Unraveling implicit knowledge in information technology jobs

By: Yuzhang Han and Prashant Palvia

Han, Yuzhang; Palvia, Prashant. 2019. Unraveling implicit knowledge in information technology jobs. 25th Americas Conference on Information Systems, AMCIS 2019.

Abstract:

Job seekers are used to looking at job postings published on the main websites like Glassdoor and Google Jobs. Typically, an online job posting provides a piece of text that describes the job in a more qualitative way. Most job seekers, who would have to view hundreds of postings every day, tend to pay attention to the explicit information exposed by the textual description, such as required skills, salary, and benefits, which are information that the author wishes to convey to the job seekers directly. However, this would lead to overlook of a large part of implicit information which is hidden deeper in the linguistic characteristics of the textual description, such as readability of the text, status of the employer, and domain-unrelated concerns of the text. These implicit aspects of the job description can give job seekers knowledge into the job culture and personal characteristics of future colleagues, helping them to prepare for job interviews more efficiently, and integrate into future job environment more smoothly. Using text mining methods, this study extracts various types of implicit information/knowledge of a collection of more than 24 thousand job postings and depicts the implicit characteristics of IT-related jobs compared to non-IT jobs. Analysis results show that IT-related and non-IT job descriptions have distinct profiles in terms of implicit characteristics.

Keywords: IT-related jobs | implicit knowledge | non-IT jobs | job postings | text mining | readability | subjectivity | sentiment | emotion | speech act

Article:

***Note: Full text of article below**

Unraveling Implicit Knowledge in Information Technology Jobs

Completed Research

Yuzhang Han
Bryan School of Business and Economics
University of North Carolina at Greensboro
y_han5@uncg.edu

Prashant Palvia
Bryan School of Business and Economics
University of North Carolina at Greensboro
pcpalvia@uncg.edu

Abstract

Job seekers are used to looking at job postings published on the main websites like Glassdoor and Google Jobs. Typically, an online job posting provides a piece of text that describes the job in a more qualitative way. Most job seekers, who would have to view hundreds of postings every day, tend to pay attention to the *explicit* information exposed by the textual description, such as required skills, salary, and benefits, which are information that the author wishes to convey to the job seekers directly. However, this would lead to overlook of a large part of *implicit* information which is hidden deeper in the linguistic characteristics of the textual description, such as readability of the text, status of the employer, and domain-unrelated concerns of the text. These implicit aspects of the job description can give job seekers knowledge into the job culture and personal characteristics of future colleagues, helping them to prepare for job interviews more efficiently, and integrate into future job environment more smoothly. Using text mining methods, this study extracts various types of implicit information/knowledge of a collection of more than 24 thousand job postings and depicts the implicit characteristics of IT-related jobs compared to non-IT jobs. Analysis results show that IT-related and non-IT job descriptions have distinct profiles in terms of implicit characteristics.

Keywords

IT-related jobs, implicit knowledge, non-IT jobs, job postings, text mining, readability, subjectivity, sentiment, emotion, speech act.

Introduction

Job postings published on large websites like Glassdoor and Google Jobs are a main source of information for job seekers. Typically, a job posting contains not only structured information, such as attributes of job title, location, and range of salary, but also a piece of textual description of the job, which is qualitative and unstructured. Thanks to the fast development of text mining techniques, people are able to extract plenty of information from unstructured text created by human. At the lowest level, linguistic information, such as word use and sentence structure, can be mined from raw text based on statistics of the text. Various types of semantic information can be further extracted, such as topics, author's subjectivity, sentiment and emotion, and text readability. So far, numerous text mining applications have been built utilizing semantic information to create knowledge about the text. Such applications include document classification, keyword identification, topic extraction, text generation and so on. In the context of understanding job postings, text mining has been used to extract useful information from text (Mooney and Bunescu 2005) and predict requirements of the job market (Karakatsanis et al. 2017).

We propose to classify the information in job description into two types, *explicit* and *implicit*. Explicit information refers to ideas and opinion the author of job description wishes to express directly in the text. Examples of explicit information are required skills, wage, benefits and responsibilities. Implicit information, on the other hand, refers to the information that the author does not mean to express in text but can be extracted from the text, such as linguistic characteristics and status of the author.

