A TOPIC MODELING BASED REPRESENTATION TO DETECT TWEET LOCATIONS. EXAMPLE OF THE EVENT "JE SUIS CHARLIE"

*Mohamed Morchid[a], Didier Josselin[c,a], Yonathan Portilla[a,b], Richard Dufour[a], Eitan Altman[b,a], Georges Linarès[a]

[a] Laboratoire Informatique d'Avignon (LIA), University of Avignon (France) - {*firstname.lastname*}*@univ-avignon.fr*
[b] INRIA, B.P 93, 06902 Sophia Antipolis Cedex (France) - {*firstname.lastname*}*@inria.fr*
[c] UMR ESPACE 7300, CNRS, UNSA (France)

KEY WORDS: Tweets location, Topic modeling, Author topic model, Twitter

ABSTRACT:

Social Networks became a major actor in information propagation. Using the Twitter popular platform, mobile users post or relay messages from different locations. The tweet content, meaning and location, show how an event-such as the bursty one "JeSuisCharlie", happened in France in January 2015, is comprehended in different countries. This research aims at clustering the tweets according to the co-occurrence of their terms, including the country, and forecasting the probable country of a non-located tweet, knowing its content. First, we present the process of collecting a large quantity of data from the Twitter website. We finally have a set of 2,189 located tweets about "Charlie", from the 7th to the 14th of January. We describe an original method adapted from the Author-Topic (AT) model based on the Latent Dirichlet Allocation (LDA) method. We define an homogeneous space containing both lexical content (words) and spatial information (country). During a training process on a part of the sample, we provide a set of clusters (topics) based on statistical relations between lexical and spatial terms. During a clustering task, we evaluate the method effectiveness on the rest of the sample that reaches up to 95% of good assignment. It shows that our model is pertinent to foresee tweet location after a learning process.

1. INTRODUCTION AND STATE OF THE ART

The exponential growth of available data on the Web enables users to access a large quantity of information. Micro-blogging platforms evolve in the same way, offering an easy way to disseminate ideas, opinions or common facts under the form of short text messages. Depending on the sharing platform used, the size of these messages can be limited to a maximum number of words or characters. Although Twitter is a recent information-sharing model, it has been widely studied. Many works have focused on various aspects of Twitter, such as social impact (Kwak et al., 2010), event detection (Zhao et al., 2011), user influence (Cha et al., 2010), sentiment analysis (Tumasjan et al., 2010), hashtag analysis (Huang et al., 2010), or theme classification (Morchid et al., 2014a).

The aim of the proposed approach is to locate a given tweet by using the tweet content (a set of words). Nonetheless, the Twitter service does not allow to send messages whose size exceeds 140 characters. This constraint causes the use of a particular vocabulary that is often unusual, noisy, full of new words, including misspelled or even truncated words (Choudhury et al., 2007). Indeed, the goal of these messages is to include a lot of information with a small number of characters. Thus, it may be difficult to understand the meaning of a short text message (STM) with only the tweet content (words). Several approaches have been proposed to represent the tweet content. The classical bag-of-words approach (Salton and Buckley, 1988) is usually used for text document representation in the context of keyword extraction. This method estimates the Term Frequency-Inverse Document Frequency (TF-IDF) of the document terms. Although this unsupervised approach is effective for a large collection of documents, it seems unusable in the particular case of short messages since most of the words occur only once (hapax legomena (Baayen, 1998)). Other approaches propose to consider the document as a mixture of latent topics to work around segments of errors. These methods build a higher-level representation of the document in a topic space. All these methods are commonly used in the Information Retrieval (IR) field. They consider documents as a bag-of-words without taking account of the words order. Nevertheless, they demonstrated their performance on various tasks. Several approaches considered a text document as a mixture of latent topics. These methods, such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990, Bellegarda, 1997), Probabilistic LSA (PLSA) (Hofmann, 1999) or Latent Dirichlet Allocation (LDA) (Blei et al., 2003), build a higher-level representation of the document in a topic space. Document is then considered as a bag-of-words (Salton, 1989) where the word order is not taken into account. These methods have demonstrated their performance on various tasks, such as sentence (Bellegarda, 2000) or keyword (Suzuki et al., 1998) extraction. LDA is a generative model of statistics which considers a document, seen as a bag-of-words, as a mixture probability of latent topics. In opposition to a multinomial mixture model, LDA considers that a theme is associated with each occurrence of a word composing the document, rather than associating a topic with the complete document.

