Interpreting Arts Audiences and Cultural Preference Through Twitter Data

Shauna Concannon and Matthew Purver

Queen Mary University of London s.concannon@qmul.ac.uk

Abstract. In this study we work with a multi-arts organisation to assess the appropriateness of natural language processing methods for the analysis of audience behaviours and interests via Twitter. We investigate supervised and unsupervised topic modelling methods to investigate whether they can: i) capture the nuanced differences between art genres/forms; ii) characterise users according to their cultural interests; and iii) help to surface the more *culturally mobile* members of the audience. We show promising results: supervised methods showed high accuracy (95%) in filtering data for relevance and pre-defined topics; unsupervised methods provided novel topic discovery, showing sophisticated genre groupings and discovering audience members with wider interests.

1 Introduction

In this paper we outline some approaches to understanding the audiences of cultural organisations by analysing public social data from Twitter as automatically as possible. Working with The Barbican Centre we sought to develop computational methods for the discovery, acquisition and summary of relevant content pertaining to the interests of their audiences.

The Barbican is the United Kingdom's largest multi-purpose arts organisation, with a programme including theatre, music, dance visual art and film; it is keen to extend reach, engage new and wider audiences and build accessibility to arts and learning. We suggest that the publicly available data on Twitter offers a novel and thus far untapped resource for empirical research into audience interests, by providing direct access to their own expressions of their interests, opinions and responses to events and exhibitions.

However, in order to make such a large, freeform dataset manageable and coherent, automated machine learning approaches are necessary; however, these must be able to adapt to a constantly changing programme, and ideally to infer some level of genre differentiation (e.g. discussion of music vs dance or theatre; sculpture vs new media art; or classical music vs hip hop). The linguistic construction of tweets does not always aid such classification: while *Looking forward to seeing Richard III tonight @theBarbican* is linguistically very similar to *Looking forward to seeing Tired Pony tonight @thebarbican*, the former discusses a play and the latter a band. Manually creating a topology of artists and genres would be labour intensive, specialist and require constant updating. Furthermore, identifying the boundary points between art forms or sub-genres is no simple task, particularly in data from a multi-arts venue that prides itself on supporting innovative work that often traverses such genre divides.

We therefore set out to test whether supervised (Naïve-Bayes classification) and unsupervised (Latent Dirichlet Allocation, LDA) machine learning approaches can provide the robustness and adaptability required while providing insights that capture differences between genres. We show that they can filter Twitter data to hone in on relevant conversational material, discover topics from within the data automatically, and ascribe tweets to both pre-defined topics and machine-discovered topics. In turn, this enables the characterisation of individual audience members by topical interest, and locate the more culturally curious, likely candidates for *cultural mobility*: crossing artistic genre boundaries.

Related Work Topic models, e.g. [1–3], are generative models for documents: a topic is modelled as a probability distribution over words, and a document as a distribution over topics. This captures the intutions that topics have distinctive but not mutually exclusive vocabularies, and that documents often discuss multiple topics. By estimating distributions from data, such models can discover the underlying topics in a collection of documents unsupervised, avoid a reliance on labeled training sets, and respond and adapt to the data.

Topic models have proven useful for categorisation and filtering in many text domains, including news articles [4], academic abstracts [5–7] and travel blogs [8]. Twitter messages (tweets) are extremely short texts (140 characters or less) and feature irregular, non-standard English, adding unique challenges to the topic modelling process. However, versions have been used successfully on Twitter data to infer user attributes such as political orientation, gender and ethnicity [9], discover dialogue structure [10], and categorise messages by style or function [11].

2 Data Acquisition and Filtering

Our first task was to automatically acquire data relevant to the Barbican. Using the Twitter API, English language tweets including mention of *barbican*, *@barbicancentre* or *#hackthebarbican* were collected. However, of the 17,812 tweets collected, a large proportion (c.20%) was not relevant to the Barbican arts centre but discussed other *barbicans*: e.g. Barbican, an area of Plymouth, UK; Barbican a non-alcoholic drink; and Barbican Road, Jamaica, home to a popular night club. Attempts to filter out these topics using pre-selected keywords did not work (for example, removing messages with the keyword 'drink' would remove tweets about the non-alcoholic drink, but also anyone drinking at a Barbican event.)