Typically, job seekers tend to pay more attention to the explicit information expressed by job description, through which they can gain answers directly to their questions about the job position. Many methods have been developed to help job seekers obtain explicit information, such as keyword extraction (Matsuo and Ishizuka 2004) and text summarization (Barzilay and Elhadad 1999). However, little attention has been paid to the implicit information/knowledge that can be extracted from the job description. In this research, we examine the implicit information of job description in the context of Information Technology (IT)-related jobs. We explore six types of implicit information: text readability, subjectivity, sentiment, emotion, speech act and domain-unrelated concerns. We demonstrate based on a collection of more than 24 thousand real job postings that IT-related jobs and non-IT jobs are statistically different with respect to these implicit features. Our findings and analysis make contribution for both research community and practitioners. We discover a set of implicit features that differentiate IT-related jobs from non-IT jobs. These features provide information systems researchers with some new perspectives from which they examine IT as a profession, augmenting their current understanding of the IT profession and the IT culture. Meanwhile, since a large part of the implicit features are related to psychology and natural language, these features also help to establish the IT profession as a research subject for psychologists and linguists. In practice, the quality of job postings substantially impacts the efficiency of recruitment and job seeking. Many recruitment efforts fail because job seekers do not hold the right expectation of the jobs, or employers do not express themselves effectively and efficiently. The implicit features can help job seekers understand job positions more comprehensively. Also, employers can also tune the style of the job postings in terms of these features in order to better exhibit their jobs openings and corporate culture. Improved description and understanding of job positions would help companies recruit employees who are truly satisfied and competent in their jobs, leading to enhanced company performance.

The rest of the paper is structured as follows: the next section describes the theoretical framework of the study based on a literature review; the following section introduces the methods for extracting, evaluating and analyzing implicit features; next we describe our analysis results; and subsequent sections provide a discussion of the results together with limitations and suggestions for future work.

Theoretical Framework and Literature Review

Text mining is a long-existing but rejuvenated branch of computer science. Usually, studies of modern text mining first convert unstructured texts created by humans into quantitative representations, such as sets of variables, and try to discover patterns and knowledge by analyzing these representations. Big data and machine learning redefined text mining as an approach to finding patterns in a massive collection of documents by studying the documents carefully. To handle a large amount of documents (called *corpus*), it is crucial to define an efficient representation, which is a set of variables, of the documents. This set of variables used in text mining is called a *feature set*, while each variable is called a *feature*. The text mining community has proposed many different feature sets to satisfy various purpose of mining. For example, the n-gram is the most popular feature set which considers every different set of n successive words in the corpus a feature (Brown et al. 1992). Frequency of each n-gram is considered the value of this feature. The n-gram feature set has been used in many cases, such as automatic identification of topics of documents (Mei et al. 2007) and extraction of semantic biomedical relations from text (Bundschus et al. 2008).

This study focuses on six types of features which are not conveyed in job descriptions directly: text readability, subjectivity, sentiment, emotion, speech act, and domain-unrelated concerns. The first five have been proven effective in either revealing possibility of events or characterizing a particular group of people. *Readability* of text is a measure of the difficulty with which a reader can understand the text. It depends on the complexity of vocabulary and syntax of the sample, focusing on word choice of the sample and how words are organized into sentences and paragraphs. Text readability has been used to assess the helpfulness of online customer reviews (Cao et al. 2011) and predict behavior of financial markets (Nassirtoussi et al. 2014). *Subjectivity* is a measure of the extent to which the author expresses own feelings and opinions rather than describing facts. It has been used in many applications such as predicting consumer's attitude towards brands (Mostafa et al. 2013) and picturing media behavior during political crisis (De Fortuny et al. 2012). *Sentiment* is a widely used feature that evaluates if a text sample expresses positive or negative attitude of the author. It has been used for online forum hotspot detection (Li and Wu 2010) and stock movement prediction (Nguyen et al. 2015). *Emotion* (such as happy, sad, anger and fear) can be seen as a finer version of sentiment, which has been gaining increasing interest. It has been used for

evaluating improvement in mental and physical health of traumatic people (Pennebaker et al. 1997) and evaluating user attitude to medical software (Bekker et al. 2003). *Speech act* is defined as an utterance that is considered as an action of author, particularly with regard to the author's intention, purpose, or effect. It evaluates the psychological gesture of the author, typically including features such as negation, tentative, certain, and exclamation. Speech act features have been used to recognize personality of the author (Mairesse and Walker 2006) and predict group project performance (Yoo and Kim 2014). These features are originally addressed in linguistics and psychology separately. In this study, they help us characterize IT-related jobs and contrast them with non-IT jobs. To our best knowledge, these have not been combined to characterize the writing behavior of a particular profession. Furthermore, the linguistic and psychological features of IT professions have not been fully explored yet.

Apart from those widely examined features, we also inspected the *domain-unrelated concerns* of job description. In our study, domain-unrelated concerns of job description refer to non-IT issues included in job description. For example, in a posting for a software engineer position, issues such as driver license and equal promotion opportunity should be counted as domain-unrelated concerns. This type of information is considered implicit because they might not be emphasized by the author and often ignored by job seekers, but at the same time they provide crucial information for job seekers.

