Social Media Character Assessment for Talent Selection using Natural Language Processing

Jovelyn C. Cuizon and Kent Ferolino
University of San Jose - Recoletos
jvlync@gmail.com

Abstract

This study aims to use social media data as corpus to assess the person's character to provide a preliminary background check on job seekers. It will provide recruiters an initial assessment of the candidates as well as supplementary information to support traditional recruitment and talent acquisition activities thereby reducing time and cost spent for character investigation. The application uses social media analytics to assign a social profile score. Unstructured text data are preprocessed to include only keywords which are relevant to the analysis. Word sense disambiguation is applied to determine the underlying meaning of the words. The bag-of-word is then checked for occurrence of associated words defined for each factor. Posts containing at least one occurrence of words associated with the factors are further tested for content polarity. Social character score is computed using proposed formula. The system recommends applicants based on skills and uses social character score for relevancy ranking of candidates relative to the job posts.

Keywords: *social media analytics, text mining, NLP*

1.0 Introduction

As social media continues to evolve and become more widespread, it presents an exciting opportunity and significant impact on talent search and the hiring process. Social networking sites (SNS) like Facebook, LinkedIn, Google Plus, and Twitter provide an abundant data on personal and professional characteristics of an individual. It gives anyone a platform to say what they think, express their sentiments in whatever manner they want, and build online presence (Boyd & Ellison, 2008). The profile they paint in social media gives HR tremendous insights and candid supplementary information about individual job seekers thus providing them speed in job processing compared to the conventional hiring process which is usually time consuming, costly, and tedious. Through social media analytics, HR will be able to assess applicant profile and use it as an effective tool to aid in the traditional forms of recruitment.

A study by Hoek, J. et al. (2016) found out that employers who use social

media in the recruitment process primarily used Facebook to identify a candidate's organizational fit while LinkedIn is used to determine their professional attributes and their fit for a job (Hoek, O'Kane, & McCracken, 2016). Kluemper and Rosen (2009) conducted a study examining the feasibility of using candidate's personal information currently available on social network sites to improve the employment screening process. They utilized judge ratings to determine if raters accurately determine personality traits and were able to positively distinguish high and low performers based exclusively on viewing candidate's social network profiles which the judges are consistent in their ratings across subjects (Kluemper & Rosen, 2009). Prescreening of applicants using social media filters the gap between the applicant's attitude and character towards employment with that of the employer. A study conducted by Kroeze identified cost reduction as an advantage in using social networking sites (SNS) because proportion of the work load has been moved from the employer to the applicants since the latter need to make sure their social media platforms look well which makes the recruiting process less expensive and faster for the recruiter (Kroeze, 2015). One tricky issue that might come out is the possibility of fabrication which put doubts on the reliability of the information present in the user's profile. Fowler suggests that it would be improbable for the majority of the information to be fabricated since other people who know the user may become anxious with the disinformation, to the point that the erroneous profile being reported (Fowler, 2013) .

In an attempt to find more detailed information about job applicants, existing works reported an increase in the use of social media by hiring managers and human resource professionals (CareerBuilder, 2016; Time, 2014). Using the internet to gather information concerning job applicants or existing employees to assess person's suitability to the position, cyber-vetting is increasingly being done as extension to traditional background investigation (Rose, A., Timm, H., Pogson, C., Gonzalez, J., Appel, E., & Kolb, N. , 2010). *Cyber-vetting* can involve (1) searching the candidate in search engines; (2) targeted examination of public social media profiles such as in Facebook and LinkedIn; and (3) perhaps a data privacy concern of requesting that the employee voluntarily divulge login credentials to give full access to profile. (Payne, 2014). Automating this tedious task of conducting background investigation would prove helpful. While several studies has been done to use social media as source of information to understand user characteristics, quantifying textual posts to automatically assign a social character score to represent an applicant character has not yet been extensively explored. This will provide the recruiting staff the ease in going through bulk of unstructured data to infer applicants' fitness of job employment with the company.

This study aims is to develop a web application which aids in the process of recruitment and talent acquisition using social media analytics. It intends to derive character assessment based on identified factors that influence employee character. Indicators that have negative impact to social profile include statements expressing approval of alcohol, guns, profanity, illegal drug references, and sexual activities. Poor spelling and grammar likewise has a negative impact to social profile. Volunteerism and charity works bring positive impact to user

social profile (Jobvite, 2014; Time, 2014; Workopolis, 2015; Mills, 2015).