For a more robust method, we experimented with both unsupervised and supervised methods. For the former we used LDA [2], requiring no training data but the specification of number of topics (35); for the latter, an initial set of 3288 tweets were manually labeled as relevant or not relevant (50:50 split) and used to train a supervised Naïve Bayes Classifier. LDA performed surprisingly well given the lack of supervision, mostly sorting irrelevant tweets into their own topics with some of the highest weights, and had the advantage of providing a useful summary of the various irrelevant topics.

Probability	Topic	Class
0.01517	drink beer apple good bottle drinking lol strawberry morning haha malt beverage peach	Irrelevant
0.00859	london create weekend park openeast festival art olympic music great east openeastfestival	Relevant
0.02425	day time good night work plymouth back lunch nice walk today lovely love haha chips hoe	Irrelevant

However, the Naïve Bayes classifier achieved very high accuracy (95.2%) in identifying irrelevant tweets (evaluated using 10-fold cross-validation), and so we used this approach to prepare our sample. With the data now filtered, we moved on to the topic and user modelling phases.

3 Topic Modelling

We now analysed the performance of the same two methods (Naïve Bayes and LDA) for classifying the filtered relevant Barbican tweets according to topic.

Supervised Approach To perform a supervised experiment, we need to topically classify the dataset and produce a set of ground truths. A subset of our dataset consisting of 5,756 randomly sampled tweets was hand-labeled by a single annotator. The categories to which tweets were allocated were taken from the Barbican's genre classification system: Art, Dance, Theatre, Music, Film and Festivals. A further two categories were added, *Location* (encapsulating talk about the physical building, architecture, and generic talk about being at the Barbican Centre, etc.) and *Hack the Barbican*, a month long technology and digital arts takeover of the venue. This was separated out into a separate category as it proved extremely difficult to classify tweets otherwise, as many of its events integrated music, arts film and theatre. The labeled data was then used to train the classifier using the one-vs-all method (taking each class in isolation to train for positive identification vs all other classes) and 10-fold cross-validation. We also conducted an all-vs-all experiment, in which approximately 234 tweets were selected from each hand labeled category to form a training set containing all classes, again using 10-fold cross-validation.

Accuracy in the all-vs-all experiment was relatively high (84%). A one-vs-all approach increases accuracy to nearly 97% in some instances (festival vs. rest). Due to the uneven topic distribution in the data, sample sizes differ for each topic. Sample size in the table below indicates total number of tweets, of which 50% form the positive class (i.e *dance*, etc.).

	All versus	**Rest versus**							
	All	Art	Dance	Film	Festival	Tech	Location	Music	Theatre
Sample size	1878	848	404	1712	456	3062	946	3642	442
Accuracy	83.97%	89.50%	92.33%	96.27%	96.94%	95.00%	93.14%	90.78%	87.33%

Unsupervised Approach We ran LDA with 30 topics, treating each tweet as a "document"; the following table shows the top 10 keywords from the 5 topics with the highest weights (probabilities of occurrence).

'Enjoyment'	'Ludovico'	'HTB'	'Gigs'	'Festival'
time	einaudi	hackthebarbican	mogwai	london
good	ludovicoeinaud	uk	tickets	create
love	night	dollop	zidane	festival
lunch	tonight	hack	tonight	openeast
day	ludovico	today	live	park
back	amazing	jackmaster	mogwaiband	weekend
walk	concert	part	devendra	olympic
night	music	loefah	london	east
nice	time	htb	banhart	music
haha	evening	free	bought	open

The 'enjoyment' category comprises general terms relating to leisure and was, perhaps unsurprisingly, the most pervasive topic in the dataset. While some topics related closely to the description of an experience, others were focused on individual events, such as the weekend festival Open East, or a month long season such as Hack the Barbican; others grouped together musical groups in a related genre. This is very promising: LDA provides a finer-grained categorisation than the manually coded topics (which were purely top level classifications e.g. Music or Theatre), and allows analysis of tweets as mixtures of topics:

