and negative emotion words (*jealous* and *bloody*) is the second most positively scored topic. The third most popular topic of this group is about entertainments (*#nowplaying*, *photo*, *:p*, *album*). As to the negatively correlated topics, the top three are about analytics, mobile advertising, and business, respectively. These topics are positively scored topics to administrators, marketing people, and people from start-ups.

Public Relation   People doing public relation show more interest in politics and social events. They also have strong sense of time, according to the high usage of temporal words in their tweets.

The five most positively distinguishing words and phrases to this occupation are *pr*, *#pr*, *pr pos*, *public relations*, and *press releases*, indicating the job content clearly. Not surprisingly, words related to social events, such as *anniversary*, *super bowl*, and *crisis* are more frequently used by them. They particularly prefer temporal words and phrases, such as *a.m.*, *p.m.*, *of the year*, *monday*. This is probably because time is a crucial factor in public relation. The high positive score of exclamation mark (*!*) implies relatively strong tone of their tweets. This group of people barely mention words about techniques. The top 200 negatively scored words of this job have a large overlapping with the top 200 most positively scored words of software engineers.

The two most positively correlated topics to this group people are formed by politics and social events vocabularies (*hillary*, *demdebate*, *#boston*, *ebola*), and emojis, respectively. The third topic is about social media platforms (*#socialmedia*, *#facebook*, *#twitter*, *#pinterest*, *#instagram*), indicating the application of social medias in the field of of public relation. The most negatively scored topic is about programming languages and tools, which agrees with the negatively scored words and phrases.

Designer   Designers use more visual-related words and compliment words, while they show less interest in business. They also express special interest in New York City.

The three most positively correlated words to designers are *illustrator*, *designer*, and *graphic*. Colors (*blue*, *red*, *black and white*) and words related to graphic designing (*font*, *logo*, *icon*) are more frequently mentioned by them. The name of a community where designers and photographers share their works, *behance*, is also positively scored. Compliment words like *cute*, *nice*, *sweet* are more frequently used by designers than people with other jobs. According to the most negatively correlated words, designers have less interest in business (*benefits*, *inverstment*, *strategies*), and companies (*ceo*, *colleagues*, *leadship*).

The most concerned topic to designers is about their job contents. The most representative words of this topic are *typography*, *fonts*, *lettering*. The second topic is mixed with emoticons, and social media platform names. In the third most positively scored topic, besides some emoticons, ex-
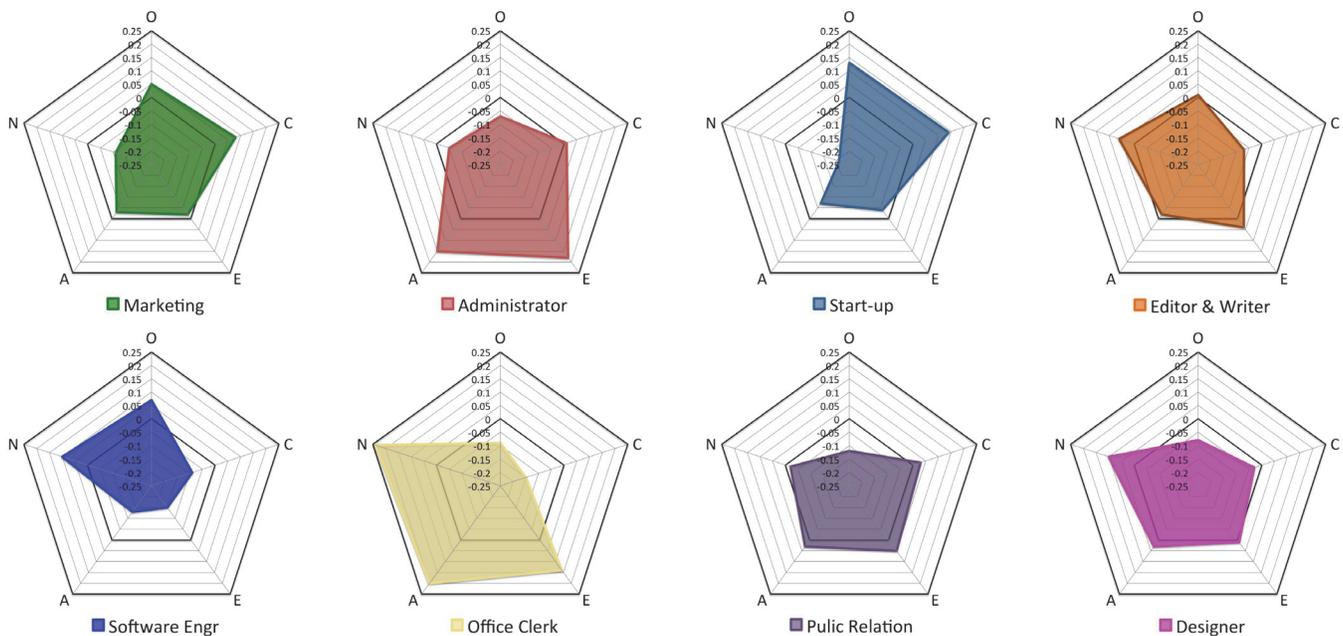
Figure 3: Radar plots of Pearson correlation coefficients between personality traits and each job, where O, C, E, A, N stands for Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, respectively. All the correlations are significant with a p-value lower than 10^{-4}.

