Temporal Word Embeddings for Dynamic User Profiling in Twitter

Breandán Kerin[1], Annalina Caputo[2], and Séamus Lawless[1]

[1] ADAPT CENTRE, School of Computer Science & Statistics
Trinity College Dublin, Ireland
kerinb@tcd.ie,seamus.lawless@scss.tcd.ie
[2] ADAPT CENTRE, School of Computing
Dublin City University, Ireland
annalina.caputo@dcu.ie

Abstract. The research described in this paper focused on exploring the domain of user profiling, a nascent and contentious technology which has been steadily attracting increased interest from the research community as its potential for providing personalised digital services is realised. An extensive review of related literature revealed that limited research has been conducted into how temporal aspects of users can be captured using user profiling techniques. This, coupled with the notable lack of research into the use of word embedding techniques to capture temporal variances in language, revealed an opportunity to extend the Random Indexing word embedding technique such that the interests of users could be modelled based on their use of language. To achieve this, this work concerned itself with extending an existing implementation of Temporal Random Indexing to model Twitter users across multiple granularities of time based on their use of language. The product of this is a novel technique for temporal user profiling, where a set of vectors is used to describe the evolution of a Twitter user's interests over time through their use of language. The vectors produced were evaluated against a temporal implementation of another state-of-the-art word embedding technique, the Word2Vec Dynamic Independent Skip-gram model, where it was found that Temporal Random Indexing outperformed Word2Vec in the generation of temporal user profiles.

Keywords: User Modelling · Word Embeddings · Random Indexing.

1 Introduction

As of the time of writing, it is estimated that approximately 4.36 billion people of an estimated 7.6 billion globally are connected to the internet[3]. Some of the most successful of the 1.9 billion live websites in 2019 are social networking sites[4], hosting Online Social Networking (OSN) platforms such as Twitter, Facebook

[3] http://worldpopulationreview.com/
[4] http://www.internetlivestats.com/internet-users/

and YouTube which allow users to connect digitally and share online content with each other. In order to capture, maintain and continuously increase the engagement of such a large user base in an incredibly complex global environment, the organisations behind these platforms are increasingly employing user profiling tactics, where the preferences and interests of the user are modelled, clustered and learned in order to deliver tailored content directly to them in a scalable manner.

As described by Kanoje et al., [6] user profiling is "the process of identifying the data about a user interest domain... [which] can be used... to understand more about [the] user and this knowledge can be further used for enhancing the retrieval for providing satisfaction to the user." It is a contentious technology from an ethical perspective: Whether it is always used in accordance with legal and ethical guidelines is highly debatable. Several multinational technology companies have come under fire for leveraging and capitalising upon their users' data without obtaining their explicit consent and knowledge, with allegations as serious as implicitly influencing the US population to elect President Donald J. Trump to the White House in 2016[5]. Regardless of this, user profiling has high potential and desirability from the perspective of improving user experience, simplifying navigation of the internet through personalisation and allowing relevant content to be delivered to users more efficiently.

The idea of temporally modelling or profiling users can be motivated by the observation that an individual user and their data are not static: Their interests and preferences evolve and vary through time, often following patterns such as trends, periodicities and spikes. This was demonstrated by Bonneville-Roussy et al. [2] who found that an individual's musical interests vary through time and tend to fluctuate and change around "particular life changes". Natural Language Processing (NLP) techniques such as word embeddings[6] are already widely applied in the analysis of users based on textual information. Since the introduction of the Latent Semantic Analysis (LSA) algorithm [3] in 1990, a wealth of word embedding techniques has been developed including Random Indexing (2005) [13], Word2Vec (2013) [11], GloVe (2014) [12], and FastText (2016) [4], all of which remain in widespread use in NLP applications.

It is clear that strong user profiling techniques should be capable of capturing temporal variances in user characteristics. The idea of capturing temporal variances whilst modelling users and their interests through time has not been the subject of a great deal of research at the time of writing this document, despite the volume of research that exists in both user profiling and NLP. Thus, exploring new viable approaches to capturing temporal variations in user profiles is the primary motivation for this research.

