Two-Layer Classification and Distinguished Representations of Users and Documents for Grouping and Authorship Identification

Haytham Mohtasseb
School of Computer Science
University of Lincoln
Lincoln, UK
hmohtasseb@lincoln.ac.uk

Amr Ahmed
School of Computer Science
University of Lincoln
Lincoln, UK
aahmed@lincoln.ac.uk

***Abstract*—Most studies on authorship identification reported a drop in the identification result when the number of authors exceeds 20-25. In this paper, we introduce a new user representation to address this problem and split classification across two layers. There are at least 3 novelties in this paper. First, the two-layer approach allows applying authorship identification over larger number of authors (tested over 100 authors), and it is extendable. The authors are divided into groups that contain smaller number of authors. Given an anonymous document, the primary layer detects the group to which the document belongs. Then, the secondary layer determines the particular author inside the selected group. In order to extract the groups linking similar authors, clustering is applied over users rather than documents. Hence, the second novelty of this paper is introducing a new user representation that is different from document representation. Without the proposed user representation, the clustering over documents will result in documents of author(s) distributed over several clusters, instead of a single cluster membership for each author. Third, the extracted clusters are descriptive and meaningful of their users as the dimensions have psychological backgrounds. For authorship identification, the documents are labelled with the extracted groups and fed into machine learning to build classification models that predicts the group and author of a given document. The results show that the documents are highly correlated with the extracted corresponding groups, and the proposed model can be accurately trained to determine the group and the author identity.**

***Index Terms*—authorship identification, similarity detection, personal blogs, users lexicon and representation, keywords extraction**

I. Introduction

The web 2.0 generation opens new directions of use and facilitates collaboration all over the world with a large number of individually written electronic texts available online. The need to authenticate those documents is becoming more important than before as the users have numerous identities in the online world and they may behave differently in each context. Identifying the author of anonymous text messages could be useful in various applications. This includes intelligence, forensic purposes, or online security where it is valuable to extract the groups of authors who may discuss similar ideas, such as terrorism for example.

One of the key problems in practically applying authorship identification is the dramatic drop in success experienced where large number of authors is considered. In this paper, we address this problem and introduce a novel solution in both authorship identification and similarity modelling using our new representation for users. Furthermore, we present a new method of authorship identification via using two classification layers. The function of the higher layer is to predict the group that a given document, written by anonymous author, belongs to. After identifying the potential group, author identification can be applied locally within that group. Identifying an author within a group that contains a limited number of authors is more accurate and practically achievable than doing the classification over the full set of authors.

In this paper, we opt to study authorship identification in personal blogs as it becomes one of the prevalent forms of users' contribution to web content with a large number of individually written electronic texts. It is not always that blogs display opinions, experience, and other useful materials. Sometimes, bloggers publish illegal contents and may run against the law [6]. People who are not socially active, find themselves more comfortable in the online world, and can do many things that are unlikely to be performed by them face-to-face. Many create a variety of personas and identities in several online environments. This anonymity may encourage harmful behaviour by some users as it is hard to catch them [24]. Bloggers may try to hide their identities using various software tools (IP hiding , proxies) and by obtaining ambiguous email addresses. This raises the need for having other technologies that are capable of capturing bloggers identity from text such as authorship identification.

The rest of the paper is organized as following. In section II, we review the recent work related to our domain. The corpus and text properties are described in section III. Section IV shows the groups extraction framework which includes the stages required to build the user representation and extract the groups. In section V, we present our two-layer authorship identification method including the experiments and results. Finally, we show the conclusions and our future work.

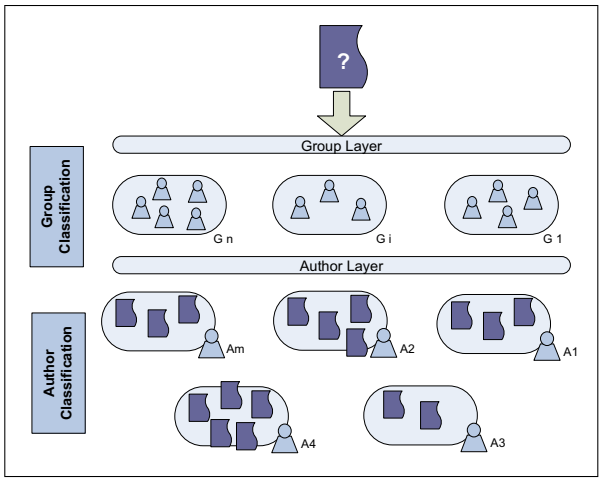

Fig. 1. Authorship identification across two layers of classification, n: number of groups, m: number of authors, and $n << m$.

