2022-10-12

XAI Analysis of Online Activism to Capture Integration in Irish Society Through Twitter

Arjumand Younus
Technological University Dublin, arjumand.younus@tudublin.ie

Muhammad Atif Qureshi
atif.qureshi@tudublin.ie

Mingyeong Jeon
Technological University Dublin, mingyeong.jeon@tudublin.ie

See next page for additional authors

Recommended Citation

Younus, A., Qureshi, M.A., Jeon, M., Kazemi, A., Caton, S. (2022). XAI Analysis of Online Activism to Capture Integration in Irish Society Through Twitter. In: Hopfgartner, F., Jaidka, K., Mayr, P., Jose, J., Breitsohl, J. (eds) Social Informatics. SocInfo 2022. Lecture Notes in Computer Science, vol 13618. Springer, Cham. DOI: 10.1007/978-3-031-19097-1_14

Funder: Irish Research Council

Authors

Arjumand Younus, Muhammad Atif Qureshi, Mingyeong Jeon, Arefeh Kazemi, and Simon Caton

XAI Analysis of Online Activism to Capture Integration in Irish Society Through Twitter

Arjumand Younus[1(B)], M. Atif Qureshi[1,2], Mingyeong Jeon[1,2], Arefeh Kazemi[3], and Simon Caton[3]

[1] eXplainable Analytics Group, Faculty of Business, Technological University Dublin, Dublin, Ireland
{arjumand.younus,muhammadatif.qureshi,mingyeong.jeon}@tudublin.ie
[2] ADAPT Centre, Dublin, Ireland
[3] School of Computer Science, University College Dublin, Dublin, Ireland
{arefeh.kazemi,simon.caton}@ucd.ie

Abstract. Online activism over Twitter has assumed a multidimensional nature, especially in societies with abundant multicultural identities. In this paper, we pursue a case study of Ireland's Twitter landscape and specifically migrant and native activists on this platform. We aim to capture the level to which immigrants are integrated into Irish society and study the similarities and differences between their characteristic patterns by delving into the features that play a significant role in classifying a Twitterer as a migrant or a native. A study such as ours can provide a window into the level of integration and harmony in society.

Keywords: Integration · Explainable artificial intelligence · Society · Twitter metadata · Textual

1 Introduction

Twitter has assumed a multi-faceted role as a communications platform while also becoming a platform for knowledge sharing, activism, journalism etc. Many researchers have used Twitter data to analyse social phenomena, opening research pathways toward data science for social good initiatives. One promi- nent area is that of migration studies, with many works relying on geo-tagged location data within tweets to make inferences about migration patterns [19,22], and mobility flows. However, we argue for a more nuanced approach to studying how immigrants have integrated into their migrated society. In this paper, we take up Ireland as a case study and utilise Twitter data from within Ireland to perform an analysis of migrants' integration into Irish society.

This publication has emanated from research conducted with the support of Irish Research Council under award no. GOIPG/2021/1354 & COALESCE/2021/112 and with the financial support of Science Foundation Ireland under grant no. 13/RC/2106 P2 at the ADAPT SFI Research Centre.

Ireland has recently grown into a multicultural society [13] with an increase in migration of foreign working professionals, international students and asy- lum seekers and refugees. Due to its welcoming policies confirming diversity and inclusion, Ireland has embraced these communities as their own [17]. However, the transformation has caused a plethora of online activism from migrant communities of Ireland and the native population of Ireland. Much of this online activism is displayed on popular social networks, such as Twitter [24]. On one end, there are activists from migrant communities, and on the other, there are those from the natives, both advocating for various social causes. There are, however, similarities and dissimilarities in the causes these activists support concerning society and the overall transformation. There is a lack of objective reasoning into the different communication patterns and causes advocated by local activists and migrant ones, and this work aims to present a computational methodology to fill that gap.