2.0 Methodology

Text mining in social media provides a mechanism to understand underlying characteristics of the person based on the textual data from unstructured free-form status posts (Aggarwal & Zhai, 2012). It encompasses the processing and representation of text for analysis to extract information useful for particular purpose. (EMC Corporation, 2012) The study employs social media analytics through text mining which involves a three-stage process: capture, understand, and present (Fan & Gordon, 2014; Ganis & Kohirkar, 2016). Text data is captured through social media APIs, understood through natural language processing and presented through computation of a personality profile score.

Data Retrieval

Social media data are collected at the time the user voluntarily registers through the system. A number of libraries and APIs are used in the back-end to perform data retrieval as shown in Figure 1. With the use of Twitter4J, the system can access Twitter API to retrieve tweets from the user's account. For Facebook, Facebook Javascript SDK is used to retrieve an access token. The access token is used by the system to request data from Facebook Graph API. Data retrieved from the APIs are returned to the system in JavasScript Object Notation (JSON), a lightweight data-interchange format which provides an easy way for the application to retrieve and parse (Ecma International, 2013) . To retrieve data from LinkedIn, JSOUP library is utilized to parse the HTML document of the user's public profile and get the list of skills. All data retrieved are saved in a database.

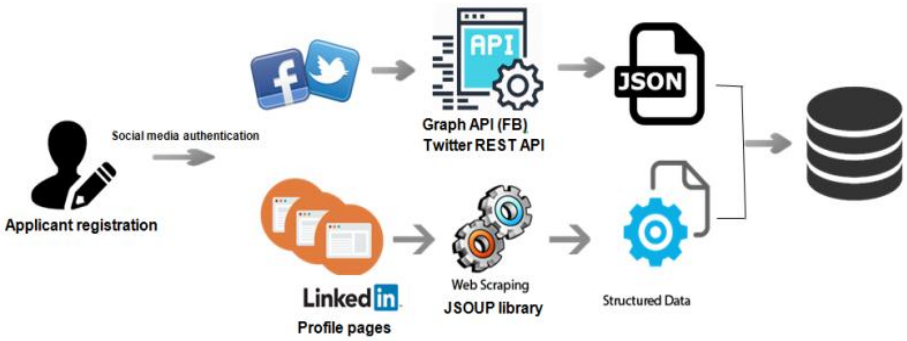

Figure 1. *Data retrieval*

Natural Language Processing

To obtain the format usable for processing, data preprocessing is performed. Pre-processing involves tokenization, removal of irrelevant characters such as emoticons, images, non-ASCII characters, stop words, and repeated characters due to word lengthening for greater emphasis and word stemming (Gurusamy & Kannan, 2014). Preprocessed data undergo key term search which checks for occurrence of pre-defined list of associated words to determine which

factors they are related with. Factors considered in evaluating each post include content related to alcohol, charity, volunteerism, firearms, illegal drugs, politics, profanity, and sexual activities. For sentences containing ambiguous words relating to a category, word sense disambiguation process is done to determine the actual sense of the word used in the sentence using WordNet.

All status posts which were identified to contain any of the categories in the text, lexicalized tree were generated using Stanford CoreNLP Lexicalized parser to show the grammatical relations of words. A type dependency list is then extracted which will be used to set the content polarity of the post, whether the post expresses approval to the content or not. Post expressing disapproval to categories which have strong negative responses from recruiters such as illegal drug references, sex, profanity, guns, and alcohol would mean positive points to the user. Spelling and grammar is also checked for scoring.

Scoring Algorithm

In order to quantify the user's character assessment, the researchers proposes a social scoring algorithm which assigns personal character score to the user based on social media activities. The study used seven categories that employers would look into, namely: Alcohol, Charity and Volunteerism, Firearms, Illegal Drugs, Poor Spelling and Grammar, Profanity, and Sexual, which would influence the character of an applicant as an employee. Each category score is computed using factor weights per polarity as shown in Table 2. The weight distribution is adapted from the survey conducted by Jobvite National Survey 2014 (Jobvite, 2014) .

Table 1. *Factors weight distribution per polarity*

Factor	Positive	Neutral	Negative
Illegal drug references	5%	22%	**83%**
Sexual posts	1%	17%	**70%**
Spelling/ grammar	3%	24%	**66%**
Profanity	5%	22%	**63%**
Guns	2%	32%	**51%**
Alcohol	2%	43%	**44%**
Volunteering /donations to charity	**65%**	25%	2%

All posts identified to contain the aforementioned factors are subjected to sentiment analysis to analyze the emotional content of the post. Posts are scored using the AFINN lexicon which rates polarity of statement from -5 to +5. Statements are further classified based on the polarity: positive (+1 to +5), neutral (0) and negative (-1 to -5). Category score is the average of each polarity score multiplied by the factor weight. The overall personality profile score is calculated by subtracting the sum of the category scores from 100 which is set as baseline score for each profile. Figure 2 shows conceptual diagram of the entire process.