We define IT-related jobs as jobs potions that are engrained in development, implementation and management of IT artifacts, including computer software and hardware, telecommunication devices, and any techniques involved. Many prior studies have demonstrated that IT-related jobs are different from non-IT jobs in many respects. An early study argued that IT profession has its own "durable domain of human concern", "codified body of principles", "codified body of practices", and "standards for competence, ethics, and practice" (Denning 2001). A follow-up study examined the special attitude of the IT industry towards women (Trauth 2002). Later, it was discovered that IT professionals' careers might follow multiple different career paths (Joseph et al. 2005). More recently, it was pointed out that IT-related jobs require particular competency in a broad range of skills, including non-technical ones (Gallagher et al. 2010). For the first time, our study highlights the characteristics of IT job employers in relation to their attitude and personality as compared to non-IT jobs. These characteristics are reflected in the way their companies compose job description.

Overall, Figure 1 depicts the conceptional model of the study. The type of job (in our study, IT-related or non-IT) influences (or is reflected by) the explicit information expressed by job descriptions, such as required skills and wage. At the same time, the type of job also influences (or is reflected by) the implicit information contained in job descriptions, such as text readability, subjectivity, sentiment, emotion, speech act, and domain-unrelated concerns of the text. The main research questions of the study are: (1) Are IT-related jobs different from non-IT jobs in relation to text readability, subjectivity, sentiment, speech act, and domain-unrelated concerns of the job description in job postings? (2) How exactly are IT-related jobs different in these implicit dimensions of the job description?

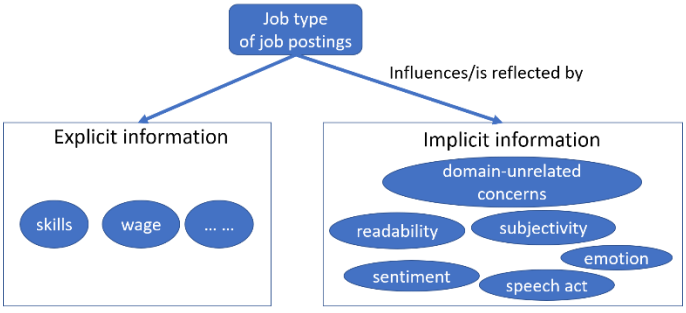

Figure 1. Conceptional Model

Methodology

Data Set

This study utilizes a secondary data collection of 24,143 job postings (after preprocessing) published by more than 300 U.S. based companies on JobsPikr.com (the largest accessible data set of such kind within

our funding limit), a job data delivery platform that "extracts job data from various company sites across the globe on daily basis" (https://www.jobspikr.com/). The original data contains several data fields, such as URL of the posting, company name, location, job title, job description and last date to apply. During preprocessing, all useless postings without job description or with the same company name, job title and job description were removed. Then, only two fields, job title and job description, were kept. Next, a key word search was done through job titles of all the postings. Postings are labeled as IT-related jobs if their job titles contain any IT-related keywords. Some keywords we employed are: names of programming languages, well-known IT-related jobs titles such as software engineer and hardware architect, IT-related tools such as database and operating system, and IT skills such as programming and machine learning.

Readability

This study employs five most popular metrics of text readability (Meade and Smith 1991). The Flesch Reading Ease test is a readability test that indicates how easy a passage in English can be read. A higher score indicates that material is of higher readability, namely easier to read and understand. Shorter words (smaller number of syllables per words) and shorter sentences (smaller number of words per sentences) are two factors that result in higher readability. In comparison, the Dale–Chall readability metric evaluates the comprehension difficulty of text readers. The metric was built based on a list of 763 words that 80% of sampled fourth-grade American students could understand. Words beyond that list are considered difficult. A text sample with more difficult words (words beyond the list) and longer sentences would be evaluated as more difficult.

The rest metrics--Flesch–Kincaid Grade Level metric, Gunning Fog index and Simple Measure of Gobbledygook (SMOG)--all represent the readability level of various books and texts by estimated number of years of U.S. school education required to comprehend them. With the first two metrics, the estimated grade level is higher when there are more longer sentences and longer, more complex words in the text sample. In contrast, SMOG, which was used for evaluating health messages initially, introduces the number of polysyllables (words of 3 or more syllables) as a factor. Higher grade levels are mainly associated with having more polysyllables per sentence.

Overall, while Flesch Reading Ease test gives higher scores for easier samples, all the other metrics give higher scores for more difficult samples. It can be seen that these metrics model difficulty of reading on different factors and express the result in different ways. To ensure rigor of analysis, we computed and analyzed all the five readability scores on the data set.

