Statistical Semantics in Context Space: Amrita_CEN@Author Profiling

Barathi Ganesh HB[1], Anand Kumar M[2], and Soman KP[2]

[1]Artificial Intelligence Practice, Tata Consultancy Services, Kochi - 682 042, Kerala, India
[2]Center for Computational Engineering and Networking, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University, India

barathiganesh.hb@tcs.com,m_anandkumar@cb.amrita.edu,kp_soman@amrita.edu

Abstract. Languages shared by people differs due to diversity in their ethnicity, socioeconomic status, gender, language, religion, sexual orientation, geographical area, accents, pronunciation and word usages. This eventually fall into hypothesis that they follow unknown hidden pattern. By using this hypothesis, determining the class of a person such as age, gender, their personality and nativity has multiple applications in social media, forensic science, marketing analysis, e-commerce and e-security. This work advances the research on author profiling much further by overcoming existing language dependent, domain dependent and lexicon based author profiling methods by finding user's sociolect aspects based on authors statistical pattern of semantics in context space. It proves to be a domain and language independent method in Author Profiling by nearing constant performance over English, Dutch and Spanish corpus.

Keywords: Author Profiling, Context Space, Distributional Representation

1 Introduction

The amount of language sharing through Internet is prevalent due to the rapid growth of the social media resources like Facebook, Twitter, LinkedIn, Pinterest and chat resources like Hike, Whatapp, Wechat [1]. This positive growth ensures and encourages the recommendation and Internet marketing among users on a particular resource. It has been used in business organization for the marketing, market analysis, advertising and connecting with customers [2]. This ensures the need of Author Profiling (AP) [15] in-order to discover user's sociolect aspects from shared language. The complication involved here is unlike natural language the language shared by the people on social media is small and unfortunate to extract information easily out of it.

People started revolving around authorship tasks, right from the ancient Greek play- wright times. Recognizing the age, gender, native language, personality and many more facets that frame the profile of a particular person. It finds application in different zones like forensic security, literary research, marketing analysis, industries, on-line messengers, e-commerce , chats in mobile applications, medical applications to treat neuroticism and many more [2]. Forensic Linguistics came into existence only after 1968. In this sector, police register is one of the area under security, in which the statements taken down by the police act as a source text for Author Profiling (AP). Legal investigation continues its examination on all fields of suspicion.

In marketing, on-line customer reviews in blogs and sites helps the consumers in deciding his/her choice about shopping a product. Detecting the age and gender of the person who posted his/her feedback paves way for the owners to improve their business strategy [2]. Industries get benefited with customer's suggestions and reviews from which they could group the most likely products based on the gender and age. Twitter and Facebook are the popularly used sites for social media. In last year survey, it shows that every month there are about 236 million users who sign up to the micro blogging site-Twitter and 1.44 billion users to Facebook but among them 83.09 millions are fake accounts [1]. The author with age under 13 and author having more than one account noted as fake account which has to be taken care. There may also be anon who tend to have many fake id's and post messages and chat with innocent people in order to trap them.

In general Machine Learning (ML) algorithm can be used to attain this objective if subjected to relevant features and most of the existing methods follow the same [3][4]. In existing methods common and most commonly used features for AP are author's style based features (punctuation marks, usage of capitals, POS tags, sentence length, repeated usage of words, quotations), content based features (topic related words, words present in dictionaries), content and typographical ease, words that express sentiments and emotions with emoticons, special words from which information could be extracted, collocations and n-grams. These features are dependent on lexicon which varies with respect to the topic, genre and language. In ML, the low dimensional condensed vectors which exhibit a relation between the terms, documents and the profile was built using Concise Semantic Analysis (CSA) in order to create the second order attributes (SOA) which was classified using a linear model and also became sensitive to high dimensional problems . This system was further extended in 2014 to make it more precise in profiling. With the generation of highly informative attributes (creating sub profiles) using Expectation Maximization Clustering (EMC) algorithm, the system built was able to group sub classes within a cluster and exhibit a relation between sub profiles. Though this system was successful,it was dependent to the language and genre [3][4].

The syntactic and lexical features utilized in earlier models vary with respect to the morphological and agglutinative nature of the language. These features also varies with respect to the domain in which the AP is performed. There exists a conflict between classifying algorithms to learn from these features in order to build a unified and effective classification model which is independent of domain and language. This can be observed from system's performance in PAN - AP shared task[3][4][5].