---

[5] `https://www.nytimes.com/2018/04/10/us/politics/mark-zuckerberg-testimony.html`

[6] Word embeddings are a means of representing the semantic properties of a vocabulary of a corpus in a vector space, and are used widely in the area of NLP.

2 Related Work

2.1 User Profiling on OSNs

User profiling using OSN data has been widely explored by the research community, focusing on problems including personality inference and expertise inference. Wald et al. [14] proposed a personality inference model for Facebook users which built user profiles based on the demographic and text-based information present on each user's Facebook profile. Similarly, Matz et al. [10] proposed an approach to analysing the personality traits of Facebook users using their determined OCEAN[7] personality traits as a basis to enable mass persuasion for marketing more effectively to these users. In the domain of expertise inference, Xu et. al. [15] developed a novel topic modelling framework to determine the expertise of Twitter users by employing an extension of the Latent Dirichlet Allocation (LDA) algorithm to produce an augmented topic model. In their research, they observed the importance of the temporal aspects of user profiles in their future works.

When it comes to temporal user profiling research, the quantity of research conducted to-date is more limited. In their 2012 paper, Zhang et. al. [17] proposed a user profiling system which modelled users of a mobile network both statically and dynamically using various different modelling techniques and compared the performance using clustering algorithms. In 2017, Liang et. al. [8] proposed a dynamic user clustering topic model which generated a vector-based model of Twitter users and clustered the results based on their cosine similarity. In both of these works, the temporal modelling approach was found to outperform its static counterparts.

2.2 Temporal Word Embeddings

It is clear that understanding the temporal aspects of words and their semantics is of major interest to fully understand the users of OSNs. Word embedding techniques, as a group of widely used NLP techniques, are a strong candidate to solve this problem. Temporality with respect to word embeddings considers the way in which word semantics vary over time. There has been an increase in interest in this variety of word embeddings, stemming from the fact that there is now an abundance of time variant data sets available from major websites and application platforms, as well as an increased appreciation of the fact that word semantics do not remain static over time. Several methods have been proposed in research which temporally extend static word embedding methods.

– Jurgens and Keith proposed Temporal Random Indexing (TRI) [5], a temporal extension of the Random Indexing word embedding technique. This

[7] OCEAN is a set of personality traits used by psychologists to measure and characterise personality traits. The abbreviation stands for Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism.

algorithm generates word embeddings as a function of time, enabling analysis and investigation into the evolution of word meanings over time. This technique was used in their research for event detection in blog posts. Subsequently, Basile et. al. [1] successfully applied TRI to event detection in news articles.

- Yao et. al. [16] proposed a "dynamic statistical model to learn time-aware word vector representation[s]", building upon the the static Word2Vec model to "learn time-aware word vector representation[s]" for a New York Times dataset. Liang et al. [9] also proposed a temporal extension of the Word2Vec technique, using it as the basis for a temporal user profiling system which modelled model Twitter users' interests through time.

As in the related literature regarding user profiling, it was concluded by these researchers that temporal models tended to outperform their static counterparts.

3 Methods

It is clear from the research literature that there remain many opportunities to further investigate potential temporal user profiling techniques, and that temporal word embedding techniques show a significant promise regarding understanding the temporal aspects of users interests based on their language usage. These observations motivated the decision to explore the application of Temporal Random Indexing to the problem of scalable temporal user profiling using OSN data.