Subjectivity and Sentiment

Subjectivity of a text indicates the extent to which the author expresses own feeling and opinion rather than describing facts. Sentiment of text refers to the attitude of author regarding like and dislike, or agreement and disagreement. In our study, we evaluated the subjectivity and sentiment of all the job descriptions using the algorithms provided in the Pattern library (Pattern Library 2019) which examine the given text for occurrence of common adjectives (e.g., good, bad, amazing, irritating, …). The produced subjectivity score of a job description is a real number in [0.0, 1.0], where 0.0 represents a fully objective attitude, and 1.0 representing fully subjective. The sentiment score is a real number in [-1.0, 1.0], where -1.0 represents a fully negative attitude, 0.0 representing neutral, and 1.0 representing fully positive.

Emotion

Emotion evaluates the writer's emotional status expressed in the text. There are several different approaches to evaluating emotion (Bao et al. 2010; Liu 2012; Ren and Quan 2012). In our study, we employ the emotion scores provided by the LIWC2015 software. Linguistic Inquiry and Word Count (LIWC) is one of the most popular text analysis software used to identify linguistic, emotional, cognitive, social and other characteristics of individuals' written text samples (Pennebaker et al. 2015). A team of linguists and psychologists were engaged to select 6,400 words, word stems and emoticons from standard dictionaries that are most related to relevant characteristics. They were converted into a hierarchy of 83 word categories. Given a sample of text, the percentage of words in these categories are computed and used as scores for measuring the interested characteristics. For example, LIWC can find that there are 84 positive emotion words in a text sample with 2000 words. It will then give this sample 4.2% as the score of positive emotion.

Our study employs the 6 emotion-related scores produced by LIWC2015, including emotional tone, positive emotion, negative emotion, anxiety, anger, and sadness. Score of emotional tone represents how emotional the text sample is, in general. The other scores represent individual types of emotion expressed in the text. A list of example words for the LIWC word categories utilized in our study can be found in the official guide of LIWC2015 (Pennebaker et al. 2015). To provide a few examples, the category positive emotion includes 620 words such as love, nice and sweet, while the category negative emotion includes 744 words such as hurt, ugly and nasty.

Speech Act

Speech act of a text sample is the linguistic gesture expressed, reflecting the purpose and attitude the author wishes to convey. Speech acts of text can be such as requests, warnings, invitations, questions, assertive, commissive, expressive, declarative, directive, promises, apologies, predictions, and the like (Millikan 1998; Mulligan 1987; Schuhmann and Smith 1990; Abbasi et al. 2018). A single text sample can have multiple speech acts. While there are several models of speech act, we employ the speech act scores provided by LIWC2015, which includes six variables: analytic, negation, tentative, certain, exclamation, and question. Each of these scores is the relative frequency of keywords in corresponding word category as described later in the results section.

Domain-unrelated Concern

In our study, domain-unrelated concerns of job description refer to non-IT issues included in the job description. For example, in a posting for a software engineer position, issues such as driver license and health should be counted as domain-unrelated concerns. These characteristics are labeled implicit because they might not be emphasized by the posting author and often ignored by job seekers. However, failing to satisfy such concerns might impact the job opportunity of the job seekers.

We use two methods to extract those domain-unrelated concerns. First, scores of five LIWC word categories are computed, which are unrelated to IT but important for job seekers. They are power, reward, risk, work, and money. Second, we use topic modeling to extract domain-unrelated concerns as topics. Topics of a sample refer to the main issues discussed by the text. In the context of text mining, topics are represented by sets of words that occur in the sample and also co-occur frequently in other text samples in the entire corpus. The most popular technique of topic modeling, Latent Dirichlet Allocation (LDA), is employed. This algorithm finds out a specified number of topics from a document corpus, each being a collection of words (Blei et al. 2003).

Analysis and Results

Two Sample Student's t-test

The study aims at showing how job description of IT-related jobs is characterized in implicit dimensions contrasted to non-IT jobs. Two-sample Student's t-test was performed on each implicit variable comparing the means of IT-related jobs and non-IT jobs. p-value and confidence interval of each test were used to decide if the difference between means of two groups was significant. This test is the classic method in many areas for contrasting two samples of unequal sizes (Kazerooni et al. 1996; Kleijnen 1999). In our analysis a 95% confidence level was employed.

Descriptive Statistics

Descriptive statistics of the preprocessed data set can be found in Table 1. We can see that while there are much less IT-related jobs than non-IT jobs (1:13.7), the number of postings of each class is large (IT: 1,640, non-IT: 22,503). IT-related jobs tend to create a much smaller part of job postings per company compared to non-IT jobs (IT: 14.14 postings per company, non-IT: 70.76 postings per company) – that is because non-IT jobs include all other jobs.

On average, job descriptions of IT-related jobs are written shorter than non-IT jobs (IT: 311.99 words, non-IT: 406.59 words, IT 23.27% shorter). But IT-related jobs tend to use longer sentences in job description

(IT: 39.79 words/sentence, non-IT: 32.48 words/sentence, IT 22.51% longer). It follows that IT-related jobs have fewer sentences than non-IT jobs (IT 37.33% fewer sentences).

