Automatic Classification of Tweets for Analyzing Communication Behavior of Museums

Nicolas Foucault, Antoine Courtin

Automatic Classification of Tweets for Analyzing Communication Behavior of Museums

Nicolas Foucault [1,2], Antoine Courtin [3,4]
[1] UMR 7114 – Modèles, Dynamiques, Corpus, Nanterre, France
[2] Julie Desk, Paris, France
[3] Labex "Les Passés dans le Présent", Nanterre, France
[4] Institut National d'Histoire de l'Art, Paris, France
nicolas.foucault@juliedesk.com, antoine.courtin@mac.com

Abstract

In this paper, we present a study on tweet classification which aims to define the communication behavior of the 103 French museums that participated in 2014 in the Twitter operation: *MuseumWeek*. The tweets were automatically classified in four communication categories: sharing experience, promoting participation, interacting with the community, and promoting-informing about the institution. Our classification is multi-class. It combines Support Vector Machines and Naive Bayes methods and is supported by a selection of eighteen subtypes of features of four different kinds: metadata information, punctuation marks, tweet-specific and lexical features. It was tested against a corpus of 1,095 tweets manually annotated by two experts in Natural Language Processing and Information Communication and twelve Community Managers of French museums. We obtained an state-of-the-art result of F_1-score of 72% by 10-fold cross-validation. This result is very encouraging since is even better than some state-of-the-art results found in the tweet classification literature.

Keywords: NLP, classification, tweets, communication, behavior, museums, annotation, community managers, corpus, Twitter, MuseumWeek.

1. Introduction

In the last decade, the two American companies, Facebook and Twitter, have captured the lion's share of the competitive international market built around the creative and participatory practices of social media. Their business model is based on the expectations and participation of audiences. Beyond studying the behavior of audiences, social media call for a deeper analysis of the actions of professionals, such as Community Managers (CMs) of museums, who use social networks for promotional/institutional purposes, in order to better understand their practices and intentions Johnson et al. (2012). To do so, it is essential to achieve a fine-grained classification of the communication content of their messages in micro-blogs such as those found on Twitter, on a large-scale.

The work presented in this paper was conducted with this aim in view and is part of the NOS project[1]. NOS aims at providing CMs of French cultural institutions with a tool dedicated to help them in the analysis of social-media messages sent by other institutions or users.

2. Literature Overview

Tweet analysis has led to a large number of studies in many domains such as ideology prediction in Information Sciences (Djemili et al., 2014), spam detection in Security (Yamasaki, 2011), dialog analysis in Linguistics (Boyd et al., 2010), and natural disaster anticipation in Emergency (Gelernter and Mushegian, 2011; Sakaki et al., 2013).

Complementary efforts have been made in Social Sciences and Digital Humanities to develop tweet classifications (Dann, 2010; Riemer and Richter, 2010; Shiri and Rathi, 2013; Stvilia and Gibradze, 2014).

However, few studies aim at classifying tweets according to communication classes. They mostly rely on small reference sets analyzed by experts in Information Communication (InfoCom) rather than by Twitter users. An exception worth mentioning is the work presented in Lovejoy and Saxton (2012), in which the authors (Twitter users) analyze the global behavior of nonprofit organizations on Twitter based on three communication classes: Information, Community and Action classes.

Recently, several studies on tweet classification have been carried out in Natural Language Processing (NLP) (Karimi et al., 2012; Kothari et al., 2013; Lin et al., 2014; Zubiaga et al., 2015). Basically, these analyses aim at categorizing open-domain tweets using a reasonable amount of manually classified data and either small sets of specific classes (e.g. positive versus negative classes in sentiment analysis) or larger sets of generic classes (e.g. News, Events and Memes classes in topic filtering). To the best of our knowledge, only Courtin et al. (2014) has classified institutional tweets in communication categories based on NLP techniques. The advantage of NLP approaches is that they can automatically classify large corpora of tweets. The most commonly used models are supervised learning, Support Vector Machine (SVM) and Naive Bayes (NB) (Sriram et al., 2010; Kouloumpis et al., 2011; Kothari et al., 2013; Malandrakis et al., 2014). In supervised learning, features are extracted from tweets and metadata and then vectorized as training examples to build models.

