Profiling Analysis Based on Social Media for Prospective Employees Recruitment Using SVM and Chi-Square

Mihuandayani$^{1,2}$, E Utami$^2$ and E T Luthfi$^3$

$^{1,2,3}$ Magister of Informatics Engineering, Universitas AMIKOM Yogyakarta, Indonesia.

E-mail: mihuandayani20@gmail.com

Abstract. Human resource management is one of the important key success of a company and a responsibility of Human Resource Department (HRD). To increase the quality of human resources, the company has made various policies in improving the capability of the employees. One of the processes is by doing the selective prospective employee recruitment. The recruitment process functioned to determine the ability, personality and psychological of the prospective employees. This research focuses on a technique to find the personality traits through the use of social media. The timeline activity of Facebook and Twitter are used for profiling analysis to find the candidate employee personalities based on the specified category namely Dominance, Influence, Steadiness, and Conscientious (DISC). Classification is performed by using Support Vector Machine (SVM) and Chi-Square to select a number of relevant attributes. The evaluation of the algorithm is used to measure the performance of SVM with Chi-Square using the parameter of accuracy, precision, and recall.

1. Introduction

Human resource is one of the company's valuable assets in any sector, whether small, medium and high. Human resource management in an organization is used to maximize the productivity by optimizing the effectiveness of its employees [1]. Based on the theory this condition used to support the organization in achieving its objectives by developing and implementing an integrated human resource strategy with the organization's business strategy. Human resources as valuable assets need to be managed properly in order to run optimally in the process of completion of work and face every activity effectively and efficiently. To carry out the work function in managing human resources of the company, personnel department and Human Resources Department make various efforts both in the process of receiving employees and in the process of completion of work by employees. In the early stages of the process of receiving the employee, the company usually performs recruitment of prospective employees openly through announcements both online and offline. In the process of distributing recruitment information through posters or social media functioned to find candidates - employees in
accordance with the position and qualifications required. This recruitment process usually consists of various tests, namely written test, practice test, personality test, and interview test. Written tests are used to determine the cognitive or level of knowledge of prospective employees. Practice tests are usually done for the needs of the practitioner's position that need certain skills in the work. Personality becomes one of the most important influencing factors in human resource management personality according to [2] is the psychological quality of a person that influences the behavior patterns of stable and different individual characteristics. Personality tests and interview tests are needed to know the personality of a person through the character possessed by the prospective employee to be considered with the required position. However, in the process of recognizing the personality of a person in a certain time, still valuable less because the personality of a person can not only be assessed for a short period of time. Therefore, it requires to be used as an alternative to knowing a person's character as supporting data from the interview test and personality tests performed.

This research proposes a technique that can be used to find a candidate personality category by utilizing social media. Social media is intended with the consideration of social activities posted on the timeline of social media can describe the character or personal nature of a person. Activities on social media like Facebook, Twitter, Instagram or YouTube become a container in people interacting and expressing feelings through status or comment column. Utilization of social media has been used as a source of research data to conduct a person's personality profiling analysis as in research [3, 4]. Profiling is a technique that automatically processes personal and non-personal data aimed at developing predictive knowledge of data in the form of building profiles that can then be applied as a basis for decision making [5]. Data in the timeline of social media users consists of data in the form of text, images, audio, and video. Many data and unstructured in social media can be processed using various techniques such as text processing, image processing, and the bag of word approach and so on. In this study, the processed data focus on text data using classification techniques to classify data based on the DISC model that divides the behaviour into four different dimensions: Dominance, Influence, Stability and Conscientious (DISC) that recognizes the feelings and behaviours of a normal person in interaction with the environment to evaluate the behaviour by the measuring stick of convention [6].

In the text classification algorithms are often used such as Naive Bayes, Support Vector Machine, and Artificial Neural Network and so on. The algorithm has advantages and disadvantages of adjusting the type of data used. Support Vector Machine (SVM) according to research [7, 8] has better performance or performance levels. But there are deficiencies in dealing with many attributes especially in the management of big data. To be able to optimize an algorithm and handle the problem of many attributes can be done through the selection of relevant and appropriate features. Selection of frequently used features such as Information Gain, Chi-Square, Backward Selection, and Particle Swarm Optimization and so on. From the existing feature selection techniques, there are advantages in the Chi-Square algorithm that has a
good level of accuracy [9, 10] in support of algorithm performance functions or SVM classification techniques performed on a number of data.

