Time Profiles for Identifying Users in Online Environments

Fredrik Johansson
Swedish Defence Research Agency (FOI)
Stockholm, Sweden
Email: fredrik.johansson@foi.se

Lisa Kaati
Swedish Defence Research Agency (FOI)
Stockholm, Sweden
Email: lisa.kaati@foi.se

Amendra Shrestha
Uppsala University
Uppsala, Sweden
Email: amendra.shrestha@it.uu.se

***Abstract*—Many people who discuss sensitive or private issues on web forums and other social media services are using pseudonyms or aliases in order to not reveal their true identity, while using their usual accounts when posting messages on non-sensitive issues. Previous research has shown that if those individuals post large amounts of messages, stylometric techniques can be used to identify the author based on the characteristics of the textual content. In this paper we show how an author's identity can be unmasked in a similar way using various time features, such as the period of the day and the day of the week when a user's posts have been published. This is demonstrated in supervised machine learning (i.e., author identification) experiments, as well as unsupervised alias matching (similarity detection) experiments.**

I. Introduction

An increasing amount of many people's life is spent online. People are using Internet and social media in order to communicate, express their opinions and beliefs, discuss topics of interest to them, etc. While much of the information is expressed publicly, there is also more sensitive information available in web forums and other social media services that potentially could be harmful to the author if it became widely known who the physical person behind the user that is posting information is in reality. There are many examples related to the analysis of terrorist activities on the Web (see e.g., [1], [2]), such as the spreading of extremism propaganda and discussions on how to make improvised explosive devices. In such settings, it can be of fundamental importance to intelligence analysts to find out what a person writes and who the physical person behind some pieces of texts really is. This is the main motivation and driving factor for the research presented in this paper. However, the Web is, fortunately, not only used for activities related to terrorism. Ordinary citizens may also want to preserve their anonymity when discussing private issues such as religion, sexual preferences, political ideas, diseases, etc. in public. Obviously, what is considered as private and sensitive information varies from country to country and individual to individual. Many people would like to be able to freely express their ideas and beliefs, while at the same time avoid revealing their true identity to e.g., friends, employers, police, or intelligence services.

A common approach to preserve anonymity is to create user accounts or aliases with no obvious connection to a person's true identity and to make use of this when discussing "sensitive issues", while using their usual user accounts when posting information they consider to be non-sensitive. A rather obvious problem with such an approach is that the used Internet service provider and social media service can log the used IP-address and identify the user from this information, unless the providers can be fully trusted by the user or if extra counter-measures are applied, such as logging in from various Internet cafes or making use of tools such as Tor[1] [3]. A less obvious problem is that it can be possible to reveal the user's identity from his or her writing style. It has for a long time been known that stylometric techniques can be used to identify an author among a small set of candidate authors given a large enough data material, but more recent research experiments presented in [4] suggest that this can be accomplished with reasonable accuracy also on large-scale datasets. A user who is aware of such techniques can in theory obfuscate their writing style intentionally, but this is probably quite unusual.

In previous work [5], we implemented a subset of the features suggested in [4] and used them for alias matching (i.e., the problem of identifying multiple aliases belonging to the same individual in an unsupervised fashion). In addition to the use of stylometric features we also used *timeprints* to increase the possibility to detect users with multiple aliases. A timeprint is a property that reflect something about the characteristics of an individual's activity. In our previous work the timeprints were based on the publishing times when a user post messages, capturing the distribution of messages over the hour-of-the-day. By using timeprints in combination with stylometry, the detection rate of finding multiple aliases increases significantly [5].

In this paper, we are exploring various time features in more depth in order to increase the quality of the used timeprints. More specifically, this is accomplished by explorative studies and experiments using the ICWSM forum dataset, containing data from the Irish forum site https://www.boards.ie/. We show that a set of time features can be powerful for unmasking an author's identity in both a supervised (author identification) and an unsupervised (alias matching or similarity detection) setting. We also use feature selection methods to find out which the most informative features are.