Compared to English, little work has been done on tweet classification in French except in sentiment analysis (Fraisse and Paroubek, 2015), and only two corpora of (non institutional) tweets exist: the FRENCH SOCIAL MEDIA BANK (Seddah et al., 2012) (sms, forums and tweet texts on general topics) and the CMR-POLITITWEETS corpus (Chanier et al., 2014) (34K tweets on politics).

[1] bit.ly/1Z5VI7x

The rest of this paper is organized as follows: in Section 3., we present the corpus of tweets we collected on Twitter and used in our experiments. In Section 4., we present the classes and features created in our work. Section 5. details the approach we followed to classify tweets automatically in communication categories. In this section, we also present the training and testing data used and the results obtained. In section 6., we present an analysis of the behavior of the 103 French museums who participated in the *MuseumWeek* in 2014 based on our classification method. Finally, we conclude and suggest some perspectives of this work in Section 7.

3. Corpus

We collected an initial corpus of 38,756 French tweets (i.e. the MW14 corpus) based the Twitter stream API [2]. We used Python's library Twython [3] to request the API based on the seven hashtags defined for the first edition of *MuseumWeek*. *MuseumWeek* is a cultural event driven by a group of twelve community managers of French museums and Twitter (Courtin et al., 2014). It took place over a period of seven days from the 24th to the 30th of March, 2014. Each day, a different theme (i.e. a hashtag) allowed any registered institution to value its collections, activities and programming while encouraging the public to share their own experiences/content on Twitter. A unique set of hashtags was defined for each country participating in the event. France's set is presented in Table 1.
The final version of the MW14 corpus contains 33,658 preprocessed tweets (9,933 tweets and 23,725 retweets). The number of tweets sent by the 103 French institutions who participated in the *MuseumWeek* was 6,691 (5,307 tweets and 1,384 retweets). Table 2 provides the number of tweets for each version of this corpus per theme of the *MuseumWeek*.

Day	Theme	Definition
1	#CoulissesMW	Discovering behind the scenes
2	#QuizzMW	Checking our knowledge
3	#LoveMW	Sharing our "coup de cœur"
4	#ImagineMW	Free imagination
5	#QuestionMW	Taking time to share
6	#ArchiMW	Appreciating museum's building architecture
7	#CreaMW	You are the artist

Table 1: Theme of the *MuseumWeek* (2014).

The preprocessing applied to obtain the final version of the MW14 corpus was threefold: data cleaning, data extraction, and data normalisation.

Data cleaning comprised two stages:

(i) Removing invalid/duplicated tweets from Twython's output (json structure);

MW14 corpus	Processing step	
	Collection	Preprocessing
CoulisseMW	6,557	6,040
QuizzMW	4,947	4,296
LoveMW	7,999	7,210
ImagineMW	5,565	4,222
QuestionMW	2,978	2,382
ArchiMW	6,051	5,522
CreamMW	4,659	3,986
Total	**38,756**	**33,658**

Table 2: Number of tweets for each theme found in the MW14 corpus and each data processing step applied on it.

(ii) Solving encoding errors so that tweets were encoded properly in UTF-8 (with Unicode characters).

Data extraction was also a two-stages process:

(i) Extraction of relevant information from the cleaned data;

(ii) Conversion to csv format (the same charset as for the cleaned data). An example of extracted data is given in Figure 1 (the first line corresponds to a csv heading).

Data normalisation is one-fold: it is an iterative process dedicated to textual normalization. We did not have to undertake a complex normalisation process as is often the case with micro-blogging data (Chanier et al., 2014). Pattern matching based on regular expressions was used to map syntactic, lexical and Twitter specific forms found in tweets to their normalized forms. Mostly, we dealt with abbreviated forms and tokenization errors related to Twitter forms as illustrated in Table 3 (examples 5 and 6 are context-dependent in this table).

```
tweet_id,date_UTC,username,tweet_text
448230958396080128,2014-03-24
22:52:42,Fred,"RT @CVersailles:  Dernier
tweet de la journée #CoulissesMW, demain
la MuseumWeek continue avec #QuizzMW,
bonne nuit !  http://t.co/MeTw3kkvBO"
```

Figure 1: Example of MW14 data (csv format).