II. Background

The area of authorship analysis includes three domains: authorship characterization or profiling, similarity detection, and authorship identification. In this study, to address the problem that exists with the increased number of authors, we combine two of those categories: the similarity detection and the authorship identification. We generalize the similarity detection from being detected between two authors to induce the similarity groups across several authors. The extracted groups are then used to build a learning model fed by the documents that have been labelled with the corresponding groups. The other category, authorship identification, is applied first in the higher layer to identify the group to which a given document belongs, then, in the second layer, to identify the potential author of this given anonymous document. In the rest of this section, we review the key related work in each of those three categories.

The third category of authorship analysis, authorship profiling, aims to discover the demographic attributes of the author. Argamon et al. in [3] made a collection of experiments to discover the gender, personality, native language, and age of the author. Identifying the native language of the author has also been investigated by Koppel et al. in [16] using a collection of idiosyncratic features.

While authorship profiling discovers the attributes of an author, the target of authorship identification is to detect the author of a given text. Several techniques and features have been used to identify authors in various text contexts. For example, Argamon et al. [4] utilized multiplicative learning with orthographic features to identify authors in a news group corpus. In a similar corpus, De Vel [7] analyzed stylistics attributes to discover forensics in email texts, while Koppel and Schler [15] depend mainly on misspelling features in addition to other lexical and syntactic sets to identify the author in email text. In the domain of web forums, Abbasi et. al. [1] used a collection of lexical, syntactical, structural, and content-specific features to find out the extreme patterns of writing on web forums.

While most of the studies utilize the same feature sets for all the authors, few studies tried to utilize customized features for each author. Li et al. [17] developed a genetic model to select the best combination of key features for author identification. Koppel et al. [14] introduced a new measure, the feature stability, to select the invariant stylistic features among the documents for a specified author. Furthermore, Abbasi et al. [2] presented the "Writeprints" technique, which separately models the features of each individual author, instead of using one model for all the authors. They build a writeprint for each author using the author's key features.

Similarity detection evaluates the similarity between different text documents regardless the author of text. It is different from authorship identification in that it employs unsupervised learning as there are no previous defined classes (authors) to be detected. Abbasi et al. [2] also used the writeprint methodology to discover the similarity between authors. They utilized, in addition to the individual writeprint features for each author, the usage patterns of other features which are less likely to be used by that author, but are important when comparing an author to another anonymous identity.

The common features in all of the above work are that they have been developed for other types of text, rather than personal blogs text. Blogs text has its own properties. It is more of informal text and contains lots of slang, words imported from other languages, and a high percentage of out-of-vocabulary words (see section III). Gehrke et al. [8] studied the authorship identification in blogs. Using Bayesian classifier, they represent each author using a probabilistic model by calculating the prior probability between the bigrams and the author, producing an individual classifier for each author. Moreover, Mohtasseb et al. have investigated the key parameters [19] and linguistic features [20] that are effective in identifying and capturing the style of authors in personal blogs. They found that the Linguistic Inquiry Words Count (LIWC) [22], is a good candidate feature set to represent the author in personal blogs, and the identification accuracy is better when authorship identification has been applied over a group of authors who share some common features. These two results motivated us to find the similarity groups among authors using the LIWC categories. Modelling users groups from text has received little or no attention as far as we know. We utilize unsupervised learning (clustering) to extract the groups relating similar authors, use supervised learning to build the two classification layers, and apply authorship identification across the two layers. The proposed framework, explained in section V, presents a new solution to the problem exhibited in most of the previous work featuring larger number of authors.