The rest of this paper is structured as follows: In Section II, we present various time features which potentially can be useful components of a timeprint, and show how the time features are varying among different users and over time. Moreover, we explain the concept of "circadian topology" or

[1] https://www.torproject.org/

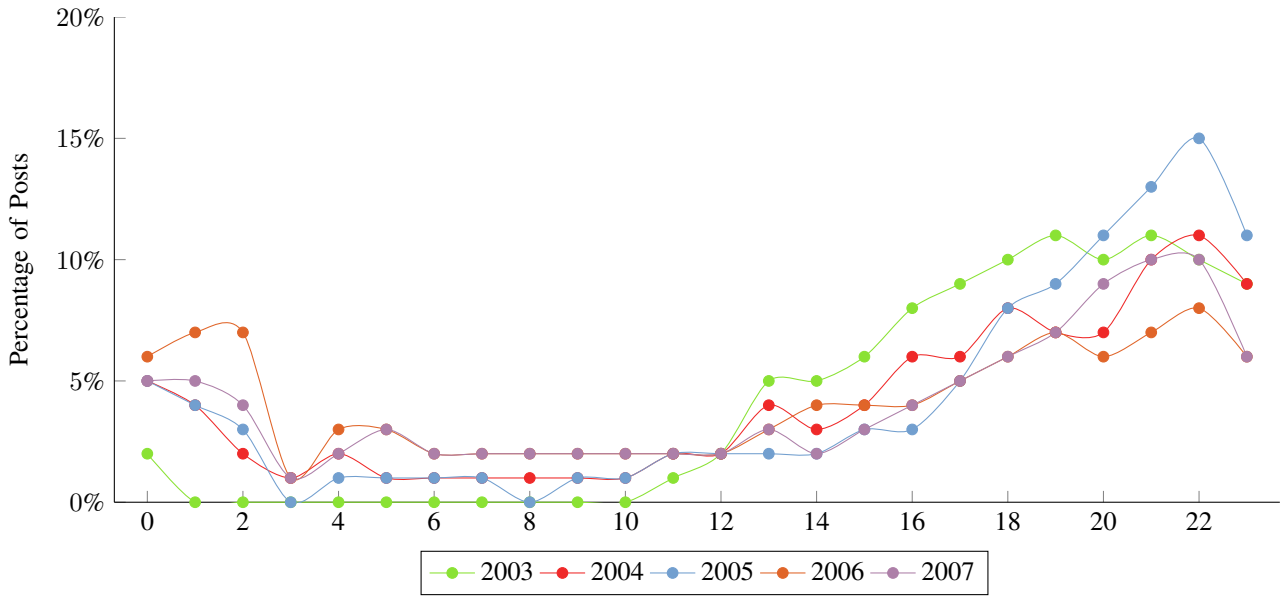

Fig. 1: Distribution of messages for a single user throughout the day for five years

"chronotype" as a motivation for why people can be expected to have timeprints which are different from other individuals' timeprints. This section is followed up with more systematic machine learning experiments, presented in Section III. In these experiments, we evaluate how well a classifier can learn to predict the correct author or user among a larger set of potential candidates by using time features. Hence, this is an example of how the classic problem of author identification can be tackled using non-textual features. In Section IV, the problem of alias matching is described. In the case of alias matching we compare each user identity to all other identities and group together users (aliases) which are more similar than a certain threshold. We have also identified the most informative features (as calculated by using information gain) and use them in the unsupervised problem of alias matching. In Section V, we briefly discuss under which circumstances the obtained results can be expected to hold in "the wild" and which implications our experimental results are likely to have. Finally, we present some conclusions and directions for future work in Section VI.

II. Timeprints and activity profiles

A chronotype or a circadian typology is an individual difference in personality, which is believed to be the cause of why some individuals prefer to work and exercise in the morning hours while other prefers evening hours. Such circadian preferences are based on genetic influence. The circadian typology classifies individuals according to three different types: morning-type, evening-type, and neither-type. The existence of such a circadian typology has been validated in several studies and several countries [6]. The circadian typology seems to have an impact on the behavior of an individual and various studies have for example suggested that evening types spend more time in front of the screen [7] or that evening types have a higher tendency for cigarette craving and alcohol usage [8].