Tweet	Topic 1 (Proportion)	Topic 2 (Proportion)
#HTB2013 is "Hack The Barbican" an experiment in creative collaboration using the Barbican Centre's public spaces	HTB (0.9722)	Enjoyment (0.0038)
Can't wait for #OpenEast (Create_London & BarbicanCentre) this wknd noordinarypark. Access is looking fantastic.	Festival (0.9716)	Enjoyment (0.0038)
Just witnessed musical perfection in the form of LudovicoEinaud BarbicanCentre Simply breathtaking #thegeniusEinaudi	Ludovico (0.9693)	Enjoyment (0.0042)

4 User Modelling

Secondly, we investigate user interest profiling by applying the resulting topics to a second corpus of user tweets. For this we take authors of Barbican-related tweets, and find the last 100 tweets from their profile (many tweets will not relate to the Barbican, but should reflect their general interests) to compile a new "document". Within this document, the topic distribution is ascertained, giving a characteristion of interests; we take individuals scoring highly for multiple topics as potential *boundary crossers*, likely candidates for cultural mobility.

Boundary Crossers Selecting users with high weightings for multiple topics identifies individuals who appear to be discussing a range of art related topics; the table below shows two such users. These indications suggest that with further development and refining of how boundary crossers are defined in regards to the number of topics they traverse and what is deemed an appropriate distribution threshold, automatic identification of boundary crossers could be developed.

Boundary crosser 1	**Boundary crosser 2**
Looking forward our next event - Haydn Chamber Orchestra concert on Jan 19th. Tickets available.	We are very much looking forward to seeing rAndomHQ in conversation BarbicanCentre this evening #rainroom pics to follow!
Calling all garden lovers - Join us at the NGS Open Garden, 7 The Grove, Highgate, N6 6JU Sunday 24th March, 2pm-5pm	Incredible #lightshow at #haywardgallery last night, inspiring works by Anthony McCall, Carlos Cruz-Diez & Olafur Eliasson @southbankcentre
Check out this #festival of music, arts, performance,film amd outdoor events June 1-8 in #holloway hollowayfest	Remember all you luxx readers to go check out the beautiful orchid festival on a kewgardens finishes tomorrow
Fancy doing some life drawing? our weekly life drawing class is happening tomorrow(Wednesday) 6.30, do feel free to drop in!	I have 2 top price tkts for #WickedTheMusical weds 15th May 2.30pm matinee. Worth 79. Any takers? I can't go coz I'm working grr.
Wednesday 12th december : special champagne & sparkling wine casino 8.00 pm45 including wine tasting and the meal 3 courses.	Lovely brunch with kelliryder & crystalg at the #whitelabelpreview San Francisco Film Centre

Acknowledgements

This work was supported by CreativeWorks London, a Knowledge Exchange Hub for the Creative Economy funded by the Arts and Humanities Research Council (AHRC). We would like to thank Ryan Nelson of the Barbican for valuable input.

References

1. Hofmann, T.: Probabilistic latent semantic indexing. In: Proc. 22nd ACM SIGIR Conference, ACM (1999) 50–57
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
3. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101**(Suppl 1) (2004) 5228–5235
4. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: Proc. 29th ACM SIGIR Conference, ACM (2006) 178–185
5. Paul, M., Girju, R.: Topic modeling of research fields: An interdisciplinary perspective. In: Proc. 7th RANLP Conference. (2009) 337–342
6. Ramage, D., Manning, C.D., McFarland, D.A.: Which universities lead and lag? toward university rankings based on scholarly output. In: Proc. NIPS Workshop on Computational Social Science and the Wisdom of the Crowds. (2010)
7. Ramage, D., Manning, C.D., Dumais, S.: Partially labeled topic models for interpretable text mining. In: Proc. 17th ACM SIGKDD Conference. (2011) 457–465
8. Paul, M., Girju, R.: Cross-cultural analysis of blogs and forums with mixed-collection topic models. In: Proc. 2009 EMNLP Conference. (2009) 1408–1417
9. Pennacchiotti, M., Popescu, A.M.: A machine learning approach to Twitter user classification. In: Proc. 5th ICWSM Conference. (2011) 281–288
10. Ritter, A., Cherry, C., Dolan, B.: Unsupervised modeling of twitter conversations. In: Proc. 11th NAACL-HLT Conference. (2010) 172–180
11. Ramage, D., Dumais, S.T., Liebling, D.J.: Characterizing microblogs with topic models. In: Proc. 4th ICWSM Conference. (2010) 130–137