Thereby, a document can belong to different topics from a word to another. However, let us notice that the word occurrences are connected by a latent variable which controls the global respect of the topic distribution in the document. These latent topics are characterized by a distribution of word probabilities which are associated with them. During the LDA learning process, distribution of words into each topic is estimated automatically. Nonetheless, the location associated with the tweet is not directly taken into account in the topic model. As a result, such a system considers separately the tweet content (words), to learn a topic model, and the labels (location) to train a classifier. Thus, the relation between the tweet content and its location (country) is crucial to efficiently locate (unknown) new tweets. In this paper, we propose to build a topic model, called author-topic (AT) (Rosen-Zvi et al., 2004, Morchid et al., 2014b) that takes into consideration all information contained in a tweet: the content itself (words), the label (country) and the relation between the distribution of

*Corresponding author

words into the tweet and the location, considered as a latent relation. From this model, a vector representation in a continuous space is built for each tweet. Then, a supervised classification approach, based on Support Vector Machines (SVM) (Vapnik and Lerner, 1963) is applied. For mathematical and methodological details, see (Morchid et al., 2014b).

Another complementary dimension of the tackled problem is geographical. Indeed, much information about located events spread over the territory in the world from their original location to other countries. An event can be considered as a flow that depicts a spatial buzz. It is then interesting to study this kind of process to assess where the impact of this specific event was the strongest and how it evolves in time. However, Twitter does not provide the entire data sets of tweets, that would be anyway unusable due to the too large quantity of data. It is expected that Twitter publishes 0.01 p.c. of located tweets from the whole data set. Moreover, it is not proved that the resulting sample is representative of all the tweets exchanged. At least, it is what we observed regarding the tweets about the "Charlie" event.

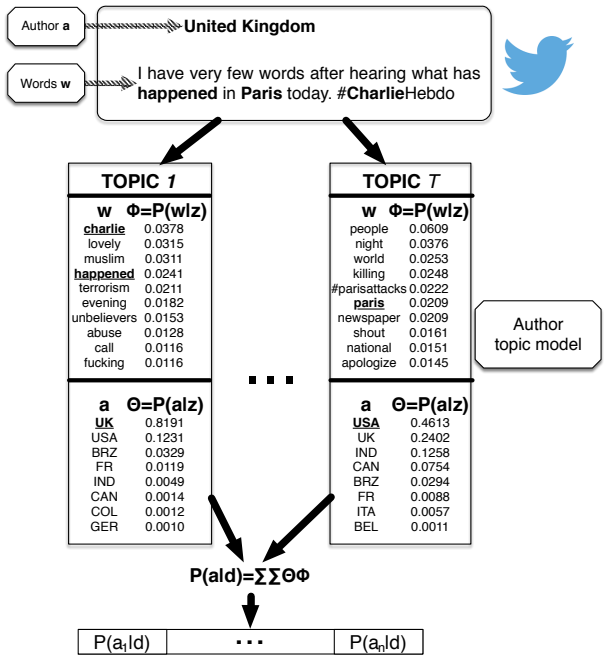

Figure 1: Example of a tweet d mapped into an author topic model of size T.

2. AUTHOR-TOPIC MODELING FOR TWEETS LOCATION

Short text messages from a sharing platform such as Twitter contain many errors due to the constraint of the tweets size. An elegant way to tackle these errors is to map tweets in a thematic space in order to abstract the document content.

To go beyond the LDA limit, the Author-topic (AT) model (Rosen-Zvi et al., 2004) was proposed. The AT model links both authors (here, the country) and documents content (words contained in a tweet). Next sections describe the AT model. An example of a tweet mapped into an AT model is presented in Figure 1, while Figure 2 represents the AT model into its plate notation. For each word w contained in a document d, an author a is uniformly chosen at random. Then, a topic z is chosen from a distribution over topics specific to that author, and the word is generated from the chosen topic.

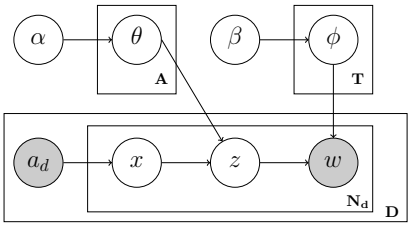

Figure 2: Generative model in plate notation of the Author-Topic (AT) model.