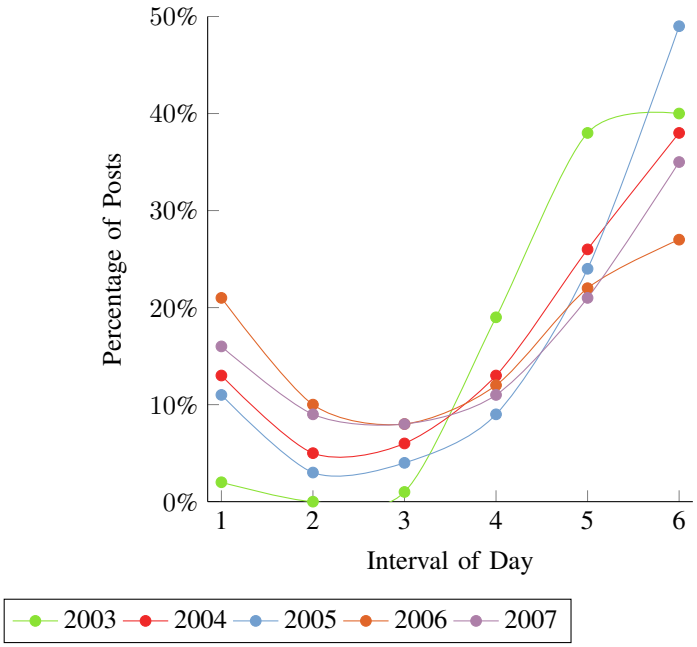

Fig. 2: Distribution of messages for a single user in four hour intervals

When we are active on social media services may be correlated to chronotypes. In a discussion forum such as boards.ie, it is possible that a morning-type individual post most of his/her messages during the morning hours, while an evening-type prefers posting messages during the evening. We believe that the chronotype is something that can be used as a mean to identify a Internet user, or rather tell different

clusters of users apart. In addition to chronotypes, there might be other distinguished features that are important and that are characteristic for a user. Examples of such characteristics are that people live in different time zones, have different working and sleeping hours, goes on vacation every summer etc.

To obtain an understanding on how users behave, and investigate if there are some features that seems to be more characteristic than others, we studied the activity of a set of randomly selected users that were active in the discussion board. Figure 1 shows a random user's distribution of posts throughout a day. As can be noted if the figure, it seems to be the case that this user has a quite similar behavior of when he/she is active and not throughout all five years that are compared. When analyzing users' activity during various time intervals we noted that the activity pattern or time profile of a user seemed to be quite specific for each of the selected users. We also noted that the activity of a user seemed to be consistent over time (using data from different years). Some of the features that we considered in our manual analysis were:

- Activity during each month
- Activity during each hour of the day
- Activity during weekdays and weekends
- Activity during four-hour intervals (early morning, morning, midday, evening, night, midnight)

Figure 2 shows the distribution of messages (posts in the discussion board) in four-hour intervals for a single user in the boards.ie dataset. Each number on the x-axis represents a four hour interval where 1 = 00-03.59, 2 = 04-07.59, and so on. The same user has been active during 2003, 2004, 2005, 2006 and 2007, and the distribution of messages is shown for each year. As can be noted in the graph, this user seems to have a consistent behavior when it comes to distribution of messages over the years.

Using activity profiles to identify individuals is something that has been considered in previous studies. In [9], a temporal analysis of the bloggosphere is done. The assumption was that each blogger has a different preference for posting. A dataset consisting of nearly 700,000 blog articles was analyzed according to two factors: (1) day of the week and (2) time of the day. One of the conclusions in the paper is that each blogger has a different temporal preference for posting which supports our thoughts that different discussion board users have different preferences for posting, and therefore will have timeprints that differ from each other.

III. Author identification

Author identification, also known as authorship attribution, can be defined as the problem of assigning a text of unknown authorship to one candidate author, given a set of candidate authors for whom texts of undisputed authorship are available [10]. Authorship identification is a fairly well-studied problem, where algorithms and various features have been extensively described in, e.g., [11], [4], [12], and [13]. However, existing approaches rely on linguistic/stylometric features (lexical, syntactic, idiosyncratic, etc.), while we here study the usefulness of time features based on when texts have been written or published. To the best of our knowledge, time features have not previously been used for author identification purposes. Clearly, information about time is not always available, but when analyzing posts from social media (e.g., Twitter, web forums, etc.), such information can often be extracted.

A. Experiments on author identification

From the ICWSM boards.ie forum dataset, we have identified and extracted the posts for the top-1000 posters from year 2007. The reason for choosing those users is that we wanted to have as large data material as possible, since a reasonable assumption is that the amount of data will have an impact on the achieved results. Each user u_i has been split into five[2] "sub-users" $u_{i1}, u_{i2}, \ldots, u_{i5}$, where the user's first post has been assigned to u_{i1}, the second post to u_{i2}, etc. The reason for using this approach is to construct several (five) training instances for each user in order to facilitate the learning phase. Based on the extracted posts, timeprint vectors have been constructed (one for each sub-user), consisting of the following sets of attributes:

- **Hour Of Day:** Hour1, Hour2, ..., Hour24,
- **Period Of Day:** MidNight, EarlyMorning, Morning, MidDay, Evening, Night,
- **Month:** Jan, Feb, ..., Dec
- **Day:** Sunday, Monday, ..., Saturday
- **Type Of Day:** WeekDay, WeekEnd

In the construction phase we first count the number of occurrences of each attribute and then express the values as relative frequencies, so that the values of each set of attributes sums to 1 (e.g., $WeekDay = 0.65$ and $WeekEnd = 0.35$). In addition to the features described above, we also incorporate the UserID u_i as the target class. Hence, we have five (different) data instances for each UserID.