A list of skills identified by the employer in the job posting is checked with each user's list of skills retrieved from LinkedIn. The users with the most number of matched skills are recommended with the personality profile score used for relevancy ranking.

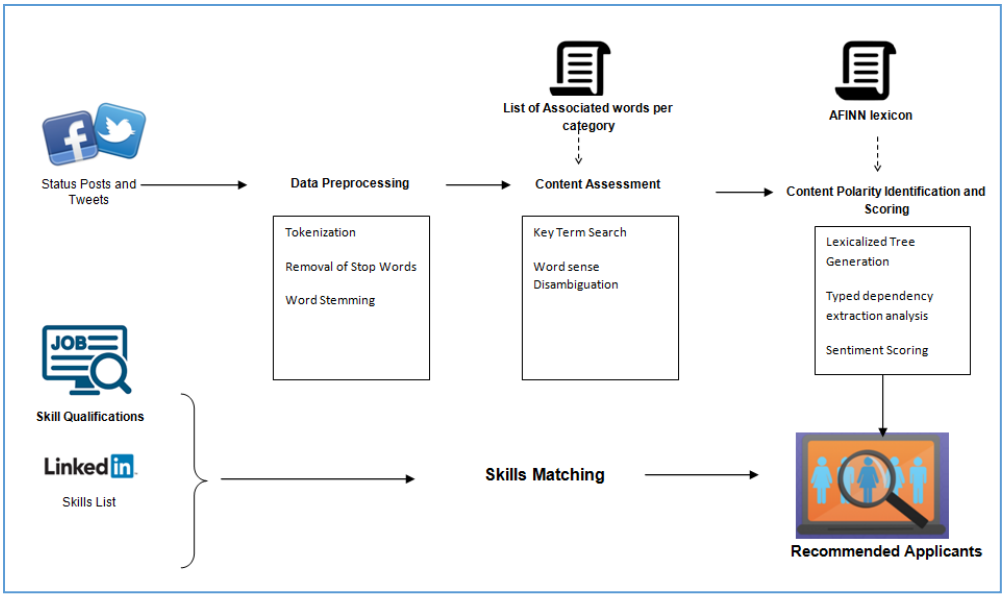

Figure 2. *Conceptual Diagram*

3.0 Results and Discussion

While issues abound with privacy and abuse of data, security measures are being adhered by social networking sites including but not limited to the restrictions of data access of different levels. Data retrieval through API for Facebook has been recently modified to only allow developers access to basic profile data. Moreover, access to posts, comments, and other data, are only granted to applications which have passed their Login Review Process. With these changes, the researchers need to register tester accounts of the application in order to get through the authentication.

One of the limitations of the application is it could only handle data in the English language. Words in other languages are not recognized when being processed which results in possible loss of data. Integration of language localization would be of much help so that status posts which are written in the local language may be included in the analysis. As of this writing, there has been no usable corpus yet developed for the Cebuano-Visayan language.

Further, the assessment score derived from this study is limited to assessing the occurrence of associated words on the status posts and tweets as defined by the researchers. Refinement of the proposed scoring mechanism through application of reject inference techniques as applied in credit scoring models may be implemented to address possible bias in the calculation of the assessment score and including applicant's manner and style of writing, the choice of words and mutual connections as basis for scoring.

User Interface

A web-based job placement portal is developed to provide an avenue for employers to create job posts and applicants to seek for jobs. The web application is developed using Java Platform, Enterprise Edition (JEE) with Tomcat V8.0 Server as application server and Spring Web model-view-controller (MVC) framework to

implement abstraction of the user interface logic, business logic and data-related logic. Hibernate Object-Relational Mapping (ORM) is used to store and retrieve objects into a relational database.

Figure 3 shows a screenshot of employer job postings. Employers post jobs specifying required candidate skill qualification to get applicant recommendation from the system.