In our application, a document d is a short text message from the sharing platform Twitter and a country is considered as an author. Thus, each tweet d is composed with a set of words w and a country a. In this model, x indicates the country providing a given word, chosen from a_d. Each country is associated with a distribution over topics (θ), chosen from a symmetric Dirichlet prior ($\overrightarrow{\alpha}$), and a weighted mixture to select a topic z. A word is then generated according to the distribution ϕ corresponding to the topic z. This distribution ϕ is drawn from a Dirichlet ($\overrightarrow{\beta}$). The parameters ϕ and θ are estimated from a straightforward algorithm based on the Gibbs Sampling (Rosen-Zvi et al., 2004) such as the LDA hyper-parameters estimation method. Figure 1 shows the mapping process of an unseen tweet d from the validation set into an author topic space of size T. Each tweet d is consists of a words set w and a label (country) a considered as the author in the AT model. Thus, this model allows to code statistical dependencies between the tweet content (words w) and label (country a) through the distribution of the latent topics in the tweet.

The generative process we use corresponds to the hierarchical Bayesian model shown, using a plate notation, in the Figure 2 (a). Several techniques, such as Variational Methods (Blei et al., 2003), Expectation-propagation (Minka and Lafferty, 2002) or Gibbs Sampling (Griffiths and Steyvers, 2004), were proposed to estimate the parameters describing a hidden space. Gibbs Sampling is a special case of Markov-chain Monte Carlo (MCMC) (Geman and Geman, 1984). It gives a simple algorithm for approximate inference in high-dimensional models such as AT-model (Rosen-Zvi et al., 2004). This overcomes the difficulty to directly and exactly estimate parameters that maximize the likelihood of the whole data collection defined as:

$$P(W|\overrightarrow{\alpha},\overrightarrow{\beta}) = \prod_{w \in W} P(\overrightarrow{w}|\overrightarrow{\alpha},\overrightarrow{\beta}) \qquad (1)$$

for the whole data collection W knowing the Dirichlet parameters $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$.
To estimate LDA, the Gibbs Sampling was firstly reported in (Griffiths and Steyvers, 2004). A more comprehensive description of this method can be found in (**?**). One can refer to these papers for a better understanding of this sampling technique. This method is used both to estimate the LDA parameters and to infer an unseen dialogue with a hidden space of T topics. Gibbs Sampling allows us to estimate the AT model parameters, in order to represent an unseen tweet d with the r^{th} author topic space of size T, and to obtain a feature vector $V_d^{a_k^r} = P(a_k|d)$ of the topic representation an unseen tweet d with the r^{th} author topic space Δ_r^n of size

T. The k^{th} ($1 \leq k \leq A$) feature is:

$$\begin{aligned} V_d^{a_k^r} &= P(a_k^r|d) \\ &= P(a_k|z_r)P(z_r|d) \\ &= \sum_{j=1}^{T} P(a_k|z_{r,j})P(z_{r,j}|d) \\ &= \sum_{i=1}^{N_d}\sum_{j=1}^{T} P(a_k|z_{r,j})P(w_i|z_{r,j}) \\ &= \sum_{i=1}^{N_d}\sum_{j=1}^{T} \theta^r_{j,a_k}\phi^r_{j,i} \end{aligned} \quad (2)$$

where A is the number of countries (authors) in our case; $\theta^r_{j,a_k} = P(a_k|z_{r,j})$ is the probability of a country a_k to be generated by the topic $z_{r,j}$ ($1 \leq j \leq T$) in the r^{th} topic space of size T. $\phi^r_{j,i} = P(w_i|z_{r,j})$ is the probability of the word w_i (N_d is the vocabulary size of d) to be generated by the topic $z_{r,j}$.

3. EXPERIMENTAL PROTOCOL

We propose to evaluate the proposed approach in the application framework of a Twitter corpus. This corpus, presented in Table 1, is composed of a set of tweets from different countries. Its was constituted during the period were the event "Charlie" occurred, corresponding to about a week in January 2015. We developed a dedicated tool able to automatically capture online the tweets emitted all over the world. A second process permitted to filter the tweets, keeping only the ones that concern the event "Charlie" and that where located in space. Using these data, a classification approach based on Support Vector Machines (SVM) was performed to find out the most likely country of a given tweet. Next sections describe the data set, the author-topic model and the SVM classification method.