III. Corpus

The style of writing in personal blogs is different from other types of text such as emails, books, or articles. The nature of personal diaries contains the personal print, details of blogger's life, and his or her experience. This type of text is rarely found

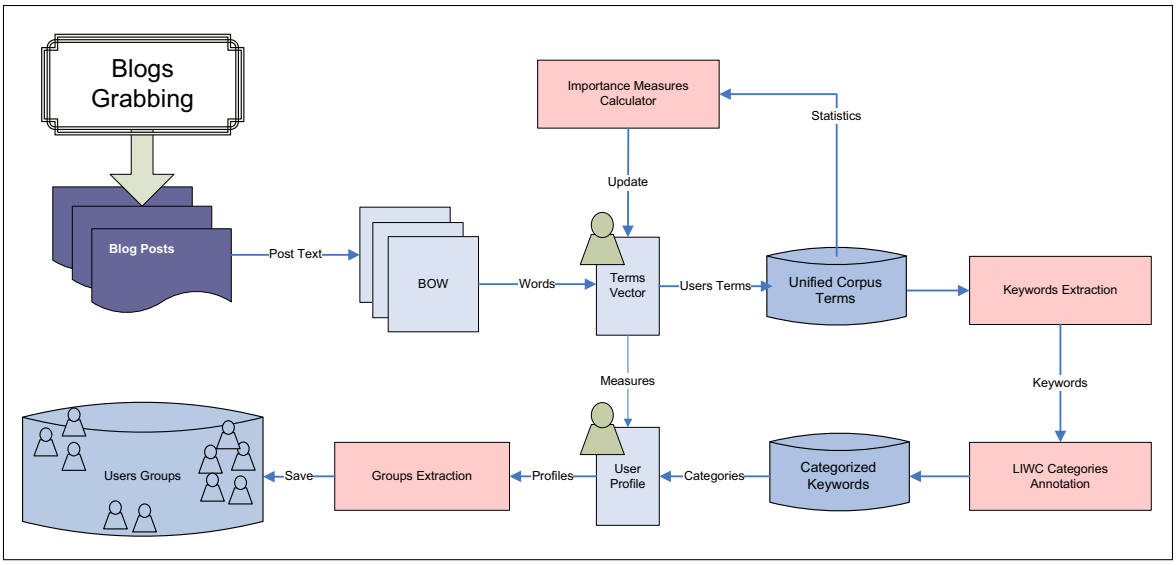

Fig. 2. Groups and Users representation and extraction framework

in other corpora. Usually, in personal blogs, there is no predetermined subject or criteria for specific readers as in news text. Blogs posts are different from emails as they are not written to a dedicated person, but they are available publicly to be accessed by everyone, sharing problems and ideas with friends and others. Our selected text is challengeable as it is informal, self-referential, combining spoken and written English, and rich in unedited content. We chose LiveJournal [1] that is a free personal blog website forming a community on the internet that contains millions of users publishing their own ongoing personal diaries. We downloaded 28183 blog posts for 100 authors with 300 posts as an average for each author. The total number of words is 9,427,293.

IV. Groups and Users Representation

The bottleneck in creating the new layer, the users groups, is to find a proper representation of authors, rather than documents, to be utilized afterwards in clustering. Applying clustering over the documents themselves will end up with clusters that contain the similar documents regardless the authors. Moreover, it is highly expected that each author will also belong to more than one cluster.

In this paper, we present a novel user representation, by creating an individual textual profile for each user based on the keywords of user. The profiles are utilized to extract the similarity and build the clusters. In this section we explain our framework which builds a profile vector for each user passing several stages. Figure 2 illustrates the details of the framework that starts by converting the blog posts to bag-of-words (BOW) representation. Next, the unique terms are extracted across the users and assigned a calculated weight that expresses the degree of importance according to the individual user and to the full set of users in the corpus.

From the weighted terms, a list of keywords is derived for each user that is utilized to create the user corresponding vector. Then the user vector is projected into a new space, the LIWC categories space, which is much smaller in its dimensionality (54 categories) compared to 1500 terms. The LIWC categories are also more descriptive as they have been built with a psychology background (sec IV-D). Afterwards, the user is represented by a vector whose entries express the importance of the selected categories. Finally, clustering is performed over the user profiles to extract the similarity groups. Following the details of each stage will be explained.