	IT	Non-IT	IT to non-IT ratio
#postings	1,640	22,503	1:13.7
#companies	116	318	1:2.74
#words/job desc.	311.99	406.59	1:1.30 (IT 23.27% less)
#words/sentence	39.79	32.48	1:0.81 (IT 22.51% more)
#sentences/job desc.	7.84	12.51	1:1.60 (IT 37.33% less)

Table 1. Descriptive statistics

Readability

Table 2 shows t-test results regarding different metrics of various implicit feature sets. Each two-sample t-test tests the significance of difference between means of non-IT jobs and IT-related jobs (non-IT mean subtracted by IT-related mean). The first column is the category of features. The second column is the p-value. The third column shows the confidence interval of the mean difference. The fourth column are these bounds expressed as percentage over the mean of non-IT jobs. The fifth column shows the increase of mean of IT-related jobs compared to non-IT jobs, expressed as percentage over mean of non-IT jobs. The last two columns show the mean of non-IT jobs and IT-related jobs, respectively. To provide an example, the first row of Table 2 presents the two-sample t-test result testing the mean Flesh Reading Ease (FRE) score of non-IT jobs subtracted by mean FRE of IT-related jobs. The p-value is extremely low, indicating that the difference of means is significant. The confidence interval of this difference is [13.68, 17.50], or [45.01%, 57.58%] with respect to the non-IT mean. IT-related mean is lower than non-IT mean by 51.31% with respect to non-IT mean. Finally, the non-IT mean is 30.394, and the IT-related mean is 14.800.

We can see in Table 2 that all the five readability scores are significant, while job descriptions of IT-related jobs are more difficult to read than those of non-IT jobs. The Flesh Reading Ease score suggests that the text of IT-related jobs are 51.31% harder than non-IT jobs. The Dale-Chall score suggests that text of IT-related jobs are 12.93% more difficult than text of non-IT jobs. Further, the Flesh-Kincad, Gunning Fog, and SMOG scores all suggest that text of IT-related jobs require more years of education to understand than text of non-IT jobs. However, the IT-nonIT difference of SMOG (15.99%) is much lower than the difference with Flesch-Kincaid (25.62%) and Gunning Fog (22.12%). This might be because the text of IT-related jobs has more words per sentence (see Section Descriptive Statistics), which is a factor that increases text difficulty with Flesch-Kincaid and Gunning Fog; but words per sentence is not considered in SMOG.

Category	Feature	p-value	CI(nonIT–IT)	CI ptg nonIT	Inc IT ptg	Mean nonIT	Mean IT
Readability	Flesch Reading Ease	1.81E-53	[13.68, 17.50]	[45.01, 57.58]	-51.307	30.394	14.800
	Flesch-Kincaid	1.11E-35	[-4.69, -3.44]	[-29.55, -21.67]	25.619	15.874	19.941
	Gunning Fog	7.02E-38	[-4.92, -3.65]	[-25.39, -18.84]	22.119	19.375	23.661
	Dale-Chall	9.61E-82	[-1.69, -1.40]	[-14.14, -11.72]	12.929	11.950	13.495
	SMOG	3.66E-50	[-2.93, -2.27]	[-14.14, -11.72]	15.993	16.275	18.878
Subjectivity	Subjectivity	2.27E-49	[0.03, 0.04]	[6.82, 9.09]	-9.036	0.440	0.400

Sentiment	Sentiment	2.56E-22	[0.02, 0.03]	[10.70, 16.04]	-11.756	0.187	0.165
Emotion	Emotional tone	1.75E-30	[5.98, 8.39]	[7.69, 10.79]	-9.238	77.761	70.578
	Positive emotion	4.42E-40	[0.57, 0.76]	[14.59, 19.45]	-17.030	3.908	3.243
	Negative emotion	1.95E-08	[-0.12, -0.06]	[-33.61, -16.81]	24.468	0.357	0.444
	Anxiety	0.391	[-0.01, 0.02]	[-13.33, 26.67]	-7.352	0.075	0.070
	Anger	0.056	[-0.07, 0.00]	[-159.09, 0.00]	18.680	0.044	0.052
	Sadness	0.157	[-0.00, 0.01]	[0.00, 32.26]	-17.671	0.031	0.025
Speech act	Analytic	8.22E-36	[-4.46, -3.27]	[-5.00, -3.67]	4.334	89.173	93.038
	Negation	1.16E-81	[0.14, 0.17]	[46.98, 57.05]	-50.762	0.298	0.147
	Tentative	0.025347	[-0.16, -0.01]	[-10.28, -0.64]	5.613	1.556	1.643
	Certain	6.68E-56	[0.66, 0.84]	[38.04, 48.41]	-43.157	1.735	0.986
	Exclamation	1.14E-219	[0.12, 0.13]	[83.92, 90.91]	-87.688	0.143	0.018
	Question	9.50E-06	[-1.94, -0.75]	[-979.80, 378.79]	682.125	0.198	1.545
Domain-unrelated concern	Power	0.014	[-0.24, -0.03]	[-5.99, -0.68]	3.336	4.077	4.213
	Reward	7.22E-13	[0.15, 0.26]	[9.29, 16.11]	-12.711	1.614	1.408
	Risk	0.072	[-0.00, 0.08]	[0.00, 13.51]	-6.838	0.592	0.551
	Work	0.004	[-0.60, -0.12]	[-4.43, -0.89]	2.660	13.539	13.899
	Money	1.24E-119	[1.14, 1.34]	[42.70, 50.19]	-46.422	2.670	1.430