In an attempt to study the various aspects of migrants' integration in Irish society, we analyse different aspects of Twitter activity (of natives and migrants) by means of Twitter metadata and the textual content of tweets. Learning user representations via the textual content they generate has gained tremendous attention due to its usefulness in solving various user classification tasks such as hate speech detection, mental illness prediction, and fake news detection [16]. We embrace a different approach, mainly relying on unsupervised learning methods for extraction of different aspects of Twitter activity, followed by usage of these features in explainable artificial intelligence based classification models [3]. We fundamentally frame the task around the prediction of whether or not a given Twitter activist is a migrant or native, along with the most informative and distinguishing features. This classification model and associated feature analysis, in turn, helps us identify the features that make sense to determine migrants' integration into society.

Our main contributions in this work are summarized as follows

– A best-of-both worlds approach is adopted by making use of rich Twitter communication patterns (such as mentions, retweets, and quotes) together with textual contents' overlap within a tabular data representation for migrant and native communities of Ireland.
– Use of two well-known explainable artificial intelligence approaches to provide a glimpse into the most discriminating features of migrants and natives, thereby providing a basic mechanism from where one can begin to assess the level of migrants' integration in society.
– A curated, custom dataset has been made publicly available to the research community to enable further development of migration studies using textual data.

The remainder of this paper is organised as follows. In Sect. 2, we discuss related work while attempting to cover migration studies along with works on user representations using social media data. Furthermore, we also present a brief explanation of the research gap and position our work concerning how it attempts to fill this gap. In Sect. 3, we present details of our collected data and methodology. In Sect. 4, we present our findings along with a discussion aroundthe implications of these findings. In Sect. 5, we conclude with a brief overview of future work.

2 Related Work

In any society where multicultural identities are present, there remains a boundary between natives and migrants in various aspects. More specifically, over the years, in cases of increased migration to European Union countries, there remains a certain level of scepticism on policies proposed for integrating migrants into society [15]. The general consensus, however, is that the policies have been ineffective in achieving the said goals. The general problem with this consensus is that there has been limited quantitative analysis of migrants' lived experiences. We aim to fill this gap using social media data for this analysis.

In studies that leverage social media, we can easily distinguish studies that investigate more large-scale macro-level events, such as political campaigning (e.g. [10]), riots and civil unrest (e.g. [7]), event detection (e.g. [1,18]), and large-scale news events like COVID-19 (e.g. [23]). Here, researchers often use social media as a lens to understand the perceptions and viewpoints of a subset of society or seek to identify key events that occur during the period(s) of observation. In these cases, it is generally easy to think about how to access and curate a meaningful and large corpus of text content. This is, however, not quite the case for studying immigrants: they are, as a hard-to-reach and arguably marginalised group, hard(er) to locate and study with social media. In fact, when it comes to marginalised groups, there is a general observation that researchers lack robust methodological guidelines [2].

There has been a lot of work (e.g. [6,8,28,31]) studying hate, toxicity, and cyberbullying, which is often directed towards immigrants and marginalised groups via social media. Though useful from a methods perspective, these studies often have a strong content bias towards content stemming from North America and the United Kingdom [21]. [9] outlines how social media can be used to specifically study marginalised groups and discusses a number of the associated challenges. The key here is to ensure that corpus is representative and not just an echo-chamber of a few (obvious) selected topics.

The literature has a number of consistent strategies for curating data sam- ples for studying marginalised communities. One common method is to select keyusers and extract their tweet history (e.g. [30]), using a set of seed users and sam-ple from "key" related users (e.g. [11,27,30]), focusing around specific hashtags (e.g. [2,27]), based on specific locations (e.g. [26])[1], and using online reposito- ries of Twitter archives for key public figures (e.g. [11]). Other exploratory data analysis techniques can also be used for vernacular discovery to identify spe-cific keywords for content search. This, as noted by [30], is useful in identifying specific sociolinguistic variations of language, slang, shorthand etc. However, as noted by [4,30], researcher training in the area of study (to generate domain

[1] It should be noted that geo-location data can be unreliable.

knowledge) is critical. An alternative method to determine a (starting) vocabulary set, as introduced by [5], is to collect anonymised data from relevant Web forums, blogs and microblogs, and ask human annotators to identify whether the content contains specific references of use.

Our work adopts the approach of selecting a manually curated set of users as proposed in [30] with the "activism" of a user being the key focus. To the best of our knowledge, our work provides a novel direction in studying a diverse range of topics emanating from marginalised communities versus privileged ones, with previous such studies limiting their study around an event, hashtag or group of similar users [11,26,27].