Data extraction was done in csv and was performed at streaming time, to avoid having to deal with heavy json structures as returned by the Twitter API later on.
The MW14 corpus is available in xslx format [4] (initial version only) but will be available (initial and preprocessed versions) on a repository at the Bibliothèque Nationale de France (BnF) and on Dataverse [5] in csv format.

[2] https://dev.twitter.com
[3] https://github.com/ryanmcgrath/twython
[4] http://bit.ly/1PM68Y2
[5] http://www.bnf.fr; http://dataverse.org

Id	Initial form	Final form	Form type
1	*#AchiMW*	*#ArchiMW*	Twitter
2	*@ leCNM*	*@leCNM*	Twitter
3	*pquoi*	*pourquoi*	Lexical
4	*collec°*	*collection*	Lexical
5	+	+ or *plus*	Lexical
6	*vs*	*vous* or *versus*	Lexical

Table 3: Textual normalisation examples.

4. Classification

The main idea of our approach is to categorize automatically tweets based on their textual content. Our classification is multi-class (i.e. any given tweet may be classified into more than one category). In this section, we introduce the classes and features used in our experiments and presented in section 5.

Both categories and features have been determined based on a small set of 200 tweets taken from the MW14 corpus. As hashtags are a distinctive characteristic of tweets (Jackiewicz and Vidak, 2014), these 200 tweets have been selected randomly based on the distribution of tweets found in the MW14 corpus over the different hashtags of the *MuseumWeek* (see Table 1).

The categories and features presented in this section have been designed by two researchers familiar with Twitter: a specialist in NLP and an expert in InfoCom who works in the cultural field.

4.1. Categories

We defined four communication categories to classify tweets in our work:

1. **Sharing experience**: sharing an experience/opinion/sentiment (e.g. *@TanjaPraske pouvez-vous nous en dire plus?*);
2. **Promoting participation**: asking users to do something and/or to participate in an activity on-line or in-real-life (e.g. *Et si vos ados venaient enregistrer leur jam session au #Studio1316 ? http://t.co/33Tq3d2H*);
3. **Interacting with the community**: hailing or replying to one or several accounts at the same time (e.g. *@NathBoiss en effet, depuis plusieurs mois la tenture est en restauration. Chaque pièce sera bichonner ainsi! ;)*);
4. **Promoting-informing about the institution**: promoting or informing people about activities, collections and practical information concerning the museum (e.g. *Découvrez les processus d'aller-retour dans l'oeuvre de #Matisse avec l'expo "Paires et séries" http://t.co/qYXpeMaK*).

These categories were defined by our expert in InfoCom and validated by two Community Managers (CMs) of French museums in Paris. They have been initially used in Courtin et al. (2014).

4.2. Features

We used 18 types of features subsumed in four main types in order to represent tweets in this study. They are based on manual analysis performed on the aforementioned set of 200 tweets by our NLP specialist and InfoCom expert, in addition of automatic term frequency extractions conducted on the MW14 corpus.

These features extend those used in Courtin et al. (2014) and draw on two previous studies done on tweet classification in sentiment analysis (Kouloumpis et al., 2011) and comment detection (Kothari et al., 2013). They were assigned boolean value, coding for the presence or absence of a specific piece of information in the textual content of tweets. The four main types of features used in this study are:

- **Metadata information**: whether the author of a tweet is a museum (`is_museum`), whether the author of the tweet is mentioned in the tweet (`@self`), whether a museum is mentioned in the tweet (`@museum`), whether another Twitter user than the author of the tweet is mentioned in the tweet `@user`);
- **Punctuation marks**: whether a tweet contains an exclamation mark (`punct!`) or a question mark (`punct?`). We noticed that these marks are particularly common in interactions between museum individuals;
- **Tweet-specific features**: whether a tweet contains url (`url`), hashtags (`#`), emoticons (`smileys`), whether a tweet is a modified tweet (`is_MT`) and whether it contains Twitter conventions such as cc (`cc`) or .@ (`.@`). The latter two features both give the tweet more visibility. The first one means: "share my tweet with all the friends mentioned explicitly after the cc reference" while the other one means: "make my tweet visible to everybody". These two features mostly appear in tweets of classes 1–3;
- **Lexical features**: each subtype of lexical features described below is detailed in Appendix A (Table A.1).
  - `cli_pro`: clitics and pronouns related to *je*, *on*, *nous* and *vous* (used in tweets of classes 1–2);
  - `greetings`: used in tweets promoting participation and by CMs to hail people;
  - `sup_aff`: superlative and affective forms (e.g. *très*, *beau*);
  - `Qtag`: question tags used by CMs to interact with people;
  - `imp_forms`: imperative forms (e.g. *tweetez* and *participez*), most likely to appear in museum tweets to make people share their experience;
  - `vocab`: museum, collection and photo-based vocabulary such as *galerie*, *exposition* and *selfie* as well as specific lemmas related to tomorrow and today (often used by museums to inform people about novelty).