3.1 The Temporal Random Indexing Technique

TRI is a word embedding technique, proposed by Jurgens and Keith [5] as a temporal extension of the Random Indexing method proposed by Salhgren [13]. In contrast to other popular embedding methods, RI-based techniques employ an implicit dimensionality reduction process which preserves the semantic information encoded within a term-term co-occurrence matrix. Rather than performing an explicit reduction upon a co-occurrence matrix as with other techniques[8], *Context Vectors* are generated incrementally by accumulating *Index Vectors* from defined *context windows*. Thus, a significant advantage of RI-based techniques is their ability to generate *Context Vectors* in an online fashion, where the model can be continuously updated as new information becomes available without requiring the model to be taken offline[9]. In contrast, other state-of-the-art methods such as Word2Vec and GloVe require training in an offline fashion: As new data is acquired, the model requires offline re-training and re-deployment.

[8] Examples of word embedding techniques which use explicit dimensionality reduction include GloVe and LSA. [3] [12]

[9] Since the Random *Index Vectors* remain constant, as new data is acquired the model can simply add the linear combination of new *Index Vectors* to the generated *Context Vectors*.

For a given corpus C consisting of n documents where document $d \in C$, a vocabulary V of m words can be extracted from C. Given this, the two steps involved in RI are as follows:

1. Assign a randomly generated *Index Vector* r_p to each word w_i in the vocabulary: $w_i \in V$.
2. Generate a semantic vector representation sv_i for each word w_i, defined as the sum of all random *Index Vectors* r_p assigned to the words that co-occur with w_i in a given *context window* given by the range $-j < p < +j$ for constant j. This is described by the following equation:

$$sv_i = \sum_{d \in C} \sum_{-j<p<+j} r_p$$

In the case of Temporal Random Indexing, before step 1 can be performed the corpus C must first be annotated to contain metadata related to the creation time and date of the data. From this time data, the corpus C can then be split into k subsets C_1, C_2, ... C_k, where k is the number of time periods to analyse. To ensure that the *Context Vectors* produced by TRI are comparable across multiple time periods, the *Index Vectors* remain constant during the entire embedding generation process across each time period T_k, and the previous time period's *Context Vectors* are re-used as the initial state of the new time period's *Context Vectors*.

The major divergence between TRI and RI occurs in step 2 of the process. In RI, all data within the corpus is used to generate the vectors as time is considered as static and doesn't vary. In contrast, time is considered dynamic by TRI and thus a separate vector space is generated for each time period T_k contained within a document d. The equation governing this second process in TRI is given as follows:

$$sv_{i,T_k} = \sum_{d \in C} \sum_{-j<p<+j} r_p$$

Using this approach, it is possible to build vector spaces for each time period T_k over a given corpus C_k annotated with creation time metadata. Each word w_i contained within the corpus has a unique *Context Vector* representation for each time period T_k considered, which are all built upon the same random *Index Vectors*. This allows for direct comparison between words within different time periods, since they are generated from a linear combination of the same random *Index Vectors*.

3.2 Augmented Temporal Random Indexing for User Profiling

In this research, the previously described TRI technique is extended to capture a vectorised representation of Twitter users based on their language usage, using the same *Index Vectors* to generate a *Context Vector* i.e. a vector representing the user. Doing this allows for a direct comparison of not only words in the same

vector space across multiple time periods, but also allows for the comparison of a single user with the shared vocabulary of all users in the dataset using vector mathematics.

This information about the user can only be captured by utilising the metadata related to the Tweets contained in the corpus: Specifically, the creation time as well as the *user id* present in Tweet metadata. The additional step required to facilitate this involves augmenting the TRI technique to generate these $user_i$ vectors, achieved by collating the text written by a user in a given time period T_k and leveraging the same random *Index Vectors* used in generating the word embeddings.

Let T_k be a time period which spans between $t_{k_{start}}$ to $t_{k_{end}}$, where $t_{k_{start}}$ is a time that predates $t_{k_{end}}$. In order to build the user vector uv_i for time period T_k, we consider all p words contained within the Tweets authored by user i within the time period T_k, where all *Index Vectors*, r_p contained within the series of Tweets are summed as follows:

$$uv_{i,T_k} = \sum_{p \in T_k} r_p$$

Applying this additional step to the TRI technique, it is possible to construct a vector space for each time period T_k from a corpus C of n documents containing metadata related to Tweet creation time and *user id*. Each user u_i has a distinct vector representation for a given time period in this vector space, given by uv_{iT_k}, which is generated by accumulating the random *Index Vectors* for each time a word p was used by a $user_i$ in the given time period T_k.