In order to overcome these conflicts, this paper proposes a model based on statistical semantics from author's digital text. Statistical semantics paves way to the advancement in research of relational similarity by including statistical features of word distribution along with traditional semantic features utilized in Latent Semantic Analysis (LSA) [6]. It is clear that sexual aspects and vocabulary knowledge of a person varies due to human's cognitive phenomena which induces and also limits people of a particular gender and age group to utilize certain range of words to convey their message. Thus by utilizing this word distribution in context space and their statistical features, the gender and age group of a particular author is identified in this work. The basic idea is to utilize the distributional representation of an author's document to aggregate the statistical semantic information and promote a set of constraints for finding related hypotheses of that author's document.

2 Related Works

John collected large number of tweets and also evaluated it with people work using Amazon Mechanical Turk (AMT). Their data included 213 million tweets on the whole from 18.5 million users. Tweets collected were multilingual. As tweets include many more contents like emoticons, images etc., feature extraction part was limited to a particular n-gram length with total distinct features of 15,572,522. Word level and character level n-grams were chosen. There was no language specific processing done but instead only n-gram counts were taken into account. Once features were extracted classifiers namely SVM, Naive Bayes and Winnow2 were evaluated out of which Winnow2 performed exceptionally well with an overall accuracy of 92%. Their work was done only for gender classifying gender information [6].

Lizi told that the entrance to colossal measure of client produced information empowers them to examine lifetime semantic variety of individuals. The center reason of the model is that age impacts the point piece of a client, and every subject has an interesting age conveyance. They made use of Gibbs EM algorithm for evaluating their model. They were able to find information of both word distribution and age distribution from the sample of twitter data they collected. They treated tweets as bag of words content thus performing well and effectively mapping the topic to ages [7].

Pastor framed their methodology by utilizing the thought of second request properties (a low dimensional and thick record representation), yet goes past consolidating data among every objective profile. The proposed representation extended the examination fusing data among writings in the same profile, this is, they concentrated in subprofiles. For this, they naturally discover subprofiles and assemble report vectors that speak to more itemized connections of archives and subprofile records. Results shows proof of the helpfulness of intra-profile data to focus sex and age profiles. The sub profile or intra profile information of each author was found using Expectation Maximization Clustering (EMC) algorithm [8].

Suraj uses MapReduce programming standard for most parts of their preparation process, which makes their framework quick. Their framework uses word n-grams including stopwords, accentuations and emoticons as components and TF-IDF (term recurrence reverse report recurrence) as the measuring plan. These were bolstered to the logistic relapse classifier that predicts the age and sexual orientation of the creators. Mapreduce distributed their tasks among many machines and made their work more easy and fast [9].

Unlike PAN 2016, in PAN 2013, PAN 2014 and PAN 2015 the training and testing were done on similar domains. In most of the work author's stylistic features, readability, specific domain features (Emoticons, Hash tags), lexical features, LSA based features along with the projection based classifiers, regression based classifiers and clustering based classifiers are used to achieve the objective. In most of the proposed systems varying in its accuracy for different domain and language [3][4][5]

3 Mathematical Background

This section first presents the problem definition followed by the mathematical modeling to the idea described in section 1 for building AP model.

3.1 Problem Definition

In general the solution is to build a training model from the given problem set $p_t = d_1, d_2, ..., d_m$ and to map each document's author to a specific gender and age group $p_t \ \epsilon \ (gender, \ age\ group)$.

3.2 Training Phase

Step 1 - Constructing document - term matrix $[V_{i,j}]_{m \times n}$, where m is total number of documents (total number of authors) in p_t, n is size of the vocabulary [10] and

$$[V_{i,j}] = term\,frequency\,(v_{i,j})\,(1 < i > m)\ and\ (1 < j > n) \tag{1}$$

$$[V] = VSM(p_t)(1 < t > m) \tag{2}$$

Step 2 - Underlying semantic information and relation between authors's documents can be obtained using latent vector by finding basis vectors of $\left[VV^T\right]$ which is column space of V. This column called as the context space with respect to the author's documents. Thus the computed basis vectors spans the context space by satisfying the following condition [11],