3 Dataset and Methodology

Our model of choice is TabNet on account of its support for multiple modalities together with the features it offers for explaining its outcomes. Explainability offers deeper insights for a task such as the one addressed in this paper of determining the levels to which migrants feel included in or excluded from society.

In this section, we first delve into the details of our dataset, followed by an explanation of the features used to make the predictions of whether a particular Twitterer from within our dataset is a migrant or a native.

Table 1. Characteristics of Curated Twitter Activists' Dataset from Both Communi- ties

Type	Mean Tweets	Mean followers	Mean following
Migrants	2506	4184	1793
Natives	3053	12116	3289

3.1 Dataset: Activists from Both Communities

For the dataset creation, we manually curate a list of Twitter activists fromwithin migrant and native communities of Ireland, with the criteria that theTwitterer must have had tweeted/retweeted four or more tweets on a social justice issue within Ireland. The curation is performed by one of the authors, familiar with Ireland's social justice landscape, and can distinguish between a native/local and a migrant. The data curation involves the following steps as an inclusion/differentiation criteria of natives vs migrants for our Irish Twitter dataset:

- Irish surname check done by means of https://www.duchas.ie/en/nom
- Reading Twitter biography field and checking for Irish terms in addition to flags of various countries[2]

[2] It has been observed that migrants within Ireland usually insert a flag representing Ireland and their country of origin.

– Reading last 20–100 tweets of an activist to determine whether or not there's any explicit mention of belonging from any country.

The above process yielded a total of 66 natives and 66 migrant activists. Despite the sample not being large, it serves as a fair representation of the Irish Twitter landscape, and Table 1 shows the characteristics of the selected activists. For each user in the curated list, we extract tweets and Twitter metadata associated with these tweets. Using Twitter API's academic research product track, the last 3200 tweets of the Twitterers in our list were crawled[3].

3.2 Twitter Metadata Features to Study Integration

As has been observed in research on Twitter communication patterns [12,33], the various forms of communication on Twitter, such as mentions, retweets, and the newly introduced feature quoted tweets, give an insight into significant aspects of a user. Regarding migrants and natives, these Twitter features specifically can help determine social bonding factors [29] between these two, which is why the following features are used in our prediction framework.

– Follower Following Ratio: The ratio between the Twitter followers and those a certain user follows serves as an important indication of online popularity. To illustrate the results in the next section, we use the notation *followerfollowingratio* for this particular feature.
– Mentions Percentage: The percentage of mentions within the last 3200 tweets of a user; this can be extracted from Twitter metadata's *replied to* _feature. To illustrate the results in the next section, we use the notation *mentionspercentage* for this particular feature.
– Retweets Percentage: The percentage of retweets within the last 3200 tweets of a user; this can be extracted from Twitter metadata's *retweeted* feature. To illustrate the results in the next section, we use the notation *retweetedpercentage* for this particular feature.
– Quoted Percentage: The percentage of quoted tweets within the last 3200 tweets of a user; this can be extracted from Twitter metadata's *quoted* fea- ture. To illustrate the results in the next section, we use the notation *quotedpercentage* for this particular feature.

3.3 Textual Features to Study Integration

Two types of features were used as the textual features, i.e., based on word embeddings and topic models. Both are textual modelling techniques well-known in the literature for making inferences concerning various classes of users (we covered some of these in Sect. 2). These techniques are able to coherently capture similar and dissimilar topics from a huge body of text which in our case is all the

[3] Note that an anonymised version of this dataset has been released for public download at https://github.com/arjumandyounus/ineire_tweetsdataset.