4 Evaluation

The evaluation of this research focused on measuring how performant TRI is at modelling a user and their interests temporally based on their use of language. A temporal extension of the Word2Vec Skip-Gram model[10] is used as a baseline comparison to the TRI user profiling method described in this paper. The DISG model utilises the time period separated data, and independently generates embeddings for the words used in each time period. To generate a single user vector using DISG, the word vectors for each word used by a given user for a given time period are summed and averaged, resulting in a user vector being produced as the centroid of the word vectors used by that user in a given time period.

For the purposes of this research, the proposed system was evaluated by following the same approach and collection of Tweets used by Liang et al. [7] The collection consists of $1,375$ randomly selected Twitter users, along with all the tweets that they have authored since the beginning of their registration and up to and including May 31, 2015. The input dataset is built by splitting all the Tweets into several different time periods: month, quarter, and semester. An

[10] Specifically, the Word2Vec Skip-Gram model used is the Dynamic Independent Skip-Gram (DISG) model, which is an inherently temporal model.

automatically generated ground truth is created by extracting all the hashtag words associated with a specific period of time. Specifically, given the collection of Tweets belonging to each one of the time periods, we extracted all of the hashtags belonging to the Tweets in each time period, removing the hash character ('#') and lowercasing the hashtag word. Finally, we ranked all of the hashtags by their $tf - idf$, where the idf was computed on all of the Tweets authored by a given user, and selected only the top 10 hashtags as ground truth. This is an important difference from the method described in [7], since in the original approach the hashtags were ranked based on their count in the tweet. We observed that many time periods contained only one occurrence of each hashtag, leading to multiple tie situations. Hence, it was decided that ranking the hashtags based on how specific they are for a given period was the most effective approach. The evaluation uses hashtags as a proxy for the most important concepts shared by a user in a specific period time and it aims at assess the capability of the system at retrieving those relevant hashtags. Since the problem is formulated as a typical retrieval task, we adopted Mean Average Precision (MAP) and Precision@k as evaluation metrics.

5 Results

By applying the techniques described in the Methods and Evaluation sections to the prepared Twitter dataset, and evaluating them against the generated ground truths using the *trec_eval* evaluation tool, the results shown in Table 1 were observed.

Table 1. Performance of TRI and baseline DISG systems with different time period splits: Month, Quarter, Semester.

	MAP		P@5		P@10	
	TRI	DISG	TRI	DISG	TRI	DISG
Month	0.0126	0.0005	0.0219	0.0007	0.0157	0.0004
Quarter	0.0142	0.0007	0.0311	0.001	0.0224	0.0007
Semester	0.0183	0.0001	0.0414	0.0002	0.0293	0.0002

As can be observed from Table 1, although poor MAP and precision values were obtained for both models, TRI was found to outperform the Word2Vec DISG model for each of the time periods considered. This is a promising result for the usability of TRI and further work into improving the data processing, implementation and evaluation methods is likely to reveal further improvements in results.

6 Conclusion

This research has demonstrated the potential of the TRI technique as an effective approach to temporal user profiling through its application in analysis of temporal variations in language use.

The results obtained clearly demonstrate that the TRI technique outperforms Word2Vec, a word embedding technique currently considered to be state-of-the-art in the NLP domain. This is a significant finding, and highlights the huge potential of applying word embedding techniques in temporal user profiling applications. As a domain which has seen little research to-date, and given the ever-increasing volume of text content generated by users of some of the world's biggest OSN platforms, the growing importance of strong NLP techniques such as TRI cannot be overstated.

It is of the opinion of the authors that with further refinement and experimentation, a temporal user profiling system employing TRI could be realised which is performant and scalable enough to be considered for use in online, web-scale environments.