A. Posts Text to BOW

The bag-of-word (BOW) representation is commonly used in information retrieval and text mining systems. The document is converted to a list or bag of weighted terms. Generally, the weighting is computed according to the frequency of word appearance inside the document. In order to reduce the dimensionality of term space, stopwords, which are topic neutral words such as articles or prepositions, and the punctuations are removed. Our concern is to find terms that unify the users not differentiate between them, so stemming has been used as an effective way to get back the multiple forms of the word to their base root or stem.

In our work, we removed the stop words according to two lists: the SMART list [2] and Onix list [3]. We used the standard Porter stemming algorithm [23] to produce the roots of the words. Finally, each blog post is represented by a vector of terms and their corresponding frequencies in the post. In the next stage, the BOW is utilized to produce the terms lexicon used by each user.

[1] www.livejournal.com

[2] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[3] http://www.lextek.com/manuals/onix/stopwords1.html

B. User Unique Terms and Importance Measures

Having the unique terms commonly used by the user will give more information about the important terms and the usage frequencies for each user individually. The user local lexicon is beneficial for several applications related to user personal modelling and online behaviour. For each term, we calculate the total frequency among the posts of the user and the number of user's documents which contain this term.

To evaluate the term importance from the user point of view, we define a measure which combines, in addition to the term frequency per user, other important values. The first one evaluates how far a specific term belongs to a particular user, in that it is rarely used by other users. We modified the *TFIDF* (term-frequency inverse-document-frequency) measure, commonly used in information retrieval and text classification, replacing documents by authors to generate (*TFIAF*) **term-frequency inverse-author-frequency measure**.

$$TFIAF(t,a) = log(TF(t,a))log(\frac{|A|+1}{AF(t)})$$

Where: *TF(t,a)* is the total frequency of term '*t*' for author '*a*', *AF(t)* is the number of authors who use term '*t*', and $|A|$ is the total number of the authors (users). *TFIAF* value increases when the other authors are less commonly using this term. This gives more information about the value of a term for a specific author. Similar modification to *TFIDF* have been explored in [5], [13]. We developed the previous measure to a new one, **author term importance measure** *ATI*, which takes into account the normalized documents usage frequency that contain this term for the corresponding author.

$$ATI(t,a) = \frac{TFIAF(t,a)DF(t,a)}{|DA|}$$

Where: *DF(t,a)* is the number of documents for author '*a*' that contain the term '*t*' and $|DA|$ is the total number of documents for author '*a*'.

C. Keywords Extraction

After converting the posts to BOW and deriving the unique terms for each user, the unified terms in the whole corpus have been induced. Each unique term in the corpus is assigned its total frequency among the posts, the total number of users who are using this term, and the total number of documents which contain this term. The extracted terms could be used as a sample of the important terms in the blogs space of "LiveJournal". We developed a new measure which combines the previous three measurement values for each term. **Term importance measure** is defined as following:

$$TI(t) = TF(t).\frac{DF(t)}{|D|}.\frac{AF(t)}{|A|}$$

Where *TF(t)* is the total frequency of term '*t*' in the corpus, *DF(t)* is the number of documents which contain term '*t*', and —*D*— is the total number of documents (posts). Term importance measure takes into account the normalized users usage frequency $\frac{AF(t)}{|A|}$ and the normalized documents usage frequency $\frac{DF(t)}{|D|}$. In order to keep the number of features reasonably small, the larger list of terms, which is about 100,000 terms, have been eliminated to a smaller one that represents the important terms or the keywords in the corpus. Keywords extraction is performed using the *TI* measure by considering the top 1500 terms.