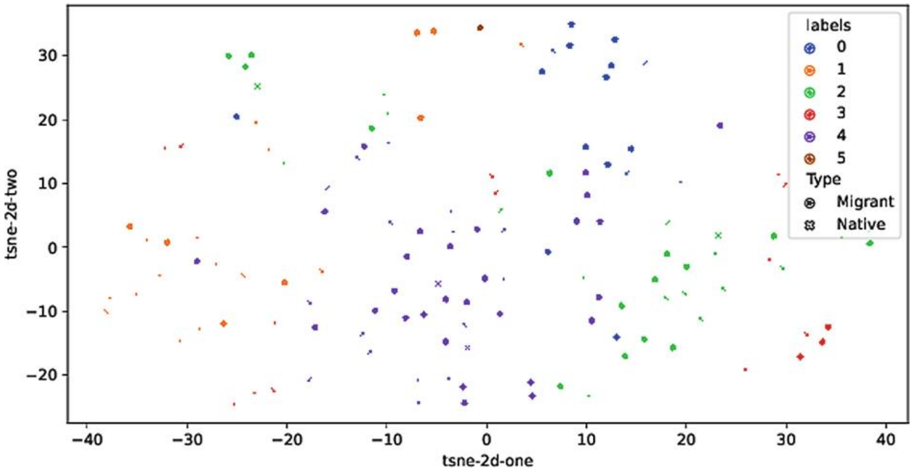

Fig. 1. TSNE visualisation of clusters obtained from KMeans over user Tweet embeddings.

tweets by migrants and activists. A particularly challenging aspect, however, in the kind of task we have proposed is the different means by which accumulation of topics for a particular user can be done on account of a user being verydiverse in tweeting habits and topical interests. We solve this challenge by takingTwitterer's most frequent words in word embeddings and the most frequent topics in the case of topic models.

For the word embedding based ones, we combine all tweet tokens of a user into a single document, thereby treating the user as a document. This is followed by a reduction process where each user document is then reduced to top-20 most frequent tokens. These tokens corresponding to each user are trained with a word2vec model [25] to obtain word embeddings and finally, KMeans clustering [14] is applied over all users[4]. The label of the cluster to which each user belongs is used as a feature in the model. Figure 1 shows the results of the KMeans clus- tering via tsne dimensionality reduction over the word embeddings of migrants and activists in our dataset. To illustrate the results in the next section, we use the notation *labels* for this feature.

For the topic modelling features, we first apply the Gibbs Sampling Dirichlet Multinomial Mixture model (i.e., GSDMM) [32] for short text clustering to all the tweets of each user. Our choice of this alternative compared to the traditional Latent Dirichlet Allocation model is motivated by its strength in modelling via collapsed Gibbs sampling of one topic per document, making it more suited to short texts such as tweets. The application of GSDMM yields a collection of topics equivalent to the number of tweets by each user, and from this collection, we derive the top-4 topics for each user. We also derive the top-4 topics for the migrants and natives in our sample. For each user, we then compute the Jaccard

4 An empirical analysis led to a stable number of 6 clusters.

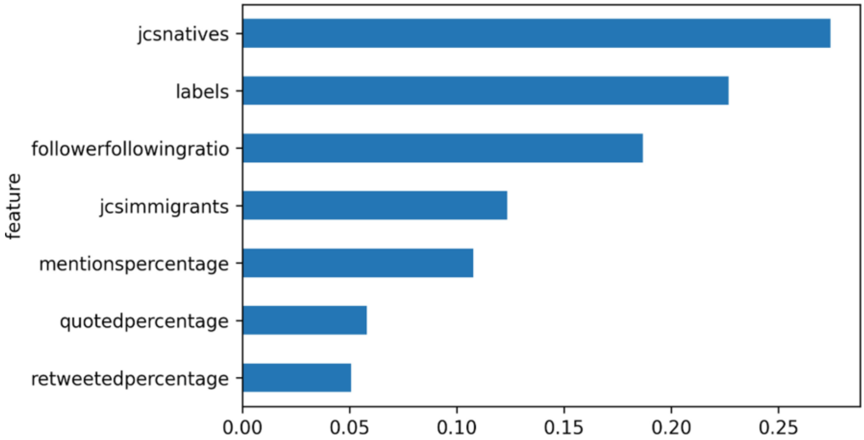

Fig. 2. Features' importance for TabNet

similarity score across the top-4 topics for migrants and natives; this score is then converted into a categorical feature whereby a score greater than 0.5 across migrants' top-4 topics is assigned a category *"SimilarToMigrants"* and a score less than or equal to 0.5 is assigned a category *"DisSimilarToMigrants"*. Note that a similar categorical assignment is performed with topics within tweets by natives. To illustrate the results in the next section, we use the notation *jcsimmigrants* and *jcsnatives* for these topic model based features.