5. Evaluation

In this section, we present the data used to train and test our classifiers, the methodology applied and the related results.

5.1. Training and Testing Data

We annotated manually a total of 1,309 tweets into categories. These data were split into four corpora: TRN, DEV, TST and CMR. Tweet distributions for each corpus are given in Table 4. All the data were selected randomly from the MW14 corpus (normalized version; no duplicates) and all annotations were done using the same guideline which describes the annotation procedure, rules and categories along with some examples.

Dist.	TRN	DEV	TST	CMR	All	%All
cat1	129	10	8	121	268	24
cat2	106	5	37	65	213	20
cat3	211	14	29	57	311	28
cat4	106	37	17	143	303	28
All	552	66	91	386	1,095	100

Table 4: Overall and per class tweet distributions.

The NLP and InfoCom experts who designed the categories and features aforementioned also annotated the three first corpora. The TRN and DEV corpora were annotated by the former expert while both experts annotated the TST corpus. The CMR corpus contains annotations made by twelve Community Managers of museums during an annotation campaign that we conducted. The campaign consisted in making six different samples of 100 tweets to be annotated by two different CMs each. So far, no inter-annotators consensus has been assigned. Therefore, the actual CMR corpus comprises only the reference annotations corresponding to straightforward agreements between annotators. The total number of references in this corpus represents 386 annotations over the 600 performed by the CMs. In addition to the other references (TRN: 552, DEV: 66 and TST: 91 tweets), the total number of references used in our experiments was 1,095 annotations.

We used Cohen's Kappa inter-annotator agreement (Cohen, 1960) as implemented in R [6]. Kappa results for the CMR and TST data are given in Table 5. According to Landis and Koch (1977) interpretation scale, these results are acceptable, lying between moderate (CMR) and strong (TST). In particular it is interesting to note that qualitatively there is no difference between those who are familiar with the annotation task (our NLP and InfoCom experts) and those unfamiliar with the task (CMs) as the inter-annotator agreements obtained on the two corpora are almost the same. Our intuition on this point is that CMs balanced their lack of experience in the annotation domain with their knowledge in institutional communication and their expertise of Twitter which was the reason why we preferred to choose CMs rather than professional annotators in the first place.

[6] http://personality-project.org/r/psych

Annotation	K	# tweets
CMR-S1	.44	100
CMR-S2	.61	100
CMR-S3	.64	100
CMR-S4	.40	100
CMR-S5	.70	100
CMR-S6	.50	100
CMR mean	.55	100
TST	.61	91

Table 5: Inter-annotator agreements (K) obtained on the Community Managers samples (CMR-S1 to CMR-S6) and the TST corpus. # tweets: number of tweets.

5.2. Methodology

We used standard precision (**P**) and recall (**R**) rather than accuracy to evaluate our approach (Ben-David, 2007). We opted for an $\mathbf{F}_{0.5}$-measure as in Kothari et al. (2013) since this measure is more appropriate to evaluate performances if precision prevails over recall as in our case [7].

We conducted global and per class evaluations and applied both direct and 10-CV (Kohavi, 1995) evaluations. In the first case (i.e. all evaluations except 10-CV), the training corpus was: TRN. Otherwise (i.e. for 10-CV evaluation), training/testing splits were defined with respect to the overall and per class proportions of tweets indicated in Table 4 (%All column). Consequently, 10-CV training and testing splits comprised 986 and 109 tweets per run respectively. In both cases, we relied on the *Scikit-learn* machine learning framework (Pedregosa et al., 2011) to build our classifiers, based on the one-against-all schema (Rifkin and Klautau, 2004) to train our SVMs and all standard parameters otherwise (i.e. for NB).