Therefore, this study uses SVM and Chi-Square to process data derived from social media Facebook and Twitter are often used to update the status and comments by users. It is used to find candidate employee personality categories based on the DISC model in support of company recruitment process. Then to be able to measure the performance of the algorithm used in data processing, the algorithm is measured using precision, recall, and F-measure parameters. The results obtained from data processing with this classification technique is used to support his worth of prospective employees in occupying the required company position or as a consideration of the company in the process of recruitment in finding the appropriate candidate personality level.

2. Related Work

Research on the use of social media as a data analysis continues to be done, including in psychological and sociological studies of profiling analysis to find the personality traits of a person. The interaction analysis of Facebook users is used to gather information about the personality and predict based on experimental 3-class models and 5-class models using machine learning especially classification tree [11]. The study shows that Facebook profile data developed with TP2010 can be utilized to predict user personality with a degree of classifier higher than 70% achieved for all personality traits. Other studies investigating personality predictability through smart-phone data based on Big Five Personality Traits are Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism using SVM with a radial basis kernel function for the classification model [12]. The study estimates the personality traits of a set of behavioral variables extracted and finds the only extroversion adapt predicted to be significantly better at 35.6% than by null model. Data derived from Twitter can be used to analyze the user's experience with personality judgment especially a person's psychological profile. Through the assessment of the Dominic, Influence, Compliance, and Steadiness assessment, a study proposes a way to predict a user's personality that is useful in career development by mapping information via Twitter data [3]. The research process data through text mining approach in analyzing DISC.

Research that utilizes Facebook as a source of data to predict the personality of a person using the machine learning algorithm and deep learning. The study [13] yielded an average accuracy rate of 74.17% using the big five model personality model. Comparison of classification techniques in [14] research using a feature-level approach to select the appropriate features of some data. The study compared seven classification techniques to measure algorithmic performance in detecting malware from Android applications that resulted in SVM after selecting features through correlation analysis with the best accuracy rate of about 97.16%. In addition to these studies, the identification of relevant factors that affect the classification performance on customer loyalty with the feature selection technique approach compared with chi-square with threshold> 0.01 shows the best result [9]. Feature selection in text mining stages
is often used to remove irrelevant features. The method of selecting such features may affect the performance of a given classification algorithm. The [15] study comparing and evaluating five classification techniques yielded SVM as the best algorithm by showing the smallest bias against the given feature selection method. In addition, the combination of the selection feature classification in the data results in a suitable combination such as chi-square and information gain identified approximately equivalent groups. The study using intrusion detection model using chi-square feature and SVM machine learning with Radial Basis Function kernel optimization resulted in better detection rate than the proposed method [10].

3. Research Methodology

In conducting a profiling for people in social media and analyzing based on the personality that is related to the recruitment process of prospective employees, this research consists of various stages in performing profiling using the SVM and Chi-Square algorithms to find appropriate target variables based on training data. The initial stage begins with knowing the process of recruitment flow used as the foundation of the company in recruiting new prospective employees. Recruitment is done starting from the request form of each division or department within the company. In the request form is explained about the qualification requirements of prospective employees who will occupy positions within the department. In addition, the time of work readiness required to occupy the position sought. Then, through the request form required approval from the Human Resource Development (HRD) and the President Director of the company. Once approved, the HRD will make an announcement through online media and print and work with certain parties in spreading recruitment information prospective employees. In this process, it takes about two weeks to collect the prospective employee file. Then the next stage is the selection process. In the selection process requires several more stages to determine the right candidates and in accordance with the required position. The selection process consists of an administrative selection, written test, practice test, and interview test. The flowchart system of this research can be shown in Figure I.

![Flowchart System of Profiling based on Social Media](image-url)

**Figure 1.** Flowchart System of Profiling based on Social Media.
Based on Figure I, the process of profiling is started by preparing the data-set from the personal page in the social media platform. In this crawling process, need to identify the kind of data such as the status updates and the comments in the user’s page. The crawling process used the Application Programming Interface (API) of Facebook and Twitter to get the authentication key and store the data in Excel format. After getting the milestone data of the personal page, the cleaning process becomes the important part to make sure the data are given the functions to replace the missing value, set role, nominal to text, replace the link, replace hash-tag symbol and then to remove the duplicates data.