$$\tilde{V} \approx [W] \times [H^T] \tag{3}$$

$$min f_r(W, H) \equiv \left\|V - WH^T\right\|_F^2 \tag{4}$$

$$s.t. \;\; W, H \geq 0$$

In equation 3, W is $m \times r$ basis matrix and H is $n \times r$ coefficient matrix. Linear combination of basis vectors (column vectors) of W with coefficients of H gives the context matrix V. While factorizing, formerly random values are assigned to W and H then the optimization function in equation 4 is applied on it to compute appropriate W and H. Where, r is the reduced dimension and F is the Frobenius norm. Here r fixed as m to have $m \times m$ context matrix. The basis vectors in W considered as the basis vector of context space, which are linearly combined with elements in the H to recompute the V. The singular or eigen vector based computation methods avoided here, since they constrained and forced to found the orthogonal basis vectors. This may not form the exact context space of the author's documents. Since the occurrence of the words in a documents cannot be a negative value, which is affordable by NMF and the non-negativity constraints makes interpretability straight forward than the other factorization methods [12].

Each element in the matrix W is a distributed representation of the semantic information of the author's documents in context space. This is known as Vector Space Model of Semantics (VSMs) [10] but in this application it captures user's cognitive ability and will be called as statistical semantics. Using these base vector it is possible to span the space, where the different representation of similar semantics lies.

$$[W]_{n \times n} = [x_1, x_2, ..., x_m] \tag{5}$$

$$[W] = VSMs([VV^T]) \tag{6}$$

Step 3 - The statistical features of semantic distribution in a context space are computed in order to build supervised classification model. Statistical features include the marginal decision boundaries with respect to word distribution in each document vector W_i based on each class which has to be classified. Performing NMF moves the values in W from discrete to continuous. Thus by taking W_i as a random variable 1 and by fixing random variable 2 as other distributions (Normal, Gamma, Chisquare, Rayleigh and Pareto Distributions), the correlation, null hypothesis between them are measured. This is expressed as,

$$[F]_{n \times s} = statistical\ features\left([W]_{n \times n}\right) \tag{7}$$

Where, s is number of statistical features and F is feature matrix for building classification model. From the above it is clear that the extracted features are only dependent on how the author's semantic distribution lies in a document.

Step 4 - In order to build classification model, the regression relation between the feature and the respective class are constructed using Random Forest tree algorithm which is a collection of Decision trees that formulates the classification rule based on randomly selected features in training set . From

$L = \{(y_i, F_i), 1 < i > n\}$ the subset of L_b is formed using $\sqrt{s}$ and b number of aggregate predictor is built [13]. Then final predictor is built by,

$$\varphi_b(F) = max_J H_J \tag{8}$$

Where, J is number of decision trees and $H_J = \{\varphi(F, L_b) = J\}$.

The gender and age group classification model is built using hierarchical method. In order to constrain the model, after finding gender information it will be fed as additional features to the age group classification model, where gender information act as a binary feature. In training there are two models built for gender and age classification. This is further detailed in following testing phase.

3.3 Testing Phase

Step 5 - As similar to training set except Step 4, the test set $p_t = \{d_1, d_2, ..., d_{n1}\}$ also follows all remaining steps to compute feature vectors. Further classification of document into gender and age group is performed using b aggregate predictors in hierarchical method. The final class is assigned based on voting method. The test features $[F_t]$ are initially classified into male/female and padding is done as an additional feature for further age group classification.

The algorithm for training and testing are shown below,

```
Input $p_i = \{d_1, d_2, ..., d_n\}$
for i=1 to n do
    $[V] = VSM\ (d_i)$
end
$[W] = NMF\ ([VV^T])$
$[F] = statistical\ feature([W])$
$model_{gen} = rft\ ([F_{final}, b])$
$model_{age} = rft\ ([F_{final}, gender])$
$y_{gen} = predict\ (model_{gen}, [F])$
$y_{age} = predict\ (model_{age}, [F, y_{gen}])$
```

Algorithm 1: Training and Testing

4 Experiment and Observations

The model diagram for performing AP is given in Figure 1. The data-set chosen for this experimentation is from the PAN CLEF AP 2016 workshop [15][16] which are built with challenges involved in real word applications. The 2016 corpus incorporates Twitter data of users for training and reviews, blog data of the authors taken as test data. This is also included in three languages (English, Dutch and Spain). Among them Dutch data-set does not have age-group information.