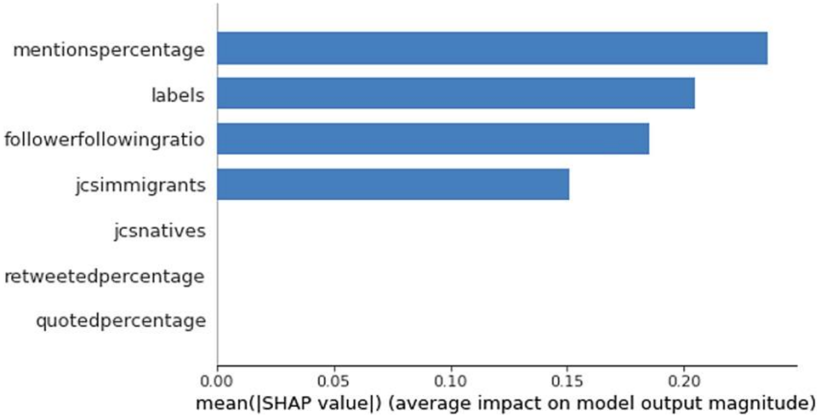

Fig. 3. Features' importance for XGBoost

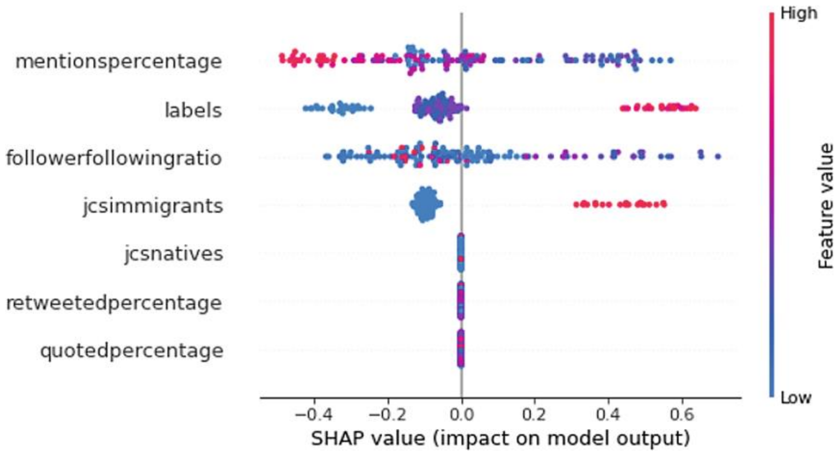

Fig. 4. SHAP Summary Plot for Tree Explainer of XGBoost

Table 2. Classification results across prediction task of Twitter activist being migrant or native

Classification approach	Accuracy	Precision
TabNet	0.85	0.78
XGBoost	0.87	0.80

3.4 Classification Frameworks and Explainable AI

The Twitter metadata features explained in Sect. 3.2 and textual content features explained in Sect. 3.3 are combined to produce a tabular representation for each user in our dataset. Two state-of-the-art tabular machine learning methods, i.e., XGBOOST with SHAP [20] and TabNet [3] built on top of the explainableAI paradigm, are then used to discover features that play a substantial role in distinguishing between migrants and natives; and this, in turn, helps characterise the level to which migrants have integrated into society. We discuss these aspects in significant detail in the next section.

4 Findings and Discussion

Table 2 presents the results for classification accuracy for both approaches. Both the algorithms exhibit similar results in terms of accuracy and precision while showing variations in feature importance (refer to Fig. 2 and Fig. 3). With respect to feature importance and associated explanations however an interesting pattern is observed from the features *labels* and *followerfollowingratio* whereby both play a significant role within both classification methods. A deeper analysis via SHAP values however reveals the stability of *labels* feature over other Twit- ter metadata features. SHapley Additive exPlanations (SHAP values) quantify