The preprocessing stage is then processed the unstructured data in order to be properly classified. Text preprocessing is a part of the natural language processing system which identifies the words in the fundamental units from analysis and tagging components [16]. This preprocessing stage consists of case folding, convert emoticon and negation, tokenizing, filtering and stemming. Preprocessing text which is used in some parts such as the case folding which transform the letters “a” to “z” in the lowercase, convert the emoticon based on the list symbol of emoticon and convert negation to change the polarity of the word. Tokenizing is used to break the status update or comments into a snippet by separating the word into every space. Then, the filtering is used to remove the irrelevant words by giving the list of stop-word removal. Stemming is to reduce every word to get the word base by using the library which contains the list of Indonesian lexical base.

Define the class label into the DISC become a very important part of this profiling because the class label should be identified through the corpus-based theory. To know every update status of the personal page into the DISC style need a proper dictionary which can identify the word. People with dominance types are highly competitive in contributing their efforts and responsibilities. Tends to be serious on the set targets because they love the challenge. People with typical dominance in a career tend to have a liberal view by frequently changing jobs because of impatience or inability to adapt. Influence types usually tend to concern around and are willing to help others because social activities are also likely to be persuasive and optimistic. The next type is Steadiness which is usually known in the patience and tends to be more calm and relaxed. Persons with this type always try to maintain the status quo by resisting change. Conscientious types tend to have a very diligent perfectionist nature in doing the work and adapt to situations to avoid hostility. The detail information about the process of data set through the update status and comments of the user in social media particularly in DISC based on theory [17] can be explained in Table 1.
Table 1. Behaviour Descriptors Based Theory.

| DISC Style | Dominance       | Influence     | Steadiness | Conscientious |
|------------|-----------------|---------------|------------|---------------|
| Decisive   | Charming        | Understanding | Accurate   |
| Competitive| Confident        | Friendly      | Precise    |
| Daring     | Convincing      | Good Listener | Analytical |
| Direct     | Enthusiastic    | Patient       | Compliant  |
| Innovative | Inspiring       | Relaxed       | Courteous  |
| Persistent | Optimistic      | Sincere       | Diplomatic |
| Adventurous| Persuasive      | Stable        | Detailed   |
| Problem Solver | Trusting | Steady    | Fact Finder |
| Results Oriented | Trusting | Team Player       | Objective |

After labeling the class data, Chi-Square is used to weight the data set and sort the weight. Chi-Square is a numerical test that measures deviations from expected distributions considering features that are independent of grade values. This measure compares the number of co-occurrences of two words with the expected number if the words were independent and also normalizes the comparison by the expected number [18]. Feature selection and ranking are crucial for scoring for every potential feature. Weighting by chi-square is done by calculating the frequency of a feature in training class samples separately to get its function. Chi-Square also selects a number of relevant attributes. The following Chi-Square’s formula is shown as follows.

$$\chi^2 = \frac{(n_{ab} - \frac{1}{N} n_a n_b)^2}{n_a n_b}$$

The classification technique classified the training data in four target variables based on the common character types of human being and the theory based DISC. SVM algorithm that aims to find the function of the separator (hyperplane) with the largest margin, so as to separate the sets of data optimally [19]. SVM is capable of working on high-dimensional datasets using kernel tricks. Radial Basis Function (RBF) kernel is the kernel used in this research. SVM used only some of the selected data points that contribute (Support Vector) to form the model used in the classification process. SVM used the kernel and identified the predictive class. The margin in vector mathematics used in SVM is shown in the following formula.

$$Margin(w) = \frac{|w.x^-| + |w.x^+|}{||w||}$$
The vector \( x^- \) and \( x^+ \) as the support vectors [18] and \( w \) is known as the linear combination vectors. To find the value of \( y \) used the following formula.

\[
x_i, w + b \geq +1 \text{ for } y_i = +1 \text{ and } x_i, w + b \leq -1 \text{ for } y_i = -1 \quad (3)
\]

After knowing the value of every support vector, to get the optimization in SVM can be used the following formula.