Table 2. Readability, Subjectivity and Sentiment

Subjectivity and Sentiment

Table 2 also shows the subjectivity and sentiment scores of the data. It can be seen that the text of IT-related jobs is 9.04% more objective (or less subjective) than the text of non-IT jobs. This is compliant with the general impression that IT practitioners spend more time in dealing with technology, doing more technical readings and writing more technical documents. More description of technology-related details might have enhanced the objectivity of their writing. As a support to this assumption, we can see that although the sentiment scores of both groups are close to neutral (0.187 and 0.165 compared to 0.0), the textual sentiment of IT-related jobs is even more neutral (IT is 11.76% lower than non-IT). This, again, reflects the general impression that IT-related writings are more technical and more neutral in stance.

Emotion

Table 2 shows the emotion scores of the data. The score of general emotional tone of IT-related jobs is lower than that of non-IT jobs (IT is 9.24% lower than non-IT). This indicates that text of IT-related jobs is less

emotional than non-IT jobs, which echoes observations in Subsection Subjectivity and Sentiment. Further, the positive emotion score is lower for IT-related jobs (IT is 17.03% lower than non-IT), while the negative emotion score is higher for IT-related jobs (IT is 24.47% higher than non-IT). It can be seen that the anger score agrees with this observation: anger score of IT-related jobs is higher than that of non-IT jobs (IT is 18.68% higher than non-IT). There is no easy explanation for this finding and future research may want to explore this phenomenon further. Although the anxiety score and sadness score seem to reflect the opposite, both scores are not significant statistically and can be ignored.

Speech Act

Table 2 shows the speech act scores. It can be seen that IT-related jobs are slightly higher than non-IT jobs in analytic score (4.33% higher) and tentative score (5.61% higher). Meanwhile, IT-related jobs are substantially lower in negation score (50.76% lower), certain score (43.16% lower), and exclamation score (87.69% lower). Finally, the question score of IT-related jobs is prominently higher (682.13% higher). This means that text of IT-related jobs shows uncertainty; it spends more words in analyzing, presenting a more tentative (even questioning) and less certain attitude. As another evidence for the uncertainty, the text of IT-related jobs is much more reluctant in negating something or expressing strong attitude by exclaiming.

Domain-unrelated Concern

Table 2 shows the scores of domain-unrelated word categories. It can be seen that IT-related jobs are slightly higher than non-IT jobs in power score (3.34% higher) and work score (2.66% higher), but significantly lower in reward score (12.71% lower) and money score (46.42% lower); the risk score is ignored for statistical insignificance. That means the job description of IT-related jobs talks more about work and power structure in working environment, but pays less attention to rewards and money.

Finally, the LDA algorithm was run to produce the top ten most significant topics in IT-related jobs, each with ten words. It is observed that four of these topics (40%) do not contain any IT-specific words, as shown in Figure 2. Tentative interpretation is made to speculate the domain-unrelated concerns expressed by these topics. Topic 1 includes words such as gender, opportunity, and disability. The topic might reflect the concern of equal job opportunity. Topic 2 includes words such as business, experience, management, skill, project, and process. These words might imply the concern about capability of business and project management. Topic 3 includes words such as marketing, campaign, loyalty, Italian, region, southern, and Hungary. There might be a concern about international marketing. Topic 4 includes words like experience, business, customer, analytics, skill, and ability. They might suggest the concern about capability of business analytics.

Topic 1	Topic 2	Topic 3	Topic 4
Words: cognizant, global, gender, provide, opportunity, disability, business, market, decision, employment	Words: business, experience, analytics, support, ability, customer, management, skill, project, process	Words: marketing, billing, analytics, campaign, loyalty, italian, region, digital, southern, hungary	Words: experience, services, business, technology, customer, solution, analytics, skill, description, ability
Interpretation: equal job opportunity?	Interpretation: business and project management?	Interpretation: global marketing?	Interpretation: business analytics?