TABLE I
TOP TEN CATEGORIES WITH THEIR TOP FIVE KEYWORDS

Category	Keywords
Affective Processes (AF.P)	love, friend, hope risk, heaven
Relativity REL)	time, day, night move, world
Cognitive Processes (CGN.P)	feel, wait, write decide, remember
Biological Processes (BLG.P)	love, life, head drink, sex
Social Processes (SCL.P)	people, talk, girl meet, mom
Positive Emotion (POS.E)	happy, fun, friend party, kiss
Negative Emotion (NEG.E)	hate,hurt, stupid devil, victim
Perceptual Processes (PRC.P)	watch, cool, voice touch, scream
Work (WRK)	school, class, busy computer, test
Leisure (LSR)	read, play, family weekend, music

D. Dimensionality Reduction and Profile Creation

This stage produces the user representation vector based on the high level categories. Although the keywords for each user were extracted, it is better for machine learning purposes (clustering) to find a new representation with a lower dimensionality. We opt LIWC that is a psycholinguistic-based lexicon and has been shown to have high correlation with the style of authors in personal blogs [20]. LIWC had been used successfully in numerous text analyses tasks for analyzing the emotions of users in blog text [10], [11], [12], identifying the gender of bloggers [21], recognizing the personality [9], [18], and for author identification [19], [20].

Using the LIWC dictionary, each term is annotated with the corresponding categories, as it is possible for the term to belong to several categories. A new dictionary has been constructed that contains the selected 54 LIWC categories with their related keywords. The dictionary is used to represent the author profile as a numerical vector $\vec{a} = (I_{c_1}, I_{c_2}, ..I_{c_i}, .., I_{c_n})$, where n is the number of LIWC categories and I_{c_i} is the importance degree of the category c_i for the specified author. The value of I_{c_i} is produced by summing the user's *ATI* values of keywords related to the category that exist in all documents of the author(the keywords that are annotated with the corresponding category) $I^a_{c_i} = \sum_{j=1}^{k} ATI(t_j, a)$, where '*k*' is the number of keywords related to category c_i. Table I illustrates the top ten LIWC categories according to the average of importance of each category across the authors. The table also shows the top five keywords of each category.

TABLE II
RESULTED CLUSTERS CENTROIDS BASED ON (I_{c_i}) OF THE TOP TEN CATEGORIES, $I_{c_i} \in [0-100]$

Cluster	AF.P	REL	CGN.P	BLG.P	SCL.P	POS.E	NEG.E	PRC.P	WRK	LSR
1	**84**	10.16	**43.83**	14.08	7.35	20.83	**38.36**	3.85	5.05	9.45
2	12.3	13.93	6.8	**83.63**	9.26	**45.23**	4.4	6.1	38.83	10.95
3	13.08	14.95	11.20	8.25	10.08	**67.67**	2.53	3.13	31.90	2.35
4	15.20	11.00	8.90	9.23	8.55	8.15	6.80	4.45	4.60	4.35
5	10.25	9.50	6.35	6.83	**30.38**	5.68	4.53	3.85	4.15	4.03
6	7.53	8.83	4.45	3.50	**29.10**	4.00	3.48	2.55	28.63	32.20
7	15.65	15.93	10.25	9.25	10.08	33.30	7.10	**32.28**	5.90	4.95
8	25.15	**22.45**	14.65	**56.13**	11.50	12.98	12.00	8.68	7.78	**77.68**
9	9.23	8.58	**75.83**	3.98	6.10	4.68	26.98	2.98	**52.83**	28.43

The resulting vectors of the users are ready to be analyzed using the clustering algorithm as explained in the following section.

E. Groups Extraction Results

Extracting the similarity is an unsupervised learning task without any previous defined classes. K-means clustering algorithm has been chosen, implemented in Weka machine learning toolbox [25], to induce the clusters/groups based on our vectors representation of users. There are two parameters that have to be initialized to start the k-means algorithm which are: the prior number of clusters and the seed value. Experimentally, for the 100 users in our corpus, the best values are to set the number of clusters to '9' and the seed to '40'. This produced well distribution of users in balanced clusters as each cluster contained reasonable number of users between 5 and 14.

Table II presents the centroids of the resulted clusters for 10 selected dimensions, as it is not possible to present results for the 54 dimensions (LIWC categories). The value of each dimension represents the importance value of the corresponding category I_{c_i} in the selected cluster. As the dimensions of the centroids are based on the expressive psychology categories of LIWC, the clusters can be described according to the high values of their attributes. For example, we notice that cluster 1 has a high value for ***affective process***, while cluster 9 is more described by the ***work*** and ***cognitive process*** categories. Some of the high values in the table are set in bold. These results are useful to add a high level knowledge that describes each group of users.