\[
\min_{w,b} L_p(w, b, \alpha) \equiv \frac{1}{2} |w|^2 - \sum_{i=1}^{n} \alpha_i y_i (x_i, w + b) + \sum_{i=1}^{n} \alpha_i \quad (4)
\]

Based on [20, 21], the SVM algorithm is very effective used to handle the text classification problem. In this study, the classification used in four class based on DISC theory and focuses on how the company strives to be able to know the psychological attitude of new prospective employees. In addition to interviews that can see a cursory style of speech and innate from employees. This research proposes the existence of profiling using Facebook and Twitter account of prospective employees through status updates and comments that are periodically done. After getting the classification with SVM, the measurements process is done by the parameters of accuracy, precision, and recall.

4. Result and Discussion

This section presented the result of profiling analysis into DISC style by giving the sample data and show the result of the measurements algorithm which is used in the process of machine learning from the social media personal page.

4.1. Dataset Collection

The collected data of social media is the update status and comments in the personal page which contains about 1000 dataset in the Indonesian Language used of Facebook and Twitter platform. Profile used as sample data that is as many as five profiles from each platform. From each profile or assumed as a candidate for a new employee. To be able to place the required position such as marketing or project officer, the theory of DISC can represent the character of the prospective employee background in terms of psychology. The result of the profiling analysis performed on five profile samples shown in the following Table 2.

| Sample  | Dominance | Influence | Steadiness | Conscientious |
|---------|-----------|-----------|------------|---------------|
| Profile A | 0.54     | 0.27      | 0.07       | 0.12          |
| Profile B | 0.16     | 0.04      | 0.55       | 0.25          |
| Profile C | 0.11     | 0.32      | 0.49       | 0.08          |
| Profile D | 0.44     | 0.28      | 0.21       | 0.07          |
| Profile E | 0.19     | 0.59      | 0.09       | 0.13          |
Based on Table 2, the five samples shown in every profile has the range value of every DISC from Facebook data and the profile presented the higher percentage whether in Dominance, Influence, Steadiness and Conscientious. For example, the Profile A shown that tend to have the Dominance characters same with the Profile D with 54 % value and 44 % respectively. Based on the theory of DISC that people with this types are highly competitive and serious on the target but sometimes that people have impatience with frequently changing a job. It can be the consideration of the company, especially for the HRD to determine the proper prospective employee related to another recruitment test. The graphic of the DISC from Facebook data is summarized in Figure 2.

![DISC of Facebook Data](image)

**Figure 2.** The Result of Profiling Based on DISC from Facebook Data.

In addition to data sourced from Facebook, profiling derived from Twitter data is also used in this study because of the number of Twitter users who do post and tweet activity quite high. The sample used is the same profile account of five samples but derived from Twitter. Here are the results obtained through profiling based on Twitter data shown in Table 3.

| Sample | Dominance | Influence | Steadiness | Conscientious |
|--------|-----------|-----------|------------|---------------|
| Profile A | 0.43 | 0.27 | 0.08 | 0.22 |
| Profile B | 0.14 | 0.16 | 0.32 | 0.38 |
| Profile C | 0.2 | 0.26 | 0.49 | 0.05 |
| Profile D | 0.37 | 0.19 | 0.33 | 0.11 |
| Profile E | 0.25 | 0.49 | 0.1 | 0.16 |

Based on Table 3, the results of profiling twitter data show strong and similar range values on Profile A, Profile C, Profile D, and Profile E if we compared with the Facebook result. While for the Profile B there is a significant value which is found that the character of Conscientious has about 38 % in Twitter data as the higher character. This condition is opposite with the Facebook data that the Profile B get the Steadiness for higher value. This case can be solved by
getting the average result to determine the type for Profile B. Here is a graphic illustration of profiling analysis obtained through Twitter data shown in Figure 3.

DISC of Twitter Data

![DISC of Twitter Data](image)

**Figure 3.** The Result of Profiling Based on DISC from Twitter Data.

To be able to know the performance of algorithms that have been applied to profiling based on DISC behavior from social media, conducted tests that show the accuracy and accuracy of the implemented algorithm.