pressions about New York City (*nyc*, *#ny*, *ny*) are also highly weighted. We believe this is because the city has special meaning to designers. As to the negatively scored topics, the tops three are about politics and social event (*hillary*, *demdebate*, *ebola*, *#sotu*, *feminist*), marketing (*seo*[9], *#marketing*, *#contentmarketing*), and business (*startups*, *league*, *silicon*), respectively.

It is effective to learn personality from language styles (Mairesse et al. 2007). Illuminated by the difference of language styles existing across jobs, we expect to observe distinctive personality traits. In the next section, we focus on the divergences of personality of people with different jobs.

Personality Analysis

In this section, we study the personality features of the people of different jobs. Previous work shows that people's personalities can be calculated from their tweets (Qiu et al. 2012). We apply the IBM Watson Personality Insights service API[10] to compute the personality traits from ones' tweets. The input is all the tweets of a person, and the service analyzes the linguistic features to infer personality, including Big Five, Needs, and Values. Due to the length limit, we only discuss the Big Five Personality Traits in this paper. Big Five is a widely examined theory of five broad dimensions describing human personality (Goldberg 1993). The five dimensions are:

- *Openness*: measures how open a person is to unusual ideas, imagination, curiosity, and variety of experience.

---

[9] meaning: Search Engine Optimization
[10] www.ibm.com/smarterplanet/us/en/ibmwatson/

In other words, a higher openness indicates a higher acceptance to new things and changes.

- *Conscientiousness*: reveals a person's self-discipline. People of higher conscientiousness tend to act in an organized and thoughtful way.

- *Extraversion*: indicates the extent to which a person prefers or enjoys being in social situations or interactions with the outside, and have company of others.

- *Agreeableness*: reflects if a person feels comfortable about compromising. Higher agreeableness implies being more cooperative toward others.

- *Neuroticism*: measures the instability of a person's emotions. It is usually easier for a person of higher neuroticism to experience negative emotions.

Our results reveal clear distinctive levels of the Big Five Traits of different jobs. In Figure 3 we report the Pearson correlation coefficients between personality traits and each job in form of radar plot. All the correlations are significant with a p-value smaller than 10^{-4}.

Among the 8 groups, people from start-ups are most open to new things, with a positive correlation of 0.13 to Openness. On the contrary, people of public relation are the most conservative. The correlation to Openness of them is -0.12. Software Engineers and marketing people are also open to new things, while office clerks and administrators people are the opposite.

Marketing people, people of start-ups have higher conscientiousness levels. In other words, they show high motivations in their work. Office clerks have the lowest conscien-

tiousness level. This agrees with our observation that office clerks post fewer content about their work. Moreover, software engineers, editors and writers, and designers also show a relatively low self-discipline.

Administrators are the most extroverted. A high positive correlation (0.18) to Extraversion indicates their strong preference for being in social situations. By contrast, the high negative correlation to Extraversion (-0.15) of software engineers shows low preference for social situations.

Office clerks are the most agreeable group of people. They have a positive correlation to Agreeableness of 0.20. Administrators also have a high Agreeableness level (0.18). Software engineers are the least agreeable. A negative correlation of -0.13 shows their low willingness to compromise.