In addition to the usefulness of groups in enhancing authorship identification as described in the next section, there are other benefits. This includes developing the blog website to *suggest friends for bloggers* that they are in the same group, or at least notify the bloggers about other people that have something *common* in their writing materials. Moreover, the user interface of each blogger could be *adapted*, in the contents or the displayed ads, according to the distinguishing features of the containing group.

V. TWO-LAYER APPROACH

Extracting the similarity groups, as we shown in the previous section, is an essential step to build the two classification layers. This is one of our contributions to improve the authorship identification accuracy considering the problem of large number of authors. In this section we present our methodology of performing authorship identification using the *two-layer approach*.

In authorship identification, given an anonymous document to find its author, the knowledge utilized in the discussed user representation is not available, as there is no previous information about the author. In other words, this means that it is not allowed to get benefits from the lexicon or the importance measures related to the author. Hence, another document representation has to be developed. A new document vector representation has been built, different from the user representation, based on the normalized frequencies of LIWC categories. Machine learning has been used to build classification models learnt from the vectors of documents that have been labelled by the corresponding group in the primary layer, and by the particular author in the secondary layer. Utilizing the machine learning is bridging the *gap* between user and document representations.

In additions to the enhancement obtained in the identification results with the larger number of authors, by using the groups' layer, we made another improvement in the classification within the group. We extended the vector of the document to include, in additions to LIWC features, other stylistic features that confirmed to give high accuracy in authorship identification in personal blogs. In the next section we present the experiment setup. Section V-B shows the identification results across the two layers.

A. Experiment Setup

The entries of document's vector contain the normalized frequencies of LIWC categories. The category c_i frequency is produced by counting the terms that belong to c_i contained in the document. $|d|$ refers to the total number of terms inside the document $\boldsymbol{d}$.

$$\vec{d} = (\frac{freq(c_1)}{|d|}, .., \frac{freq(c_i)}{|d|} .., \frac{freq(c_n)}{|d|})$$

We used Weka toolbox [25] for machine learning task and chose the Support Vector Machines algorithm (SVM) which is effective in text classification tasks. Working with blog text is challenging as there are no standardization or rules for the textual content. Hence, we intended to divide the documents

vectors in the corpus into 10 dataset according to 10 different post lengths. For each dataset (from the 10 datasets), we evaluated authorship identification across the classifiers of the two layers using 10-fold cross validation. In each fold, 90% of the dataset is used for training and the remaining 10% is used for testing. In the next section, we show the experimental classification results of the two layers using the overall accuracy across the 10 folds according to the selected testing parameters.

Having the appropriate features is a key for the success in achieving high classification accuracy. Although we used only LIWC to model the users and extract the similarity group among them, we intend in this experiment to add more stylistic features that can represent more the different styles across the documents of author. For group and author identification which are based on a single document, we need to get more linguistic features that distinguish a single document. We select to utilize two more feature sets [20] that have been used before on authorship identification in personal blogs and gave high identification accuracy. The first feature set is the *syntactic* features set which count the number of sentences, words, abbreviations, and punctuations. In contrast to what we did in user representation, we keep the punctuations to use this features set. The second features set is the *misspelling errors characterization* which extracts couple of features that represent various categories of misspelling behaviours of the author. We repeat the previous experiment, extending the numerical vector of document by adding the new features. In the following section, we show the accuracy results of the two experiments.

B. Results

Figure 3 shows the accuracy results of group identification according to the utilized features. First note, that using LIWC in all cases, we obtained classification accuracies higher than 50% in most ranges, even with short documents, establishing the validity of the method by predicting the group using only single document. The results confirm that having more words inside the query's document improved the identification result up to 72%. Adding the two feature sets is clearly enhanced the identification results raising the accuracy to exceed 80% for group identification in some ranges.