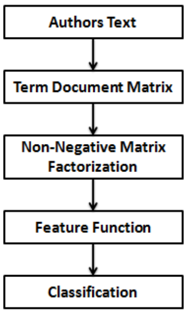

Fig. 1. Model Diagram

During pre-processing, author's text alone extracted for further actions. As detailed in problem definition section, the total documents (Author's tweets) represented as a Document-Term matrix ($m \times n$). This matrix multiplied with its transposed version to get the document-document co-occurrence matrix of size $m \times m$. Further NMF applied on the document-document matrix to get the basis vector matrix (Context Matrix) with r = m. The basis vector of the each author's document considered as a random variable and its correlation with other distribution random variables mentioned in the section 2 are measured as the features. This final, feature matrix is utilized to construct the classification model which is built using Random Forest Tree classifiers. Here the classification model is built based on 100 decision trees, constructed to form the random forest tree. Initially gender classification is performed and by feeding its result to the feature matrix, age group is classified. Same process applied on the three languages without any change. All the above done in Python and its packages (Scikit Learn and Scipy).

Language	**Total #**	**10-Cross Validation**	**10-Cross Validation**
	Documents	**Gender %**	**Age %**
English	436	61.50	53.35
Dutch	250	59.25	–
Spanish	686	63.89	54.60

Table 1. Training Performance

10 fold cross validation is performed to measure the training performance and given in the Table 1. The measures performed on individual (English, Dutch and Spanish) and combined data-sets (English and Spanish). Thought the proposed performance not greater in accuracy, from results it can be observed that, proposed model shows constant accuracy over the all the language and genre. This ensures that proposed model act as the language and domain independent method.

5 Conclusion

With the global need for author profiling system this experimentation has brought forth a simple, unified and reliable model for finding the demographic features of an individual by extracting statistical semantics property of context space. This is achieved by incorporating the Document - Term Matrix, Non - Negative Matrix Factorization and statistical features along with the Random Forest Tree classifier. From the results it can be concluded that this serves as the domain and language independent method, however there is still room for improvement. The future work will be extending and implementing proposed algorithm on distributed computation frameworks like Apache Hadoop and Apache Spark.

References

1. Andrew Perrin: Social Media Usage: 2005-2015. 2015
2. Mangold, W. Glynn, and David J. Faulds: Social media: The new hybrid element of the promotion mix. Business horizons, (2009)
3. Rangel, Francisco, Efstathios Stamatatos, Moshe Moshe Koppel, Giacomo Inches, and Paolo Rosso: Overview of the author profiling task at pan 2013. In CLEF Conference on Multilingual and Multimodal Information Access Evaluation, (2013)
4. Rangel, Francisco, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans: Overview of the 2nd author profiling task at pan 2014. CLEF Evaluation Labs and Workshop, (2014)
5. Rangel, Francisco, P. Rosso, M. Potthast, B. Stein, and W. Daelemans: Overview of the 3rd Author Profiling Task at PAN 2015. In CLEF, (2015)
6. Barathi Ganesh HB, Reshma U, and Anand Kumar M: Author identification based on word distribution in word space. Advances in Computing, Communications and Informatics (ICACCI), (2015)
7. Burger, John D., John Henderson, George Kim, and Guido Zarrella: Discriminating gender on Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2011)
8. Liao, Lizi, Jing Jiang, Ying Ding, Heyan Huang, and Ee Peng LIM: Lifetime lexical variation in social media. (2014)
9. Lpez-Monroy, Adrin Pastor, Manuel Montes-y-Gmez, Hugo Jair Escalante, and Luis Villaseor Pineda: Using Intra-Profile Information for Author Profiling. In CLEF (Working Notes), (2014)
10. Maharjan, Suraj, Prasha Shrestha, and Thamar Solorio: A Simple Approach to Author Profiling in MapReduce. In CLEF (Working Notes), (2014)
11. Turney, Peter D., and Patrick Pantel: From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, (2010)
12. Lee, Daniel D and Seung, H Sebastian: Learning the parts of objects by non-negative matrix factorization. Nature Publishing Group, (1999)
13. Xu, Wei, Xin Liu, and Yihong Gong: Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, (2003)
14. Leo Breiman: Random forests. Machine learning, (2001)
15. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Evaluations Concerning Cross-genre Author Profiling. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016)

16. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PANs Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268299. Springer, Berlin Heidelberg New York (Sep 2014)