4.2. Result of Classifier

To perform calculations on the algorithm used is SVM with parameters \(c\) and \(\gamma\). The value parameter \((c, \gamma)\) is determined first by the range of values 0.1 to 0.9, then pairing each parameter value \((c, \gamma)\) so that the parameter pair that produces the highest accuracy is used in the test scenario based on the percentage of training data and test data. The values of the hyper-parameters \(C\) and \(\gamma\) which given the best is shown in Table 4.

|   | Dominance | Influence | steadiness | Conscientious |
|---|-----------|-----------|------------|---------------|
| \(c\) | 0.75      | 0.61      | 1.50       | 2.41          |
| \(\gamma\) | 0.18      | 1.95      | 0.75       | 0.65          |

Table 4. Value of Hyper Parameters \(c\) and \(\gamma\)

Measurement is done by using the confusion matrix technique which contains the True Positives (TP), True Negatives (TN), False, False Positive (FP) and False Negative (FN). From this TP, TN, FP and FN can be calculated the accuracy, precision, and recall. Accuracy means the percentage of the algorithm correctly does the prediction by using the following formula.

\[
Accuracy = \frac{TP+TN}{n}
\] (5)
While the precision means the percentage of the predicted data as positive is correct by using
the following formula.

\[
Precision = \frac{TP}{TP+FP}
\]  

(6)

Recall means the percentage of positive data predicted as positive or how much the selected
relevant item by using the following formula.

\[
Recall = \frac{TP}{TP+FN}
\]  

(7)

The result of the performance measurements by using the parameters of accuracy, precision, and
recall which is implemented in the Facebook data is shown in Table 5.

| DISC Style  | Accuracy | Precision | Recall |
|-------------|----------|-----------|--------|
| Dominance   | 0.78     | 0.88      | 0.54   |
| Influence   | 0.61     | 0.72      | 0.92   |
| Steadiness  | 0.47     | 0.76      | 0.85   |
| Conscientious| 0.72     | 0.89      | 0.72   |
| Average     | 0.65     | 0.81      | 0.76   |

Based on Table 5, the profiling analysis with the SVM algorithm and chi-square in Facebook
data shown the average of accuracy, precision and recall was 65 %, 81 %, and 76 % respectively
achieved. While the highest value of accuracy, precision, and recall was in 78 % Dominance, 89
% Conscientious and 92 % influence. The measurements graphic of the Facebook data is illustrated in Figure 4.

![Figure 4. Facebook Measurements with Confusion Matrix](image)
In addition to the measurements of Facebook data, the testing is also done by the Twitter data to test the performance algorithm. The result of measurements in Twitter data can be shown in Table 6.

**Table 6. Performance Measurements of Twitter Data**

| DISC Style  | Accuracy | Precision | Recall |
|-------------|----------|-----------|--------|
| Dominance   | 0.43     | 0.82      | 0.87   |
| Influence   | 0.65     | 0.47      | 0.82   |
| Steadiness  | 0.65     | 0.66      | 0.91   |
| Conscientious | 0.83  | 0.79      | 0.54   |
| Average     | 0.64     | 0.69      | 0.79   |

Based on Table 6, the algorithm shown the average of accuracy, precision and recall were 64 %, 69 %, and 79 % respectively achieved for the profiling. The highest value of accuracy, precision, and recall were in 83 % Conscientious, 82 % Dominance and 91 % Steadiness. The graphic is illustrated in Figure 5.

![Performance Measurements of Twitter Data](image)

**Figure 5. Facebook Measurements with Confusion Matrix**

The measurement by confusion matrix is done for the profiling analysis based on DISC behavior in social media Facebook and Twitter data. The algorithm of SVM and Chi-Square performance can be calculated from Facebook and Twitter and the overall average has shown about 64.5 % accuracy, 75 % precision and 77.5 % recall achieved.

5. Conclusion

Profiling analysis conducted on social media data is done to assist the recruitment process of new employee candidates that can be used as consideration in terms of personality of prospective employees. In doing profiling is based on the DISC theory that divides the personality of a person based on Dominance, Influence, Steadiness and Conscientious. From the
sample used from the social media account five accounts from Facebook and Twitter are done from the stages of crawling data, cleaning data, preprocessing text, labeling, weighting by chi-square, classifier by SVM and perform measurements through accuracy, precision and recall parameters. The test results show the overall average of about 64.5% accuracy, 75% precision and 77.5% recall achieved. This research can be used as data processing support for the company in the recruitment process in addition to finding the ability in the field or company position required.