Moving to lower level classification inside the group, to capture the particular author, we can see in table III, using only LIWC, that in addition to the post length parameter, the number of authors within the group is affecting author identification result, which emphasizes one of the motivation of our research. The clustering results ended up with different numbers of authors inside the resulted groups between 5 - 14 authors. The drop in the authorship identification result is normal, but we also addressed this and improved the identification result within the group by adding two more feature sets as shown in table III. The full sets of features bring the minimum accuracy above 60% in the worst cases for the two layers. Moreover, according to our design and the distribution of users within the groups/clusters, we have an upper limit in the number of authors to not exceed 14 authors and stop further decline in the result of author identification by dividing the authors into groups. This allows us to deal with 100 users and extend for larger number, which is a limitation in all previous work in this field.

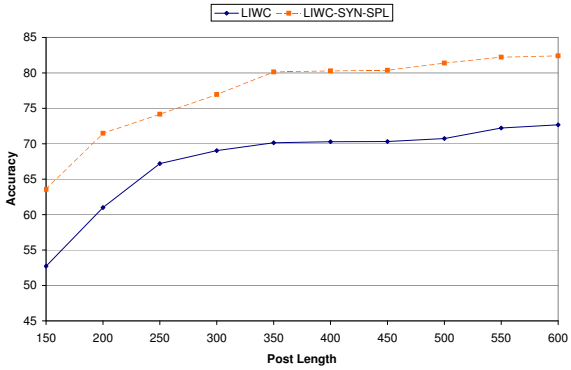

Fig. 3. Group identification overall accuracy according to features

TABLE III
EXPERIMENTAL RESULTS (% ACCURACY) FOR AUTHOR IDENTIFICATION WITHIN GROUP

L	LIWC				LIWC-SYN-SPL			
	No. Authors				No. Authors			
	5	8	11	14	5	8	11	14
150	62.2	59.0	50.5	48.0	76.0	69.1	63.4	60.1
200	66.9	59.6	56.3	50.9	76.5	74.6	69.5	63.6
250	76.5	62.3	58.6	52.8	84.3	77.2	71.3	65.4
300	76.8	65.6	60.9	53.1	87.8	77.7	74.9	66.6
350	77.6	66.6	63.8	56.4	88.2	80.7	75.2	71.0
400	78.2	69.0	64.2	60.2	88.8	81.9	76.0	72.5
450	78.8	70.3	64.3	60.5	88.8	83.8	77.0	72.8
500	81.9	74.3	64.8	62.1	89.0	85.0	79.5	76.4
550	82.5	78.0	68.5	64.2	89.6	85.1	80.3	77.0
600	84.0	78.1	70.8	66.5	91.6	87.5	81.0	77.3

VI. CONCLUSIONS AND FUTURE WORK

In this paper we present our novel distinguished representations of users and documents for grouping and authorship identification. This facilitates constructing our two-layer framework that enables us to apply authorship identification over larger number of authors (100). The framework is also easily scalable to deal with more than 100 authors. Most of previous work reported a dramatic drop in the identification result when the number of authors exceeds 20-25.

The proposed two-layer solution divides the large number of authors into smaller groups that contain reasonable number of authors (5 - 14 authors) and modify the classification to be performed across two stages, by attributing first the appropriate group and then identifying the particular author within the group. The two-layer approach noticeably enhances the identification results compared with the previous studies, especially with allowing for large number of authors to be considered which is a common limitation in most of the previous work in literature. The framework is improving the authorship identification in two sides. The first one is by

stopping the decline in the classification result via limiting the number of authors to not exceed 14 authors inside each group. The second one is by improving the authorship identification accuracy within the group via adding the appropriate stylistic features to the document representation.

Extracting the groups across the authors is not easy, as there are different contents and styles contained in the documents of authors. In this paper, we introduced our novel user representation extracted from all the documents of each author, and apply clustering over users' vectors to extract the similarity groups. The vector of user was constructed based on a variety of measures resulted from the lexicon, keywords, and level of importance of each user.

The groups/clusters are better described and more understandable as the features of the user vector are psychology based. This high level description of groups give an indication about the orientation of bloggers community and is useful in other applications like customized web interfaces and advertisement adapted by the group of the blogger. Moreover, the proposed system is useful for security purposes to detect the increased unsafe anonymous identities in blogosphere. There are more signals about the development of blog space to contain extreme and terrorism patterns raising the risk effect of that online environment.