5.3. Results

Overall and per class results are presented respectively in Tables 6 and 7 while the impact of each main kind of feature is described in Section 4. on the classification performances is given in Table 8.

Eval.	**P**	**R**	$\mathbf{F}_1$	$\mathbf{F}_{0.5}$
TRN	.746	.768	.757	.750
DEV	.774	.742	.757	.767
TST	.541	.780	.639	.577
CMR	.693	.841	.760	.718
10-CV	.680	.768	.721	.696

Table 6: Overall results (all features).

In Table 6, can be seen that the result performed on the TRN and DEV corpora are the same ($\mathbf{F}_1$ score of .757 for each corpus). This table also shows that the result obtained by CV on all the corpora merged together was very close to the previous result with an overall $\mathbf{F}_1$ score of .721. This

[7] However, the tables of results give $\mathbf{F}_1$ scores for the sake of comparison with other studies.

result is rather surprising since CV evaluation usually leads to lower results than direct evaluation (as it takes into account 10 times more results on smaller and much more varied samples of data) and means that the categories, features and corpora produced are consistent enough to automate the classification process of tweets in communication classes according to their textual content. The corollary of this result is that the different corpora produced, and especially the TRN and DEV corpora annotated by the same person (our NLP specialist) and the CMR corpus annotated by a large number of non professional annotators (twelve CMs) in a very segmented way (six subcorpora of 100 tweets, each subcorpus annotated by two different CMs), present annotations of quality. Even the CMR corpus presents the best results with an F_1 score of .760 which is very promising for future work as it is a state-of-the-art result (Pak and Paroubek, 2010; Kouloumpis et al., 2011; Maynard et al., 2012; Zubiaga et al., 2014; Zubiaga et al., 2015).

Eval. - CMR	**P**	**R**	F_1	$\mathbf{F}_{0.5}$
cat1	.618	.847	.715	.653
cat2	.576	.853	.688	.616
cat3	.809	.750	.778	.796
cat4	.822	.870	.845	.831

Table 7: Results per class (all features).

The lower performances obtained on the TST data as shown in Table 6 can be attributed to the number of tweets in the first two categories (respectively 8 and 37 tweets) compared to the small amount of data it contains. In fact, these categories are the most difficult to distinguish for classification purposes as they obtained the lowest F_1 scores in the per class evaluation results presented in Table 7 (.653 and .616 for each category respectively). Further investigations are necessary concerning this point.

Eval. - CMR	**P**	**R**	F_1	$\mathbf{F}_{0.5}$
lex	.567	.745	.644	.595
lex+punct	.580	.752	.655	.607
lex+punct+meta	.701	.832	.761	.723
lex+punct+meta+tweets	.693	.841	.760	.718

Table 8: Results per main types of features (CMR corpus).

Finally, Table 8 shows that metadata information features are the most discriminating for the classification of tweets in our communication categories since the overall performance gain is about 12% in terms of F_1 score (lex+punct: .607 versus lex+punct+meta: .723). The performance gain with punctuation marks is much smaller (about 1%) whereas there is no gain concerning the tweet-specific features; the result is even slightly lower compared to the lex+punct+meta combination (lex+punct+meta+tweets: .718). This is very surprising as this kind of features always provides a classification benefit in the literature (Kouloumpis et al., 2011) and therefore needs further investigation. It would be interesting for example to carry out the same kind of analysis as here but for each subtype of features in order to filter out irrelevant features.

6. Communication Behavior

The global behavior of the 103 Museums which participated in the *MuseumWeek* is given in Figure 2. This analysis was obtained by applying our classification method presented in the previous section to the 5,307 tweets (no retweets) sent by the 103 French museums during the *MuseumWeek*. In this case, we trained our classifiers on the TRN corpus.