The people from start-ups have highest stability level of emotion (-0.21 correlated to Neuroticism), followed by marketing people (-0.11). By contrast, a 0.24 correlation to Neuroticism of office clerks indicates they have relatively unstable emotions. Software Engineers and Designer also have positve correlations.

Occupation Prediction

Motivated by the observations on the correlation between human traits and occupations, we build a classifier to predict jobs based on features extracted from tweets. Each person is assigned a normalized weight on each job. We label the person with the job of the highest weight if the weight is larger than 0.8. If none of the weights is larger than 0.8, we remove the person from our data set. In each prediction task, we collect the people from a job as positive data, and then sample a equal number of people from other jobs as negative data. We tried 4 sets of features: LIWC features, 2,000 tweets topics, words and phrases, and all above. The values of words and phrases features are the TF-IDF values calculated for terms based on their frequencies in ones' tweets. The values of tweet topic features are the weights that each person is assigned on the topics. We plot the results in Figure 4. The results are evaluated using 5-cross validation.
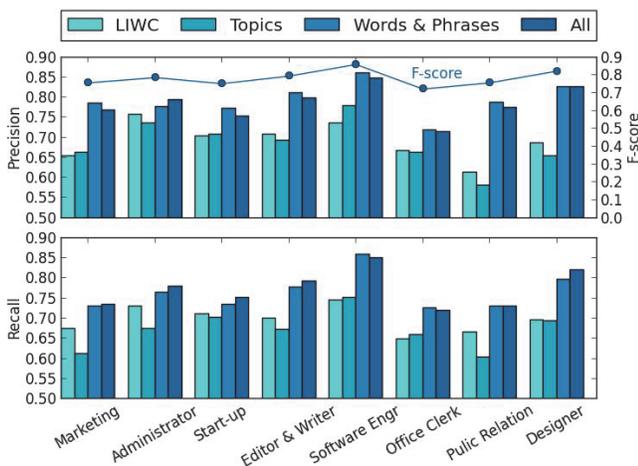

Figure 4: Precisions and recalls of predicting 8 jobs using 4 sets of features. F-scores of using all features is also plotted.

The average F-score of all eight jobs is 0.78, indicating the strong discrimination power of language features on job prediction tasks.

Software engineers (precision = 87%, recall = 86%) and Designers (precision = 82%, recall = 83%) have better prediction results. This is because software engineers and designers have more distinguishing language patterns and unique interests. People with both jobs talk much more about techniques, such as programming languages, online tools, while these technique terms are seldom used by people with other jobs. Because of the same reason, editors also have a relatively high precision (81%) and recall (79%). They mention a lot about reading and writing, and show stronger interest in politics and social events. Office clerks have relatively low results (precision = 73%, recall = 73%). This is because although these group of people have unique preferences (strong self-awareness word like *i*, *my*, *my life*), these features are also usually used by other people. Marketing, Administrator, Start-up, and Public Relation, have median performances, with precisions around 78% and recalls around 75%. This is because some language patterns and interests are shared among these four jobs. For example, people with these jobs are all interested in business, companies, and marketing. The prediction results align with the fact that Software Engineers and Designers are people whose skills are more specialized, whereas Marketing, Administrator, Start-up, and Public Relation require skills that are less specialized but more holistic, especially for the case of office clerk.

Words and phrases related features have the best performance among the three single type features. This is because words and phrases are the most informative type of features among three. They cover all the terms used in tweets with the cost of high dimensions of features. The lexicon LIWC uses is fixed and relatively small, while topic features are based on word clusters. Although these two types have fewer dimensions, the information is also reduced. When combining the three together, we observe a small boost of performance, especially on recall. This suggests that extra information brought by LIWC and topics is helpful the in classifying task.

Conclusions and Future Work

In this paper, we investigate the divergences across occupations. From multiple platforms, we gathered user information of several aspects. To overcome the ambiguity and uncertainty of self-reported information, a soft clustering approached was applied to extract occupations from crowdsourcing data. Linguistic styles were described using the most positively and negatively correlated words and phrases to people, while people's interests are learned by extracting significant topics from their tweets. The Big Five Traits are also inferred using the tweet texts. We used Person Correlation Coefficients to uncover the differences of above human characteristics across jobs. The results indicate that people with different jobs have unique preferences to certain language styles and interests. Our results also reveal clear divergent levels of the Big Five Traits of different jobs. A classifier was built to predict people's job based on the features

extracted from tweets. A high accuracy indicates the strong discrimination of language features on job prediction tasks.