Table 3. Description of topics in top six clusters

Cluster Number (No. of Tweets Inside)	Top four words
Number 5 (80201)	Covid cases hospital people
Number 7 (45319)	People today Ireland Government
Number 8 (44564)	Work stamp people visa
Number 9 (40977)	Dublin great day weather
Number 1 (40192)	Black garda race need
Number 2 (37319)	Women trans Ireland rights

the contribution of each feature corresponding to the predictions made by a given machine learning model, and thereby gives deeper insights into local interpretability of a predictive model. The insights derived from the SHAP summary plot of Fig. 4 point to the observation that the *labels* feature derived from word embeddings of textual content within migrants' and natives' tweets leads to stable predictions, and a coherent level of linearity. This points to a significantaspect with respect to migrants' integration into society where the topics and interests of both communities are defining factors. As an example we revisitthe tsne visualisation of Fig. 1 whereby labels 0 and 2 in particular show a dis-tinct pattern of differences in content posted by migrants and natives on Twitter; *label 0* is derived from words textit"stereotype, racism, black, twitter" while *label 4* is derived from words textit"savenmh, abortion, women, repealed" whereby migrants have a large proportion of belonging to *label 0* while natives have a large proportion of belonging to *label 4*. From this pattern, there is a clear demar-cation of interests and concerns of both communities depicting a probable lackof migrant integration into Irish society specially given the fact that *"savenmh"* made headlines for many weeks.

To dig deeper into the differences in topical interests of migrants and natives, we show the top four words within the top six topics obtained via GSDMM in Table 3. The topics are in descending order by the number of documents/tweets in each topic. As with the labels obtained from word embeddings, there is an apparent topical difference, particularly in Clusters 2, 8, and 9 whereby most tweets in Cluster 8 and 9 are those from migrant activists, while most in Cluster2 are those from native activists.

Our research confirms the following findings

- The topical distribution of tweets by migrants and natives shows significant differences in terms of ideas and interests.
- Natives exhibit a higher degree of content-based homophily in their interactions as compared to migrants specifically in relation to friends.

5 Conclusion

This paper presents an analysis via explainable artificial intelligence of features that play a critical role in distinguishing between migrant and native activists on Twitter. Both Twitter metadata and textual contents of tweets were used as features in two state-of-the-art machine learning classification models; and labels derived by application of KMeans clustering over word embeddings are determined to be stable features for the prediction task of migrant activists vs native activists. This essentially points towards a lack of migrants' integration into Irish society in terms of the concerns and interests of the two different communities, where the identification of concerns and interests was captured using gathered topics embodied in word embeddings. The work in this paper can serve as a first step in utilising advanced textual features in solving the societal challenge of assessing migrants' integration into society. In future work, we aim to incorporate mention and retweet networks into the model separately while also incorporating a detailed analysis of conversation threads by both nativesand migrants. To summarise, an analysis such as ours can aid policymakers in gaining deeper insights into the causes behind societal conflicts introduced by inward migration into their countries; it also has significant implications from a European Union perspective.

References

1. Aiello, L.M., Petkos, G., et al.: Sensing trending topics in Twitter. IEEE Trans. Multimedia **15**(6), 1268–1282 (2013)
2. Arastoopour Irgens, G.: Using knowledgeable agents of the digital and data feminism to uncover social identities in the# blackgirlmagic twitter community. Learn. Media Technol. **47**(1), 79–94 (2022)
3. Arık, S.O., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: AAAI, vol. 35, pp. 6679–6687 (2021)
4. Bunting, A.M., Frank, D., Arshonsky, J., Bragg, M.A., Friedman, S.R., Krawczyk, N.: Socially-supportive norms and mutual aid of people who use opioids: An anal- ysis of reddit during the initial Covid-19 pandemic. Drug Alcohol Depend. **222**, 108672 (2021)
5. Burnap, P., Colombo, G., Amery, R., Hodorog, A., Scourfield, J.: Multi-class machine classification of suicide-related communication on twitter. Online Soc. Netw. Media **2**, 32–44 (2017)
6. Burnap, P., Williams, M.L.: Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. Policy Internet **7**(2), 223–242 (2015)
7. Burnap, P., et al.: Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack. Soc. Netw. Anal. Min. **4**(1), 1–14 (2014)
8. Calvete, E., Orue, I., Gámez-Guadix, M.: Cyberbullying victimization and depression in adolescents: the mediating role of body image and cognitive schemas in a one-year prospective study. Eur. J. Crim. Policy Res. **22**(2), 271–284 (2016)
9. Caton, S.: Using social media and text analytics to investigate marginal phenomena. Eurofound working paper WPEF21052 (2022)
10. Caton, S., Hall, M., Weinhardt, C.: How do politicians use Facebook? an applied social observatory. Big Data Soc. **2**(2), 2053951715612822 (2015)
11. Coe, K., Griffin, R.A.: Marginalized identity invocation online: the case of president Donald trump on twitter. Soc. Media Soc. **6**(1), 2056305120913979 (2020)