Figure 2 reveals for instance that French museums were communication-centric since most of their tweets concerned promotion of their own institution (in pink on the figure). It also shows that very few participants exchanged with people on Twitter since the proportion of tweets categorized as Interacting with the community (in blue on the figure) is only the third most representative behavior, except for the most active museums, i.e., the twelve museums that sent more than 100 tweets during the week like the Centre Pompidou (302 tweets) and Musée du Quai Branly (263 tweets).

7. Conclusion and Perspectives

In this work, we collected a corpus of 30K tweets from Twitter during the first edition of the *MuseumWeek* and used it to create the first event-based corpus (available online) of French tweets categorized in communication classes. Our approach is multi-class and combines SVMs and Naive Bayes classifiers based on four kinds of features: metadata information, punctuation marks, tweet-specific and lexical features. We evaluated it on 1,095 tweets annotated by twelve community managers of museums, one NLP specialist and an expert in Information Communication. We obtained a state-of-the-art result of .695 $F_{0.5}$-score.

We are still completing our reference corpus (CMR corpus) and we plan in future work to extend our classifier using n-grams and POS tag features. In the meanwhile, this study was also applied on a new corpus of data collected during the *MuseumWeek* last year (600K tweets) in order to carry out a comparative study of the *MuseumWeek* event between 2014 and 2015. This study was recently published officially for the French Minister of Culture (Courtin and Foucault, 2015).

8. Bibliographical References

Ben-David, A. (2007). A Lot of Randomness is Hiding in Accuracy. *Engineering Applications of Artificial Intelligence*, 20(7):875–885.

Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *Proceedings of the 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10.

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C., Hriba, L., Longhi, J., and Seddah, D. (2014). The CoMeRe corpus for French: Structuring and Annotating Heterogeneous CMC Genres. *JLCL*, 29(2):1–30.