Acknowledgement. This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106 at the ADAPT SFI Research Centre at Trinity College Dublin and Dublin City University and by the European Union's Horizon 2020 (EU2020) research and innovation programme under the Marie Skodowska-Curie grant agreement No.: EU2020 713567. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant # 13/RC/2106.

References

1. Basile, P., Caputo, A., Semeraro, G.: Temporal random indexing: a tool for analysing word meaning variations in news. In: Martinez-Alvarez, M., Kruschwitz, U., Kazai, G., Hopfgartner, F., Corney, D., Campos, R., Albakour, D. (eds.) Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016. CEUR Workshop Proceedings, vol. 1568, pp. 39–41. CEUR-WS.org (2016), `http://ceur-ws.org/Vol-1568/paper7.pdf`
2. Bonneville-Roussy, A., Rentfrow, P.J., Xu, M.K., Potter, J.: Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood. (2013). https://doi.org/10.1037/a0033770
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science **41**(6), 391–407 (1990). https://doi.org/10.1002/(SICI)1097-4571(199009)41:6¡391::AID-ASI1¿3.0.CO;2-9
4. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter

of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics, Valencia, Spain (Apr 2017), https://www.aclweb.org/anthology/E17-2068
5. Jurgens, D., Stevens, K.: Event detection in blogs using temporal random indexing. In: Proceedings of the Workshop on Events in Emerging Text Types. pp. 9–16. Association for Computational Linguistics, Borovets, Bulgaria (Sep 2009), https://www.aclweb.org/anthology/W09-4302
6. Kanoje, S., Girase, S., Mukhopadhyay, D.: User profiling trends, techniques and applications (2015)
7. Liang, S.: Dynamic user profiling for streams of short texts. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 5860–5867. AAAI Press (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16646
8. Liang, S., Ren, Z., Zhao, Y., Ma, J., Yilmaz, E., Rijke, M.D.: Inferring dynamic user interests in streams of short texts for user clustering. ACM Trans. Inf. Syst. **36**(1), 10:1–10:37 (Jul 2017). https://doi.org/10.1145/3072606, http://doi.acm.org/10.1145/3072606
9. Liang, S., Zhang, X., Ren, Z., Kanoulas, E.: Dynamic embeddings for user profiling in twitter. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1764–1773. KDD '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3219819.3220043, http://doi.acm.org/10.1145/3219819.3220043
10. Matz, S.C., Kosinski, M., Nave, G., Stillwell, D.J.: Psychological targeting as an effective approach to digital mass persuasion. Proceedings of the National Academy of Sciences **114**(48), 12714–12719 (2017). https://doi.org/10.1073/pnas.1710966114, https://www.pnas.org/content/114/48/12714
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
12. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
13. Sahlgren, M.: An introduction to random indexing. In: Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering (2005)
14. Wald, R., Khoshgoftaar, T., Sumner, C.: Machine prediction of personality from facebook profiles. In: 2012 IEEE 13th International Conference on Information Reuse Integration (IRI). pp. 109–115 (Aug 2012). https://doi.org/10.1109/IRI.2012.6302998
15. Xu, Z., Ru, L., Xiang, L., Yang, Q.: Discovering user interest on twitter with a modified author-topic model. In: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. vol. 1, pp. 422–429 (Aug 2011). https://doi.org/10.1109/WI-IAT.2011.47
16. Yao, Z., Sun, Y., Ding, W., Rao, N., Xiong, H.: Dynamic word embeddings for evolving semantic discovery. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. pp. 673–681. WSDM '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3159652.3159703, http://doi.acm.org/10.1145/3159652.3159703

17. Zhang, C., Masseglia, F., Zhang, X.: Modeling and clustering users with evolving profiles in usage streams. In: 2012 19th International Symposium on Temporal Representation and Reasoning. pp. 133–140 (Sep 2012). https://doi.org/10.1109/TIME.2012.16