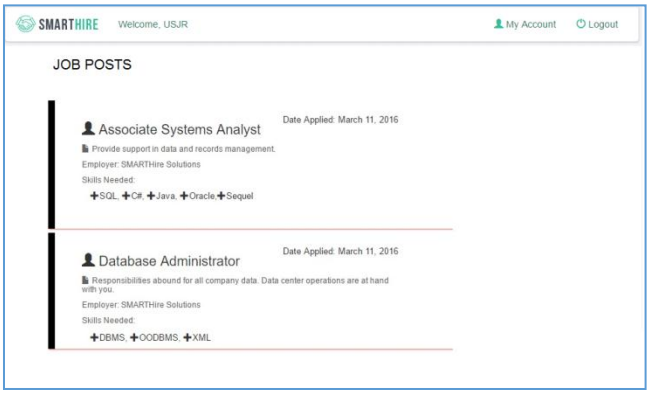

Figure 3. *Employer Job Posts*

The user voluntarily permits the application to retrieve social media data for assessment purposes. Applicants register to the site to be able to apply, receive notifications on job postings which match their skills and be included in the system's recommendation to the specific job posts. During the registration process, applicants specify social media accounts to allow the application retrieve relevant information. Figure 4 shows a screenshot of the user registration form.

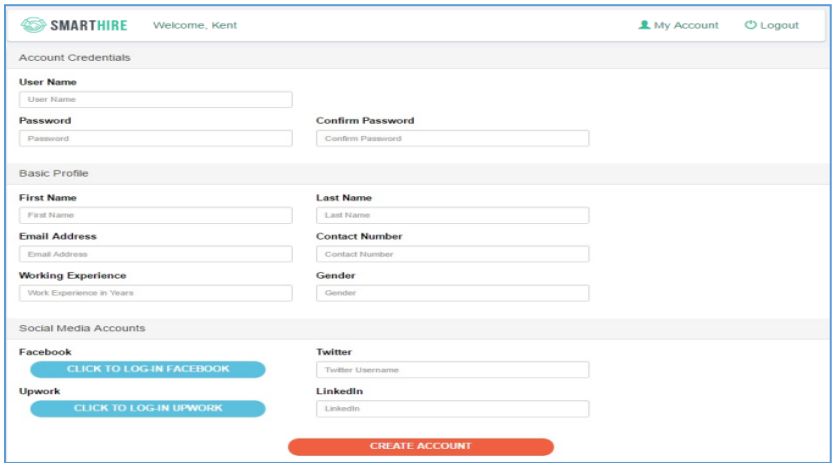

Figure 4. *Applicant Registration Form*

For each job post, the system recommends qualified applicants based on skills requirement with relevancy ranking based on computed social character score. Figure 5 shows list of recommended job applicants for a job post with their

corresponding character score. Skills list from LinkedIn is used to identify competent applicants based on job qualification.

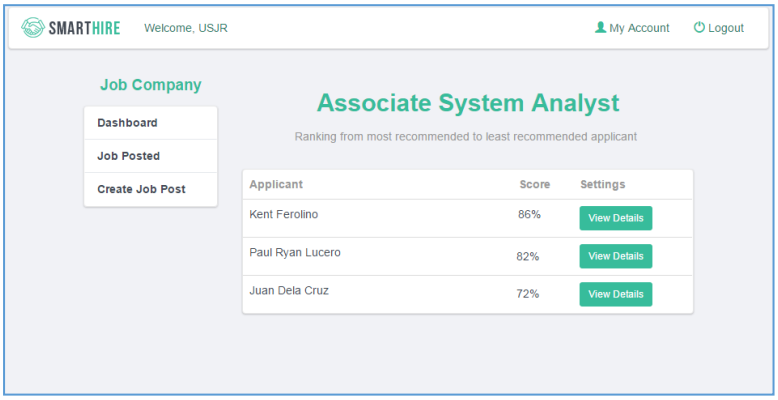

Figure 5. *Recommended Job Applicants*

Performance Testing

In order to determine the response time of the social media API on client requests, data retrieval is tested on actual user accounts with varying number of posts/tweets for a span of 3 years and 3 months. The time elapsed to fetch the data is recorded. The average ping on the servers during the entire testing process for Facebook and Twitter is at 316ms and 386ms respectively.

Table 2. Data Retrieval Performance Testing Summary

Social Media Platform	Data retrieved	Time elapsed
Facebook Data Retrieval from Jan 1, 2013 to March 1, 2016	Number of posts retrieved: 279	21.90 seconds
	Number of posts retrieved: 250	10.52 seconds
	Number of posts retrieved: 279	21.90 seconds
	Number of posts retrieved: 1136	45.12 seconds
	Average time per post	0.0498560120 sec/post
Twitter Data Retrieval from Jan 1, 2013 to March 1, 2016	14 tweets	0.840 seconds
	99 tweets	4.201 seconds
	Average time per tweet	0.051215 sec/tweet

The system was able to retrieve more than 3-year worth of data for less than a minute for both Facebook and Twitter. Fetching this much data is only done during the first login process. Updates were incremental therefore; response time would be shorter for subsequent posts.