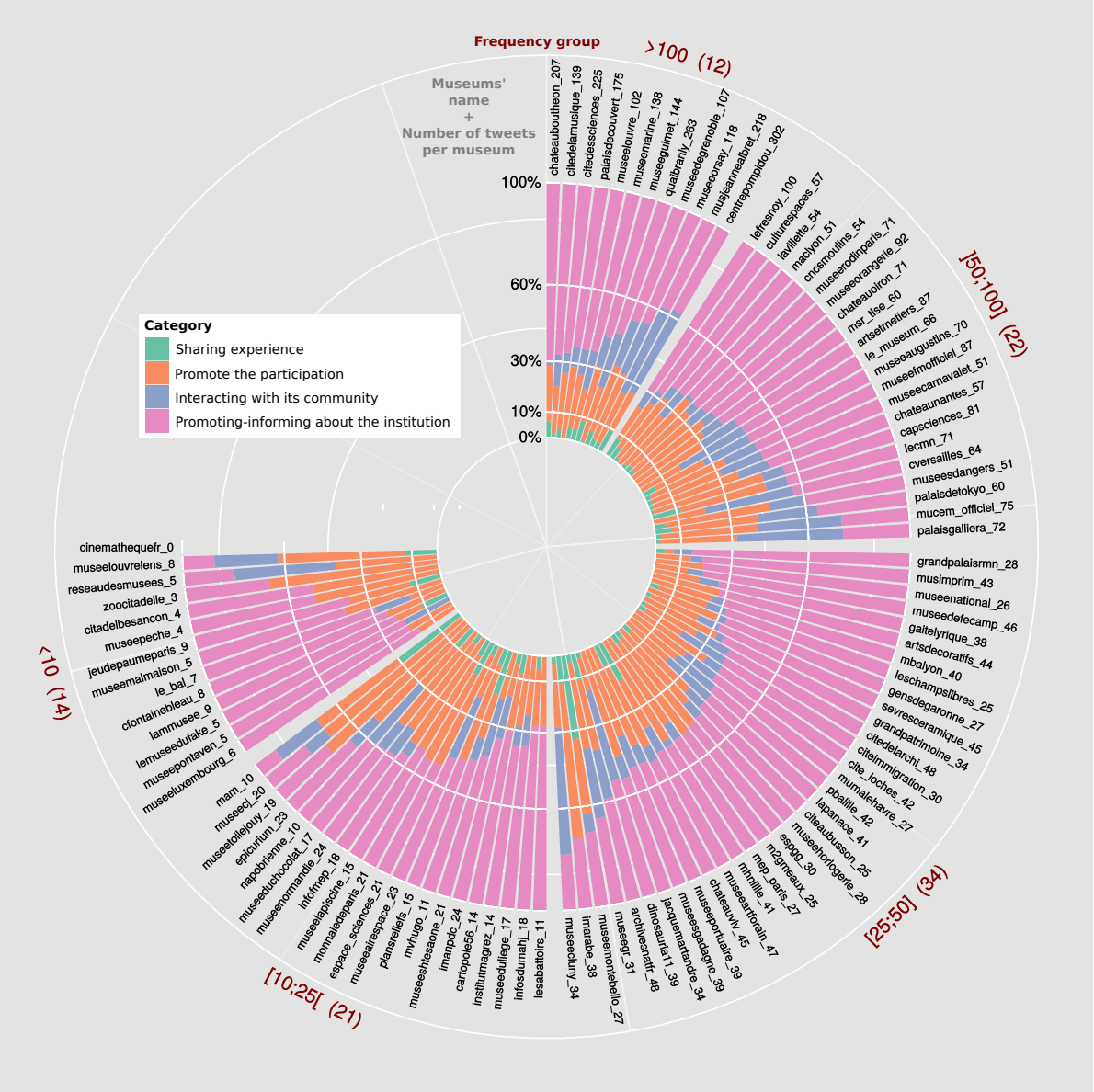

Figure 2: Behavior of the 103 French museums which participated in the *MuseumWeek* (2014). Behavior is defined according to our 4 communication categories. Museums are presented by frequency groups in descending order, based on the total number of initial tweets (i.e. no retweets) they sent during the event.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Courtin, A. and Foucault, N. (2015). Quantitative and Categorial Analysis of Tweets Sent During the #MuseumWeek2015 Event. Research Report V0.2, Labex les passés dans le présent, October.

Courtin, A., Juanals, B., Minel, J., and de Saint Léger, M. (2014). A Tool-Based Methodology to Analyze Social Network Interactions in Cultural Fields: The Use Case "MuseumWeek". In *Proceedings of the 6th International Conference on Social Informatics (SocInfo'14)*, pages 144–156.

Dann, S. (2010). Twitter Content Classification. *First Monday*, 15(12).

Djemili, S., Longhi, J., Marinica, C., Kotzinos, D., and Sarfati, G.-E. (2014). What does Twitter have to say about ideology? In *NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication / Social Media*, volume 1, pages 16–25.

Fraisse, A. and Paroubek, P. (2015). Comparable Microblogs for Multilingual Affective lexicons. *Natural Language Engineering*, 1(1):1–20.

Gelernter, J. and Mushegian, N. (2011). Geo-parsing Messages from Microtext. *Transactions in GIS*, 15(6):753–773.

Jackiewicz, A. and Vidak, M. (2014). Étude sur les mots-dièse. *SHS Web of Conferences*, 8:2033–2050.

Johnson, L., Adams, S., and Cummins, M. (2012). NMC Horizon Report: 2012 Higher Education Edition. Technical report, New Media Consortium.

Karimi, S., Yin, J., and Thomas, P. (2012). Searching and Filtering Tweets: CSIRO at the TREC 2012 Microblog Track. In *Proceedings of The 21st Text REtrieval Confer-*

ence (TREC'12).

Kohavi, R. (1995). A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 1137–1143.

Kothari, A., Magdy, W., Darwish, K., Mourad, A., and Taei, A. (2013). Detecting Comments on News Articles in Microblogs. In Emre Kiciman, et al., editors, *Proceedings of the 7th International Conference on Web and Social Media (ICWSM)*. The AAAI Press.

Kouloumpis, E., Wilson, T., and Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! In *Proceedings of the 5th International Conference on Web and Social Media (ICWSM)*.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Shou-De Lin, et al., editors. (2014). *Proceedings of the 2nd Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics and Dublin City University, Dublin, Ireland.

Lovejoy, K. and Saxton, G. D. (2012). Information, Community, and Action: How Nonprofit Organizations Use Social Media. *Journal of Computer-Mediated Communication*, 17(3):337–353.