In our first experiment, we have varied the number of potential authors from 100 to 1000 in steps of 100, and compared the accuracy for two popular supervised learning algorithms: a naive Bayes (NB) classifier and a support vector machine (SVM) classifier. For the SVM, we have made use of a linear kernel since this was shown to give better results than a radial basis function in our initial experiments. In each step we have performed 10-fold cross validation and the results from the ten folds have been averaged into a single accuracy value. The results from the experiment are shown in Figure 3.

As can be seen, both classifiers perform relatively well on the classification task. The SVM classifier is consistently outperforming the NB classifier, but this comes with a price. The training and evaluation phase of the NB classifier took a few minutes while the last steps took days to perform for the SVM classifier on the standard computer we used for the experiments.

Using the timeprint that only contains information about the activity of a user, the correct user is almost always selected when having 100 potential authors, and the accuracy is still over 90% for the SVM classifier when increasing the number

[2]The number of slices has been arbitrarily selected, but turned out to work well.

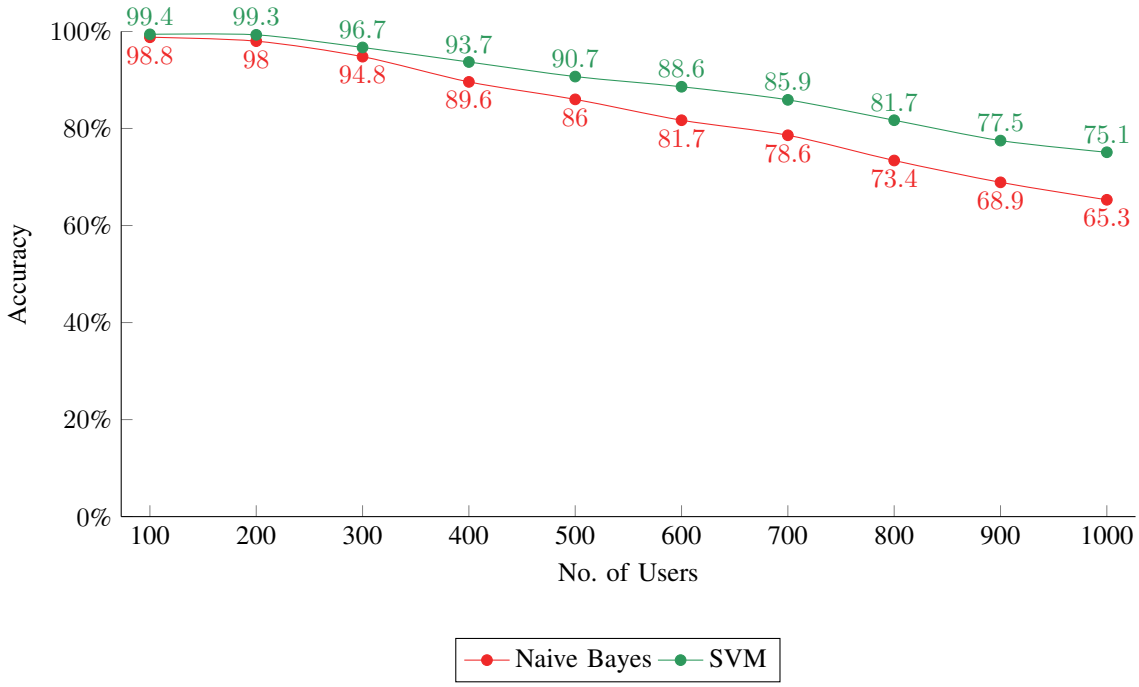

Fig. 3: Accuracy results from classification with a naive Bayes classifier and a support vector machine (SVM) classifier. Experiments conducted on the 1000 users that made most posts in year 2007.

of users to 500. The accuracy decreases when even more potential authors are added, but it is still over 65% when reaching 1000 potential authors also for the quite simple NB classifier, while the corresponding accuracy for the SVM classifier is 75%. The achieved results imply that time features can be very useful for author identification when having access to large amounts of data material. Those results are significantly higher than those obtained for author identification with textual (stylometric) features on a forum dataset reported in [11]. It should however be noted that it is not the same forum datasets that have been used in those experiments.

An important part of the explanation to the decrease in accuracy when the number of potential authors is increased is obviously that there are more candidates to chose among for the classifiers, but a contributing factor may also be that there is less data material for the users further down in the list. To get a better understanding of what impact the amount of posts has on the results, we have in a second experiment modified the original dataset so that each user's timeprint vector is built from the user's first 444 posts instead of all its posts (the threshold has been based on the amount of messages posted by the thousandth user). When adjusting the experiment in this manner, the results shown in Figure 4 were obtained.