Accuracy Testing

In order to validate the accuracy of the system, 70 text input, 10 from each category labeled through human interpretation were tested through system and human interpretation for accuracy in content assessment and polarity. The result of human interpretation is compared with the systems result. The initial test result shows 88% (62 out of 70) accuracy in content assessment while 82% (58 out of 70) in content polarity. Table 3 presents the accuracy testing summary.

Table 3. *Accuracy Testing Summary*

Category	Content Assessment		Content Polarity	
	Human	System	Human	System
Alcohol	10	10	10	9
Charity and Volunteerism	10	8	10	9
Firearms	10	8	10	6
Illegal Drugs	10	10	10	9
Profanity	10	9	10	9
Sexual	10	7	10	9
Poor Spelling and Grammar	10	10	10	7
Total	**70**	**62**	**70**	**58**

Based on the testing result, the system was able to correctly identify posts which contain any of the pre-defined categories and the polarity of the statements.

4.0 Conclusion

The study is an attempt to show the use of text mining through natural language processing to provide a mechanism to automatically assign a social character score through textual posts of emotion and sentiments in social media. The development of the application presented does not intend to replace actual personal encounter as a means to assess a person but rather provide a tool to support HR in sourcing the right candidates by leveraging through technology to streamline the lengthy process. As a screening method preceding interview and other recruitment process, this software would aid recruiters in being selective of a better pool of candidates to carry on, thus reducing the time and cost spent on the succeeding processes.

References

Aggarwal, C., & Zhai, C. (2012). Mining Text Data. *Science & Business Media*.

Boyd , D., & Ellison, N. (2008). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication , Vol. 13*, pp. 210-230.

CareerBuilder. (2016, April 28). *Number of Employers Using Social Media to Screen Candidates Has Increased 500 Percent over the Last Decade*. (CareerBuilder) Retrieved from careerbuilder.com

Ecma International. (2013). The JSON Data Interchange Format.

EMC Corporation. (2012). Data Science and Big Data Analytics.

Fan , W., & Gordon, M. D. (2014). The Power of Social Media Analytics. *Communications of the ACM 57 , Vol. 6*, pp. 74-81.

Fowler, K. M. (2013). Scanning Social Networking Sites as Part of a Hiring Process (Masteral Thesis). Graduate School of Angelo State University.

Ganis, M., & Kohirkar, A. (2016). *Social Media Analytics Techniques and Insights for Extracting Business Value Out of Social Media.* IBM Press Pearson.

Gurusamy, V., & Kannan, S. (2014). Preprocessing Techniques for Text Mining. *ResearchGate.*

Hoek, J., O'Kane, P., & McCracken, M. (2016). Publishing personal information online: How employers' access, observe and utilise social networking sites within selection procedures. *Personnel Review , Vol. 45* (Issue 1), pp. 67 – 83.

Jobvite. (2014). *2014 Social Recruiting Survey.* Retrieved July 10, 2016, from http://www.jobvite.com/

Kluemper , D., & Rosen, P. (2009). Future employment selection methods: evaluating social networking web sites. *Journal of Managerial Psychology , Vol. 24* (Issue: 6), pp.567 - 580.

Kroeze, R. (2015). *Recruitment via Social Media Sites: A critical Review and Research Agenda.*

Mills, H. (2015, June 14). *The influence of social media on job applications*. Retrieved June 14, 2015, from http://heleenmills.com

Payne, E. (2014). *Think before you post... Your future employer may be watching.* Doctoral Research at University of College Cork.

Rose, A., Timm, H., Pogson, C., Gonzalez, J., Appel, E., & Kolb, N. . (2010). *Developing a Cybervetting Strategy for Law Enforcement.* U.S. Department of Defense.

Time. (2014). *The 7 Social Media Mistakes Likely to Cost You a Job.* Retrieved June 14, 2015, from http://www.time.com/

Time. (2014, June). *The 7 Social Media Mistakes Likely to Cost You a Job*. Retrieved June 14, 2015, from http://www.time.com/

Workopolis. (2015). *The top three things that employers want to see in your social media profiles.* Retrieved June 14, 2015, from http://careers.workopolis.com/