Malandrakis, N., Falcone, M., Vaz, C., Bisogni, J. J., Potamianos, A., and Narayanan, S. (2014). SAIL: Sentiment Analysis using Semantic Similarity and Contrast Features. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval'14)*, pages 512–516, Dublin, Ireland.

Maynard, D., Bontcheva, K., and Rout, D. (2012). Challenges in developing opinion mining tools for social media. In *Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at International Conference on Language Resources and Evaluation, LREC 2012, 26 May 2012, Istanbul, Turkey*. European Language Resources Association.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Riemer, K. and Richter, A. (2010). Tweet Inside: Microblogging in a Corporate Context. In *Proceedings 23rd Bled eConference eTrust: Implications for the Individual, Enterprises and Society*.

Rifkin, R. and Klautau, A. (2004). In Defense of One-Vs-All Classification. *Journal of Machine Learning Research*, 5:101–141.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2013). Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):919–931.

Seddah, D., Sagot, B., Candito, M., Mouilleron, V., and Combet, V. (2012). The French Social Media Bank: a Treebank of Noisy User Generated Content. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*, pages 2441–2458.

Shiri, A. and Rathi, D. (2013). Twitter Content Categorisation: A Public Library Perspective. *Journal of Information & Knowledge Management (JIKM)*, 12(04).

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). Short Text Classification in Twitter to Improve Information Filtering. In Fabio Crestani, et al., editors, *SIGIR*, pages 841–842. ACM.

Stvilia, B. and Gibradze, L. (2014). What do Academic Libraries Tweet About, and What Makes a Library Tweet Useful? *Library & Information Science Research*, 36(3–4):136–141.

Yamasaki, S. (2011). A Trust Rating Method for Information Providers over the Social Web Service: A Pragmatic Protocol for Trust among Information Explorers and Information Providers. In *Proceedings of the 11th Annual International Symposium on Applications and the Internet (SAINT'11)*, pages 578–582.

Zubiaga, A., Vicente, I. S., Gamallo, P., Campos, J. R. P., Loinaz, I. A., Aranberri, N., Ezeiza, A., and Fresno-Fernández, V. (2014). Overview of tweetlid: Tweet language identification at SEPLN 2014. In *Proceedings of the Tweet Language Identification Workshop co-located with 30th Conference of the Spanish Society for Natural Language Processing, TweetLID@SEPLN 2014, Girona, Spain, September 16th, 2014.*, pages 1–11.

Zubiaga, A., Spina, D., Martínez, R., and Fresno, V. (2015). Real-time Classification of Twitter Trends. *Journal of the Association for Information Science and Technology*, 66(3):462–473.

A Lexical Features for Tweet Classification

cli_pro	**Clitics and pronouns related to:**	
	vous	"vous", "votre", "vos", "votres"
	nous	"nous", "notre", "nos"
	on	"on"
	je	"je", "j", "moi", "me", "ma", "mes", "m", "mon", "mien"
greetings	**Greetings:**	
	merci	"merci", "bravo"
sup_aff	**Superlative and affective forms:**	
	preferer	"préférer", "préfère", "préféré"
	positif	"aime"," belle" , "beau", "bel", "plaisir", "amour", "trésor", "adore", "adoré", "<3", "coup de cœur", "cœur", "superbe", "sublime", "merveille", "bien"
	emphase	"très", "trop", "plus", "bcp", "beaucoup", "encore", "mieux"
Qtag	**Question tags:**	
	question	"combien", "parce que", "pourquoi", "question"
	interrogation	"quel", "quoi", "qui", "que", "quand"
imp_forms	**Imperative forms:**	
		"faites", "aimez", "tweetez", "suivez", "racontez", "publiez", "participez", "dites", "inventez", "échangez", "rendez"
vocab	**Museum, collection and photo-based vocabulary:**	
	photo	"photo", "selfie", "vidéo"
	collection	"œuvre", "expo"
	museum_word	"musée", "visite", "salle", "galerie", "tableau", "collection"
	aujour_demain	"aujourd'hui", "demain", "journée", "jour", "bonjour", "thème"
	voici_voila	"voici", "voilà"

Table A.1: Detail of each subtype of lexical features used in this study, subdivided into finer subgroups of features.