As seen in Figure 4, the results become significantly lower when adjusting the available data material in this way. This suggests that the success of the used features are sensitive to how much posts that are available. However, the obtained results are still highly compatible compared with the results presented for stylometric features in [11] (this comparison is only valid for 100 users since Abbasi and Chen did not test the algorithms on larger problem instances). As can be seen, the SVM classifier is outperforming the NB classifier also in this experiment.

IV. Alias matching

In the author identification problem we compare each anonymous user (the users present in a test set) to a fixed set of pre-defined known entities (the UserIDs present in the training set). In this way, we assume that the anonymous user is one of the exhaustive list of candidate authors present in the training set. However, in an alias matching setting (described in more detail in [5]), we can't assume that we have knowledge of all potential authors. The problem is instead to compare each anonymous identity to all other identities and group together users (aliases) which are more similar than a certain threshold. Hence, while author identification can be seen as a supervised machine learning problem, alias matching is an unsupervised problem where the same algorithms cannot be used. Instead, we are for the alias matching problem using a distance function (in this case Manhattan distance) to calculate the similarity among user profiles.

In addition to publishing time, i.e., timeprints, there are also other kinds of features which can be used for alias matching. One obvious candidate is the use of stylometric features. In some contexts, it can also be useful to include string-based features (for matching based on alias names) and social network-based features (for matching based on thread or friend information). The usefulness of such features is described in more detail in [5] and [14]. However, in the experiments presented here we restrict our focus to time-based features.

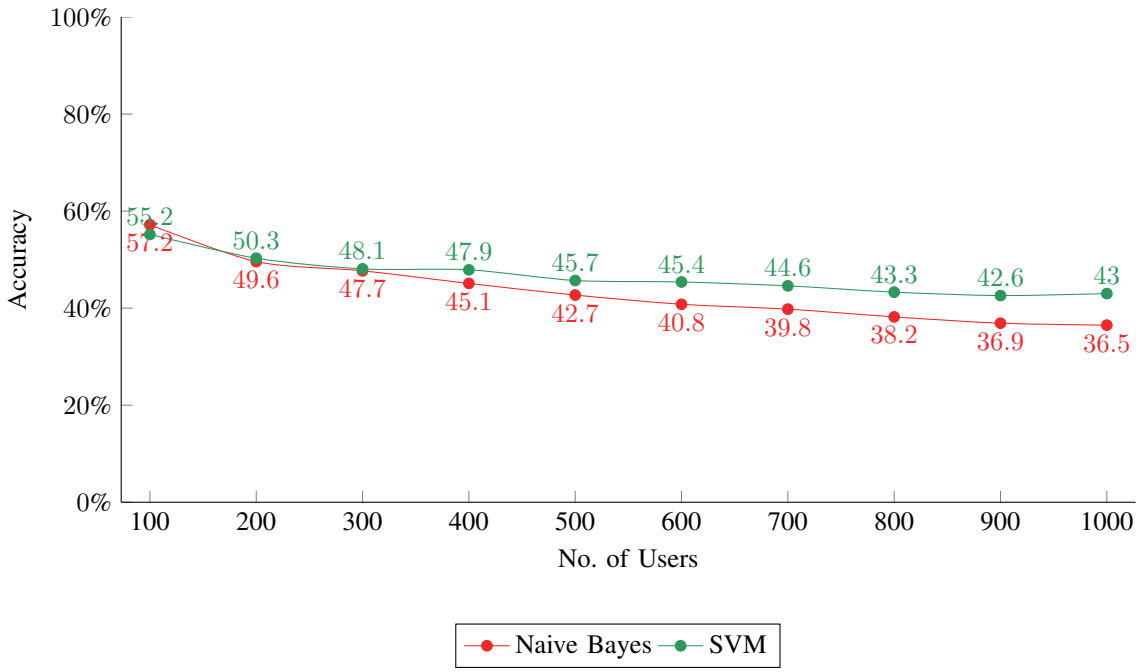

Fig. 4: Accuracy results from classification with a naive Bayes classifier and a support vector machine (SVM) classifier. Experiments conducted on the 1000 users that made most posts in year 2007, using only the first 444 posts.