Acknowledgments
Universitas AMIKOM Yogyakarta for the support by giving all the facility which is needed for the study. This research also can be done by a good cooperation and big effort to provide a paper which hopes can be useful for others.

6. References

[1] Armstrong M and Taylor S 2014 Armstrong’s handbook of human resource management practice, 13th edition (London: Kogan Page)
[2] Huczynski A A and Buchanan D A 2007 Organizational Behaviour FT Prentice Hall, Harlow Essex sixth edition; Instructor’s Manual available online p 138.
[3] Ahmad N and Siddique J Personality Assessment using Twitter Tweets International Conference on Knowledge-Based and Intelligent Information and Engineering Systems KES2017, 6-8 September 2017 (Marseille, France)
[4] Farnadi G, Tang J, Cock M D and Moens M F User Profiling through Deep Multimodal Fusion WSDM’18 February 5-9 2018 (Marina Del Rey, CA, USA) DOI: 10.1145/3159652.3159691
[5] Ferraris V, Bosco F, Cafiero G, D’Angelo E and Suloyeva Y 2013 Defining Profiling Available at SSRN: https://ssrn.com/abstract=2366564 or http://dx.doi.org/10.2139/ssrn.2366564
[6] Marston W M 1928 Emotions of Normal People (London: K. Paul, Trench, Trubner & Co. Lt)
[7] Sheshasaayee A and Thailambal G 2017 Comparison of Classification Algorithms in Text Mining International Journal of Pure and Applied Mathematics 2017 Vol 116 No 22 p 425-433.
[8] Latah M and Toker L 2018 A Novel Intelligent Approach for Detecting DoS Flooding Attacks in Software-Defined Networks International Journal of Advances in Intelligent Informatics. Vol 4 No 1 p 11-20.
[9] H Sulistiani and A Tjahyanto Comparative Analysis of Feature Selection Method to Predict Customer Loyalty Journal of Engineering Vol 3 No 1 e-ISSN:2337-8557
[10] S T Ikram and Cherukuri A K 2017 Intrusion Detection Model Using Fusion of Chi-Square Feature Selection and Multi-Class SVM Journal of King Saud University – Computer and Information Sciences Vol 29 p 462-472.
[11] Ortigosa A, Carro R M, Quiroga J I 2014 Predicting User Personality by Mining Social Interactions in Facebook Journal of Computer and System Sciences 80 p 57-71.
[12] Monsted B, Mollgaard A, and Mathiesen J 2018 Phone-based Metric as a Predictor for Basic Personality Traits Journal of Research in Personality Vol 74 p 16-22.
[13] Tandera T, H D Suhartono D, Wongso R. and Prasetio Y L 2017 Personality Prediction System from Facebook Users 2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13 – 14 October 2017 (Bali, Indonesia)

[14] Sing L and Hofmann M 2017 Dynamic Behavior Analysis of Android Applications for Malware Detection. 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT) (Manipal University Jaipur) December 22-23

[15] Vora S and Yang H 2017 A Comprehensive Study of Eleven Feature Selection Algorithms and Their Impact on Text Classification Computing Conference 2017 18 – 20 July 2017, (London, UK)

[16] N Indurkhya and F J Damerau 2010 Handbook of Natural Language Processing. Second Edition (Chapman & Hall/CRC 2010 by Taylor and Francis Group, LLC)

[17] DiSC Classic Personal Profile System (The DiSC Model: Theory and Background, The Story Behind DiSC – it is older than you think, Elaboration of D, i, S, and C-behaviour) 2001 Inscape publishing.

[18] W B Croft, D Metzler, T Strohman 2010 Search Engines Information Retrieval in Practice (New York: Addison Wesley).

[19] H Jiawei, M Kamber, and J Pei 2012 Data Mining: Concepts and Techniques Third Edition. Waltham (MA: Morgan Kaufmann)

[20] Mihuandayani, E Utami, E T Luthfi 2018 Text Mining Based on Tax Comments as Big Data Analysis Using SVM and Feature Selection 2018 International Conference on Information and Communications Technology (ICOIACT) DOI: 10.1109/ICOIACT.2018.8350743.

[21] Ronen F and James S 2006 The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data (Cambridge University Press)