Overall, our study has revealed interesting patterns of how individuals' characteristics and daily tweets are connected to their job profiles. By showing the similarities and dissimilarities between job categories based on those characteristics and communication styles, our findings suggest that some individuals can possess set of characteristics that fit with multiple jobs, while others might possess a set of characteristics that are more unique to one specific type of job. In the future, we would like to introduce more features to categorize people into different job profiles, such as using their job titles and looking into their employment histories. We are also interested in extracting more human characteristics, besides personality and use of language, from their tweets; for examples, their life styles, habits, leisures. By obtaining richer information from people's online profiles, a more comprehensive study could be performed to uncover the deeper connection between people's personal lives and their jobs.

Acknowledgment

References

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; Verkamo, A. I.; et al. 1996. Fast discovery of association rules. *Advances in knowledge discovery and data mining* 12(1):307–328.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Bradley, J. R., and Cartwright, S. 2002. Social support, job stress, health, and job satisfaction among nurses in the united kingdom. *International Journal of Stress Management* 9(3):163–182.

Golbeck, J.; Robles, C.; Edmondson, M.; and Turner, K. 2011. Predicting personality from twitter. In *PASSAT and 2011 IEEE SocialCom*, 149–156. IEEE.

Goldberg, L. R. 1993. The structure of phenotypic personality traits. *American psychologist* 48(1):26.

Hurtz, G. M., and Donovan, J. J. 2000. Personality and job performance: the big five revisited. *Journal of applied psychology* 85(6):869.

Judge, T. A.; Heller, D.; and Mount, M. K. 2002. Five-factor model of personality and job satisfaction: a meta-analysis. *Journal of applied psychology* 87(3):530.

Lindquist, T. L.; Beilin, L. J.; and Knuiman, M. W. 1997. Influence of lifestyle, coping, and job stress on blood pressure in men and women. *Hypertension* 29(1):1–7.

Mairesse, F.; Walker, M. A.; Mehl, M. R.; and Moore, R. K. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research* 457–500.

Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. " how old do you think i am?"; a study of language and age in twitter. In *ICWSM*. AAAI Press.

Payne, N.; Jones, F.; and Harris, P. 2002. The impact of working life on health behavior: the effect of job strain on the cognitive predictors of exercise. *Journal of occupational health psychology* 7(4):342.

Pennebaker, J. W., and King, L. A. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology* 77(6):1296.

Qiu, L.; Lin, H.; Ramsay, J.; and Yang, F. 2012. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality* 46(6):710–718.

Quercia, D.; Kosinski, M.; Stillwell, D.; and Crowcroft, J. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *PASSAT and 2011 IEEE SocialCom*, 180–185. IEEE.

Reichel, A., and Pizam, A. 1984. Job satisfaction, lifestyle and demographics of us hospitality industry workersversus others. *International Journal of Hospitality Management* 3(3):123–133.

Salgado, J. F. 1997. The five factor model of personality and job performance in the european community. *Journal of Applied psychology* 82(1):30.

Schau, H. J., and Gilly, M. C. 2003. We are what we post? self-presentation in personal web space. *Journal of consumer research* 30(3):385–404.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9):e73791.

Silva, T. H.; de Melo, P. O.; Almeida, J.; Musolesi, M.; and Loureiro, A. 2014. You are what you eat (and drink): Identifying cultural boundaries by analyzing food & drink habits in foursquare. *arXiv preprint arXiv:1404.1009*.

Staiano, J.; Lepri, B.; Aharony, N.; Pianesi, F.; Sebe, N.; and Pentland, A. 2012. Friends don't lie: inferring personality traits from social network structure. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, 321–330. ACM.

Strully, K. W. 2009. Job loss and health in the us labor market. *Demography* 46(2):221–246.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29(1):24–54.

Vinciarelli, A., and Mohammadi, G. 2014. A survey of personality computing. *Affective Computing, IEEE Transactions on* 5(3):273–291.

Voss, M.; Floderus, B.; and Diderichsen, F. 2004. How do job characteristics, family situation, domestic work, and lifestyle factors relate to sickness absence? a study based on sweden post. *Journal of Occupational and Environmental Medicine* 46(11):1134–1143.