A. *Experiments on alias matching*

In our first alias matching experiment we have chosen the same set of users as in the previously described author identification experiments. From this set of users, we first selected a smaller set of users ($n = 100$) (where the selection is based on the descending order of the users' amount of posts). Each of these users have been split into two separate users u_{ia} and u_{ib}, where $1 \leq i \leq 100$ and odd posts are assigned to user u_{ia} and even posts to u_{ib}. Now, each user in the set $\{u_{1a}, u_{2a}, u_{3a}, ..., u_{na}\}$ is compared, one at a time, with all the users in the set $B = \{u_{1b}, u_{2b}, u_{3b}, ..., u_{nb}\}$. Based on the results from the time-based matching we rank the members of set B according to how similar they are to the selected user (where the similarity among two vectors is calculated using Manhattan distance). The most similar member of the set B is ranked as number one, the next most similar as number two and so on.

The reported accuracy is calculated as the fraction of times the index of the selected alias is found within the top-N rankings (where the results for $N = 1$ and $N = 3$ are reported). This kind of experiment has then been conducted for increasing values of the number of users n, where we have varied n from 100 to 1000 in steps of 100.

The used methodology is the same that previously has been reported in [5], except for that we now use a threshold of a minimum of 200 posts instead of 60 posts and use data from 2007 instead of 2008. The motivation for those changes is to study the impact of the number of posts on the results.

The results from the experiment are summarized in Figure 5. As can be seen, high accuracies are obtained, irrespectively if we look at the top-1 or top-3 statistics. Expectedly, there is a decrease in accuracy as the number of users is increased, but the top-1 accuracy is almost 80% for 1000 users and the corresponding top-3 accuracy is almost 90%. These numbers are significantly higher than the results reported in [5] (where the corresponding numbers for the time profile were 33% and 47%, respectively), which once again shows the importance of having a lot of data available when building the time profiles. A change from Euclidean to Manhattan distance has also been improving the results somewhat, but only a few percentages.

In our last experiment, the same experimental setup as in our former experiment has been used, except for a change in the time features that have been utilized. Instead of just using the hour-of-day features, we have initially included a large range of features (the same as described in Section III-A). In next step, we have ranked the usefulness of the features using information gain, which is an entropy-based feature selection method [15]. The results from the information gain indicated that there were two sets of time features that seemed to be most useful:

- **Period of Day**
- **Month**.

One possible explanation to why the feature **Period of Day** is useful is that dividing the day into four hour blocks captures the chronotype of a user better than considering the activity of a user each hour. The feature **Month** captures the difference in behavior over a longer time period. Since the dataset we have used is an Irish forum, we can assume that most active users have their origin in Ireland and that the country's seasonal shifting may influence how much time is spent in front of the computer. In the summer it is common to have vacation and in some cases that affect the behavior of a user significantly.

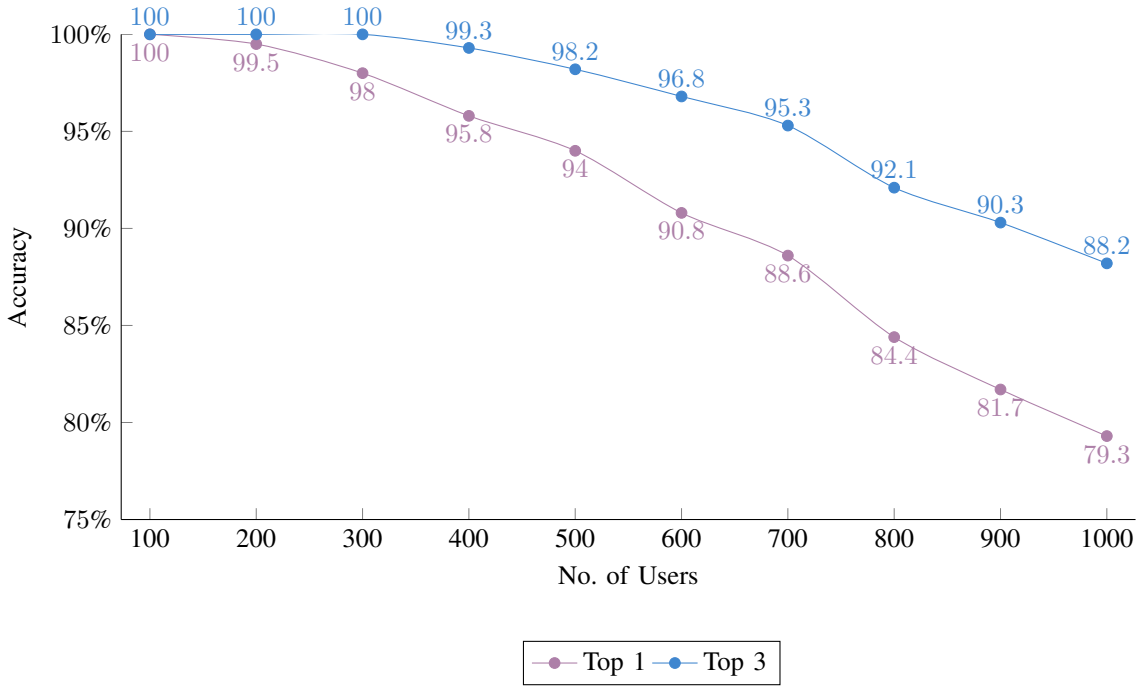

Fig. 5: Accuracy result from experiments with alias matching using a timeprint with the feature **Hour Of Day** as in [5]