Figure 2. Domain-Unrelated Topics

Discussion

The results of this study help us capture several patterns of how IT practitioners describe their jobs. These patterns reflect some important features of IT profession and IT industry. These patterns provide additional knowledge about IT jobs than previously available.

First of all, IT companies make relatively fewer job postings (see Subsection Descriptive Statistics and Table 1). This might imply that IT industry is reaching saturation, or the fluidity of IT employees is lower.

Secondly, IT employers tend to use more complicated language. For example, they write fewer words in total but longer sentences, while their writings are hard to read. This might be because the daily work of IT-related jobs is of higher complexity and difficulty. Thus, IT practitioners have lower sensitivity to complexity and difficulty. On the other hand, it may also mean that the IT employers need to simplify their language to reach a larger audience.

Further, IT employers tend to describe their jobs objectively, neutrally, analytically, tentatively, and less emotionally. On the one hand, this phenomenon correlates with the nature of their jobs which rely on technology, machines and algorithms. Their jobs might transform their linguistic style towards being more analytical and neutral. On the other hand, the IT occupational culture values caution and analytical thinking. This culture might encourage them to adopt the observed style of writing.

IT employers pay more attention to work, but less attention to money and reward. This implies that both employers and employees are relatively more work-centric. They look more at the joy and satisfaction brought by their work, but deem material reward of their work less important. As an evidence, many IT professions are willing to work all day and night, in office and from home. In turn, many IT companies have to improve their recreational and dining facilities to keep their employees' living style healthy.

Finally, IT-related jobs seem to care more about equal job opportunity, skills of management, marketing, and business analytics. This observation implies that IT employers are transitioning from purely valuing the technical capabilities of their employees, to valuing more and more their business capabilities. The role of IT employees is also transitioning from worker to cooperator. This observation bodes well for the alignment of IT and business, which has been a perennial concern for management (Kappelman et al. 2018).

Limitation and Future Work

In this first paper, we limited our analysis to fairly simple statistical tests. For example, the interaction of implicit features was not tested; and regression analysis of implicit features on job type was not conducted. Furthermore, more theoretical justifications need to be provided for supporting the use of implicit features. Preliminary statistical evidence has been found to show that IT-related jobs have special characteristics on implicit dimensions. But psychological and culture theories can provide more support for our findings. On a more ambitious note, our findings can help develop new theories.

In the future, studies can be crafted to help build a decision support system that will assist job seekers in filtering job postings. Both implicit and explicit features should be included and integrated into this system. Besides, a broader conceptual model/theory can be built which involves both explicit and implicit features as well and other constructs, such as environment and IT culture. Their relationships to key outcome variables, such as job performance and satisfaction would be worthwhile to explore.

REFERENCES

2019. "Pattern Library." from https://www.clips.uantwerpen.be/pages/pattern-en

2013. *Speech Act and Sachverhalt: Reinach and the Foundations of Realist Phenomenology*, K. Mulligan (eds.), Springer Science & Business Media.

Abbasi, A., Zhou, Y., Deng, S., and Zhang, P. 2018. "Text Analytics to Support Sense-Making in Social Media: A Language-Action Perspective," *MIS Quarterly* (42:2), pp. 427-464.

Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., and Yu, Y. 2012. "Mining Social Emotions from Affective Text," *IEEE transactions on knowledge and data engineering* (24:9), pp. 1658-1670.

Barzilay, R., and Elhadad, M. 1999. "Using Lexical Chains for Text Summarization," *Advances in automatic text summarization*, pp. 111-121.

Bekker, H. L., Hewison, J., and Thornton, J. G. 2003. "Understanding Why Decision Aids Work: Linking Process with Outcome," *Patient education and counseling* (50:3), pp. 323-329.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of machine Learning research* (3:Jan), pp. 993-1022.

Bundschus, M., Dejori, M., Stetter, M., Tresp, V., and Kriegel, H.-P. 2008. "Extraction of Semantic Biomedical Relations from Text Using Conditional Random Fields," *BMC bioinformatics* (9:1), p. 207.

Cao, Q., Duan, W., and Gan, Q. 2011. "Exploring Determinants of Voting for the "Helpfulness" of Online User Reviews: A Text Mining Approach," *Decision Support Systems* (50:2), pp. 511-521.

De Fortuny, E. J., De Smedt, T., Martens, D., and Daelemans, W. 2012. "Media Coverage in Times of Political Crisis: A Text Mining Approach," *Expert Systems with Applications* (39:14), pp. 11616-11622.

Denning, P. J. 2001. "The Profession of It: Who Are We?," C*ommunications of the ACM* (44:2), pp. 15-19.

Denning, P. J., and Frailey, D. J. 2011. "Who Are We---Now?," *Communications of the ACM* (54:6), pp. 25-27.