As for future work, we are continuing our experiments to consolidate and enhance our framework and its scalability, with the increase in the number of authors. We are also investigating the various parameters such as the number of clusters and number of users per cluster, and the alternative techniques to better enhance the performance of our framework. This includes using different clustering algorithms and adding more extra layers if needed. We also plan to include semantic representation of the authors to enhance clustering and similarity extraction. Furthermore, we explore the possibility of testing our system on other corpora as we believe that our two-layer framework is extendable and applicable to other corpora and domains. However, we need to find the appropriate feature sets, mainly for grouping, for the selected corpus. Finally, we are planning to get benefits of the extracted groups to build individual feature sets for each group that will improve authorship identification within the group.

ACKNOWLEDGMENT

The authors would like to thank Dr. David Cobham for his cooperation and help to produce this work.

REFERENCES

[1] A. Abbasi and H. Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, pages 67–75, 2005.

[2] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions Information Systems*, 26(2):1–29, 2008.

[3] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52, 2009.

[4] S. Argamon, M. Saric, and S. S. Stein. Style mining of electronic messages for multiple authorship discrimination: first results. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480, 2003.

[5] S. Chan, R. K. Pon, and A. F. Cardenas. Visualization and clustering of author social networks. In *Distributed Multimedia Systems Conference*, pages 174–180, 2006. http://www.cs.ucla.edu/ cardenas/cardenas2.html.

[6] G. S. Dardick, C. R. La Roche, and M. A. Flanigan. Blogs: Anti-forensics and counter anti-forensics. In *Proceedings of The 5 th Australian Digital Forensics Conference*, page 199, 2007.

[7] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *ACM SIGMOD Record*, 30(4):55–64, 2001.

[8] G. T. Gehrke, S. Reader, and K. M. Squire. Authorship discovery in blogs using bayesian classification with corrective scaling, 2008.

[9] A. Gill. Personality and language: The projection and perception of personality in computer-mediated communication, 2003.

[10] A. J. Gill, R. M. French, D. Gergle, and J. Oberlander. The language of emotion in short blog texts. In *Proceedings of the ACM 2008 conference on Computer supported cooperative work*, pages 299–302. ACM New York, NY, USA, 2008.

[11] J. T. Hancock, K. Gee, K. Ciaccio, and J. M. H. Lin. I'm sad you're sad: emotional contagion in cmc. In *Proceedings of the ACM 2008 conference on Computer supported cooperative work*, pages 295–298. ACM New York, NY, USA, 2008.

[12] J. T. Hancock, C. Landrigan, and C. Silver. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 929–932. ACM New York, NY, USA, 2007.

[13] Y. He, S. C. Hui, and A. C. M. Fong. Citation-based retrieval for scholarly publications. *IEEE Intelligent Systems*, 18(2):58–65, 2003.

[14] M. Koppel, N. Akiva, and I. Dagan. Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology*, 57(11):1519–1525, 2006.

[15] M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, pages 69–72, 2003.

[16] M. Koppel, J. Schler, and K. Zigdon. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628. ACM New York, NY, USA, 2005.

[17] J. Li, R. Zheng, and H. Chen. From fingerprint to writeprint. *COMMUNICATIONS- ACM*, 49:76–82, 2006.

[18] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.

[19] H. Mohtasseb and A. Ahmed. Mining online diaries for blogger identification. In *The 2009 international conference of Data mining and Knowledge engineering (ICDMKE'09)*, pages 295–302, 2009.

[20] H. Mohtasseb and A. Ahmed. More blogging features for author identification. In *Proceeding of the 2009 international conference on Knowledge discovery (ICKD'09)*, pages 534–539, 2009.

[21] S. Nowson and J. Oberlander. The identity of bloggers: Openness and gender in personal weblogs. In *Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs*, 2006.

[22] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc. *Mahway : Lawrence Erlbaum Associates*, 2001.

[23] M. Porter. The porter stemming algorithm. *Accessible at http://www.tartarus.org/martin/PorterStemmer*.

[24] N. Willard and D. JD. Educator's guide to cyberbullying addressing the harm caused by online social cruelty. *Accessible at http://cyberbullying.org*, 19, 2005.

[25] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2005.