When the features **Period of Day** and **Month** are used the accuracy becomes as high as 100% for all the tested user sizes (in both top-1 and top-3 statistics). This is a very strong result and a comparison with the use of a timeprint with the feature **Hour Of Day** (top-1 statistics) is shown in Figure 6. As can be noted, the accuracy for using the features **Period of Day** and **Month** outperform the feature **Hour Of Day**.

V. Discussion

The experiment results presented in the previous sections indicate that timeprints can be very useful for both author identification and alias matching. However, it should be noted that the results have been obtained in quite well-controlled experimental settings which does not necessarily hold true in a real-world environment. In our alias matching experiments, we have been able to control so that posts have been evenly distributed among sets A and B. "In the wild", posts from two aliases belonging to the same individual could potentially be more unevenly distributed, so that the posts for one alias have been created during a completely other time period than the posts for the first alias. If this would be the case, this would have a negative impact on the obtained accuracy.

For the author identification problem, we have split the available posts for a user into five separate training samples. This has proved to work quite well, but the number of training samples per user has been arbitrarily selected. The optimal value of training samples is probably dependent upon the number of potential authors as well as how much posts we have available for each user, but finding such an optimal value has been outside the scope of this paper. However, as a rule of thumb, the more posts we have for a certain user, the more high-quality training samples we can create.

One positive interpretation of the obtained results is that police and intelligence services around the world can become more effective in finding the author of large quantities of terrorist propaganda and other crime- or terrorism-related content. A more negative interpretation is that the anonymity of ordinary citizens in worst case may be weakened. This raises the question of whether the use of time-based features can be defended against by an individual who wants to preserve his or her anonymity. A potential solution could be to use software which does not publish posts directly as they are written, but rather delay the creation time of new posts randomly.

VI. Conclusions and Future Work

In this paper, we have presented the idea that a user's timeprint (which can be extracted from the publishing times of a large number of social media services) can be useful for identifying users who make use of multiple aliases. This idea has been motivated by arguments such as the existence of individual differences in personality preferences related to time (morning-type, evening-type, neither-type) and the fact that people have different working hours and sleeping hours. By selecting a few users and looking at their behavior over time we have noted that many users seem to have a quite stable activity behavior over time. This information can be captured in what we refer to as a timeprint. Our initial manual analysis has indicated that there might be a possibility to tell individuals apart based on their timeprints. However, by just looking at a set of users' behavior over time we can not say much about how unique a timeprint is.

To get a better understanding of the uniqueness of individuals' timeprints, we have made supervised machine learning experiments where we have attempted to learn classifiers to

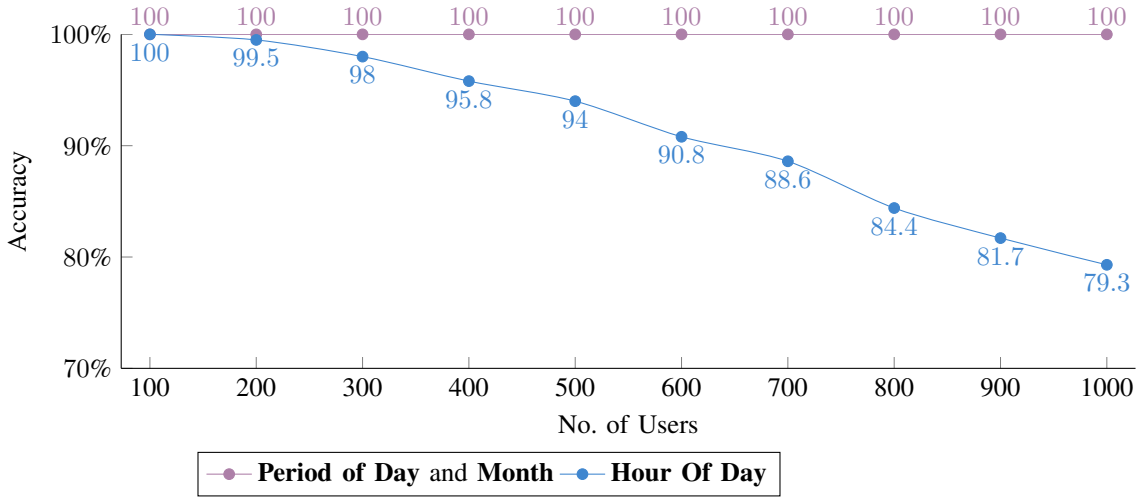

Fig. 6: Accuracy for alias matching using two different timeprints. One with the features **Period of Day** and **Month** and one with the feature **Hour Of Day**.