12. Fan, Y., Zhang, Y., Ye, Y., Li, X.: Automatic opioid user detection from twit- ter: transductive ensemble built on different meta-graph based similarities over heterogeneous information network. In: IJCAI, pp. 3357–3363 (2018)
13. Fanning, B.: Multiculturalism in Ireland. In: Racism and Social Change in the Republic of Ireland, pp. 178–198. Manchester University Press (2018)
14. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics **21**, 768–769 (1965)
15. Gońda, M., Lesińska, M., Pachocka, M.: Relations Between Immigration and Integration Policies in Postwar Europe, pp. 24–45 (2020)
16. Islam, T., Goldwasser, D.: Analysis of twitter users' lifestyle choices using joint embedding model. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, pp. 242–253 (2021)
17. Trinity College Dublin: Equality, diversity and inclusion strategy (2022). https://www.justice.ie/en/JELR/Pages/DoJ EDI Strategy
18. Kilroy, D., Caton, S., Healy, G.: Finding short lived events on social media. In: 28th AIAI Irish Conference on Artificial Intelligence and Cognitive Science (AICS) (2020)
19. Kim, J., Sˆırbu, A., Giannotti, F., Gabrielli, L.: Digital footprints of international migration on twitter. In: International Symposium on Intelligent Data Analysis. pp. 274–286. Springer (2020)
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
21. Matamoros-Fernandez, A., Farkas, J.: Racism, hate speech, and social media: a systematic review and critique. Televis. New Media **22**(2), 205–224 (2021)
22. Mazzoli, M., et al.: Migrant mobility flows characterized with digital data. PLoS ONE **15**(3), e0230264 (2020)
23. de Melo, T., Figueiredo, C.M.: A first public dataset from Brazilian twitter and news on Covid-19 in Portuguese. Data Brief **32**, 106179 (2020)
24. Michael, L.: Reports of racism in Ireland (2020). https://inar.ie/wp-content/uploads/2020/03/2019 iReport Final.pdf
25. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, vol. 26 (2013)
26. Murthy, D., Gross, A., Pensavalle, A.: Urban social media demographics: an exploration of twitter use in major American cities. J. Comput.-Mediat. Commun. **21**(1), 33–49 (2016)
27. Nartey, M.: Centering marginalized voices: a discourse analytic study of the black lives matter movement on twitter. Crit. Discour. Stud. **9**(2), 1–16 (2021)
28. Naseem, U., Razzak, I., Eklund, P.W.: A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. Multim. Tools Appl. **80**(28), 35239–35266 (2021)
29. Norris, P.: The bridging and bonding role of online communities. Press Polit. **7**, 3–14 (2002)
30. Patton, D.U., Lane, J., Leonard, P., Macbeth, J., Smith Lee, J.R.: Gang violence on the digital street: case study of a south side Chicago gang member's twitter communication. New Media Soc. **19**(7), 1000–1018 (2017)
31. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. Lang. Resour. Eval. **55**(2), 477–523 (2020). https://doi.org/10.1007/s10579-020-09502-8

32. Yin, J., Wang, J.: A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD International Confer- ence on Knowledge Discovery and Data Mining, pp. 233–242 (2014)
33. Younus, Arjumand: Use of microblog behavior data in a language modeling frame- work to enhance web search personalization. In: Ma, S., et al. (eds.) AIRS 2016. LNCS, vol. 9994, pp. 171–183. Springer, Cham (2016). https://doi.org/10.1007/ 978-3-319-48051-0 13