Gallagher, K. P., Kaiser, K. M., Simon, J. C., Beath, C. M., and Goles, T. 2010. "The Requisite Variety of Skills for It Professionals," *Communications of the ACM* (53:6), pp. 144-148.
García-Crespo, Á., Colomo-Palacios, R., Gómez-Berbís, J. M., and Tovar-Caro, E. 2008. "The It Crowd: Are We Stereotypes?," *IT Professional* (10:6), pp. 24-27.
Joseph, D., Ang, S., and Slaughter, S. 2005. "Identifying the Prototypical Career Paths of It Professionals: A Sequence and Cluster Analysis," *Proceedings of the 2005 ACM SIGMIS CPR conference on Computer personnel research: ACM*, pp. 94-96.
Kappelman, L., Johnson, V., McLean, E., and Maurer, C. 2018. "The 2017 Sim It Issues and Trends Study," MISQ Exec (17:1), pp. 53-88.
Karakatsanis, I., AlKhader, W., MacCrory, F., Alibasic, A., Omar, M. A., Aung, Z., and Woon, W. L. 2017. "Data Mining Approach to Monitoring the Requirements of the Job Market: A Case Study," *Information Systems* (65), pp. 1-6.
Kazerooni, E. A., Lim, F. T., Mikhail, A., and Martinez, F. J. 1996. "Risk of Pneumothorax in Ct-Guided Transthoracic Needle Aspiration Biopsy of the Lung," *Radiology* (198:2), pp. 371-375.
Kleijnen, J. P. 1999. "Validation of Models: Statistical Techniques and Data Availability," in *Proceedings of the 31st conference on Winter simulation: Simulation---a bridge to the future*-Volume 1: ACM, pp. 647-654.
Li, N., and Wu, D. D. 2010. "Using Text Mining and Sentiment Analysis for Online Forums Hotspot Detection and Forecast," *Decision support systems* (48:2), pp. 354-368.
Liu, B. 2012. "Sentiment Analysis and Opinion Mining," *Synthesis lectures on human language technologies* (5:1), pp. 1-167.
Mairesse, F., and Walker, M. 2006. "Words Mark the Nerds: Computational Models of Personality Recognition through Language," in *Proceedings of the Annual Meeting of the Cognitive Science Society*.
Matsuo, Y., and Ishizuka, M. 2004. "Keyword Extraction from a Single Document Using Word Co-Occurrence Statistical Information," *International Journal on Artificial Intelligence Tools* (13:01), pp. 157-169.
Mei, Q., Shen, X., and Zhai, C. 2007. "Automatic Labeling of Multinomial Topic Models," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining: ACM*, pp. 490-499.
Millikan, R. G. 1998. "Language Conventions Made Simple," *The Journal of Philosophy* (95:4), pp. 161-180.
Mooney, R. J., and Bunescu, R. 2005. "Mining Knowledge from Text Using Information Extraction," ACM SIGKDD explorations newsletter (7:1), pp. 3-10.
Mostafa, M. M. 2013. "More Than Words: Social Networks' Text Mining for Consumer Brand Sentiments," *Expert Systems with Applications* (40:10), pp. 4241-4251.
Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. 2014. "Text Mining for Market Prediction: A Systematic Review," *Expert Systems with Applications* (41:16), pp. 7653-7670.
Nguyen, T. H., Shirai, K., and Velcin, J. 2015. "Sentiment Analysis on Social Media for Stock Movement Prediction," *Expert Systems with Applications* (42:24), pp. 9603-9611.
Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. 2015. "The Development and Psychometric Properties of Liwc2015."
Pennebaker, J. W., Mayne, T. J., and Francis, M. E. 1997. "Linguistic Predictors of Adaptive Bereavement," *Journal of personality and social psychology* (72:4), p. 863.
Ren, F., and Quan, C. 2012. "Linguistic-Based Emotion Analysis and Recognition for Measuring Consumer Satisfaction: An Application of Affective Computing," *Information Technology and Management* (13:4), pp. 321-332.
Schuhmann, K., and Smith, B. 1990. "Elements of Speech Act Theory in the Philosophy of Thomas Reid," *History of Philosophy Quarterly* (7:1), pp. 47–66.
Tausczik, Y. R., and Pennebaker, J. W. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of language social psychology* (29:1), pp. 24-54.
Trauth, E. M. 2002. "Odd Girl Out: An Individual Differences Perspective on Women in the It Profession," *Information Technology & People* (15:2), pp. 98-118.
Yoo, J., and Kim, J. 2014. "Can Online Discussion Participation Predict Group Project Performance? Investigating the Roles of Linguistic Features and Participation Patterns," *International Journal of Artificial Intelligence in Education* (24:1), pp. 8-32.
Meade, C. D., and Smith, C. F. 1991. "Readability Formulas: Cautions and Criteria," *Patient education and counseling* (17:2), pp. 153-158.