tell users apart based on various time features. This can be thought of as author identification based on activity rather than textual style. The results suggest that high accuracy can be obtained also for large number of potential authors (over 90% up to 500 users), but that the accuracy is highly dependent upon the number of posts from which the timeprints are created. In a second set of experiments, we have tested the usefulness of various time features for the unsupervised alias matching problem. We show that by using information gain and selecting the most informative features, good performance can be achieved. Once again, the results are highly dependent on the amount of posts that are used to construct the timeprints.

The results in the paper are encouraging from an intelligence and security perspective, but they might pose a threat towards privacy and online anonymity. If this kind of techniques can be used to reveal the true identity of a potential terrorist, there is a risk that the same techniques can be used also for other purposes, even though the usefulness of the technique decline as the number of users is increased. If time features are combined with textual features (such as in [5]), the classification accuracy can be expected to become higher than what has been reported here. One way to defend against the use of ”timeprint attacks” could be to use tools that automate the process of publishing. A more drastic defense could be that some individuals choose to stop posting sensitive information at all, but this would obviously have potentially severe consequences for democracy and individuals' right to freedom.

A. *Future work*

In this paper we have only considered users in a discussion forum, but it is possible that the results can be transferred to other social media services as well. As future work we plan to test the usefulness of the developed timeprints on other social media services such as Twitter. We also aim at cross-platform experiments, in which correlations among discussion forums and other social media services can be explored. We also would like to carry out large-scale experiments like those in [4], where the full set of their stylometric features are combined with the timeprint features developed in this paper.

Another direction for future work is to do more experiments on how much data that is needed to create a timeprint that is useful for identification of users. Our experiments shows that the amount of data is significant for the results. Another factor that influences the experiments we have conducted is the experimental setup. Hence, we also would like to conduct more experiments using different setups in the future.

ACKNOWLEDGMENT

This research was financially supported by the Swedish Armed Forces Research and Development Programme.

REFERENCES

[1] A. Abbasi and H. Chen, "Affect intensity analysis of dark web forums," in *Proceedings of the 5th IEEE International Conference on Intelligence and Security Informatics*, 2007.

[2] J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Mårtenson, and P. Svenson, "Analysis of weak signals for detecting lone wolf terrorists," in *Proceedings of the 2012 European Intelligence and Security Informatics Conference*, 2012, pp. 197–204.

[3] D. Goldschlag, M. Reed, and P. Syverson, "Onion routing," *Communications of the ACM*, vol. 42, no. 2, pp. 39–41, 1999.

[4] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, E. Shin, and D. Song, "On the feasibility of internet-scale author identification," in *2012 IEEE Symposium on Security and Privacy (SP)*, may 2012, pp. 300 –314.

[5] F. Johansson, L. Kaati, and A. Shrestha, "Detecting multiple aliases in social media," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. ACM, 2013, pp. 1004–1011.

[6] A. Adan, S. N. Archer, M. P. Hidalgo, L. Di Milia, V. Natale, and C. Randler, "Circadian typology: a comprehensive review," *Chronobiology international*, vol. 29, no. 9, pp. 1153–1175, 2012.

[7] R. Urban, T. Magyarodi, and A. Riga, "Morningness-eveningness, chronotypes and health-impairing behaviors in adolescents." *Chronobiology International*, vol. 28, pp. 238–247, 2011.

[8] A. Adan, "Chronotype and personality factors in the daily consumption of alcohol and psychostimulants." *Addiction*, vol. 89, pp. 455–462, 1994.

[9] B. Lee, "A temporal analysis of posting behavior in social media streams," *International AAAI Conference on Weblogs and Social Media*, 2012.

[10] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[11] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Trans. Inf. Syst.*, vol. 26, no. 2, pp. 7:1–7:29, Apr. 2008.

[12] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 378–393, 2006.

[13] P. Juola, "Authorship attribution," *Found. Trends Inf. Retr.*, vol. 1, no. 3, pp. 233–334, 2006.

[14] J. Dahlin, F. Johansson, L. Kaati, C. Mårtenson, and P. Svenson, "Combining entity matching techniques for detecting extremist behavior on discussion boards," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, 2012, pp. 850–857.

[15] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Formal Methods for Information Retrieval*, vol. 42, pp. 155–165, 2006.