3.1 Tweets dataset

Concerning the event "Charlie", tweets are automatically labeled with one of the 16 countries presented in Table 1 which sums up the corpus obtained in the sample available from the Twitter servers. This data set is split in three parts depending on the tweets emission day δ in 2015:

- January $7^{th} \leq \delta \leq$ January 8^{th} (887 tweets),
- January $9^{th} \leq \delta \leq$ January 10^{th} (471 tweets),
- January $11^{th} \leq \delta \leq$ January 14^{th} (881 tweets),

We used $1,520$ tweets for the training phase of the AT models and 669 for the validation (testing) phase which corresponds to a corpus of $2,189$ tweets for the whole 16 countries (roughly 137 tweets for each country).

3.2 Author-Topic model

The number of topics contained in the AT model strongly influences the quality (called perplexity) of this model. Indeed, an AT model with only few topics will be more general than a one with a large number of classes (granularity of the model). For a sake comparison, a set of 100 AT models of size T was learnt ($5 \leq T \leq 105$).

Table 1: Number of tweets for each period of January 2015.

Country name	7^{th} **to** 8^{th}		9^{th} **to** 10^{th}		11^{th} **to** 14^{th}	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
France	287	124	171	74	259	111
United-Kingdom	90	39	58	25	99	43
United-States	77	34	58	26	110	48
Brazil	30	13	11	5	32	15
Italia	28	12	16	7	20	9
Spain	20	9	14	6	14	6
Turkey	14	7			13	6
Nederland	16	8			7	3
Canada	9	5			7	4
Belgium	9	5				
Mexico	8	4				
Colombia	9	5				
Philippines					9	4
Argentina					8	4
Germany	8	4				
India	9	4				
Total	614	273	328	143	578	253

3.3 SVM classification

As the classification of tweets requires a multi-class classifier, the SVM *one-against-one* method is chosen with a linear kernel. This method gives a better accuracy than the *one-against-rest* (Yuan et al., 2012). In this multi-class problem, A denotes the number of countries and $t_i, i = 1, \ldots, A$. A binary classifier is used with a linear kernel for every pair of distinct country. As a result, binary classifiers $A(A-1)/2$ are constructed all together. The binary classifier $C_{i,j}$ is trained from example data where t_i is a positive class and t_j a negative one ($i \neq j$).

For a vector representation of an unseen tweet d ($V_d^{a_k^r}$ for an AT representation), if $C_{i,j}$ means that d is in the country t_i, then the vote for the class t_i is incremented of one. Otherwise, the vote for the country t_j is increased by one. After the vote of all classifiers, the tweet d is assigned to the country having the highest number of votes.

4. RESULTS

A major event named "Je suis Charlie" happened in Paris in January, 7^{th} to 14^{th} 2015 and immediately appeared on Internet sharing platforms. Figure 3 shows the accuracy of this bursty event (percentage of countries found) obtained during the country location task presented above, for different time periods ($7^{th} \rightarrow 8^{th}$ January 2015 (a), $9^{th} \rightarrow 10^{th}$ January 2015 (b), $11^{th} \rightarrow 14^{th}$ January 2015 (c) and $7^{th} \rightarrow 14^{th}$ January 2015 (d).

The first remark regarding the curves in Figure 3, is that the higher the number of topics of the AT model, the better the location accuracy. Indeed, regardless the time period of tweet emissions, the best accuracy is reached with an AT model of size 96 or 98 topics. Contrariwise, the worst results are observed with AT models having a small number of topics (5).

The approach proposed to automatically locate a tweet obtains very promising results (more than 95% for the $9^{th} \rightarrow 10^{th}$ January 2015 time period as shown in Figure 3-(b), Figure 3-(d) presenting the results observed regardless the epoch ($7^{th} \rightarrow 14^{th}$ January 2015)). In a same manner, one can easily notice that the more precise the AT model (high number of topics), the higher the accuracy. A topic model with a thin granularity allows us to better characterize the meaning of a given message.

Finally, we can point out that the best result is reached during the $9^{th} \rightarrow 10^{th}$ January 2015 period presented in Figure 3-(b) with an accuracy of 95.7% which corresponds to the second attack in Paris.

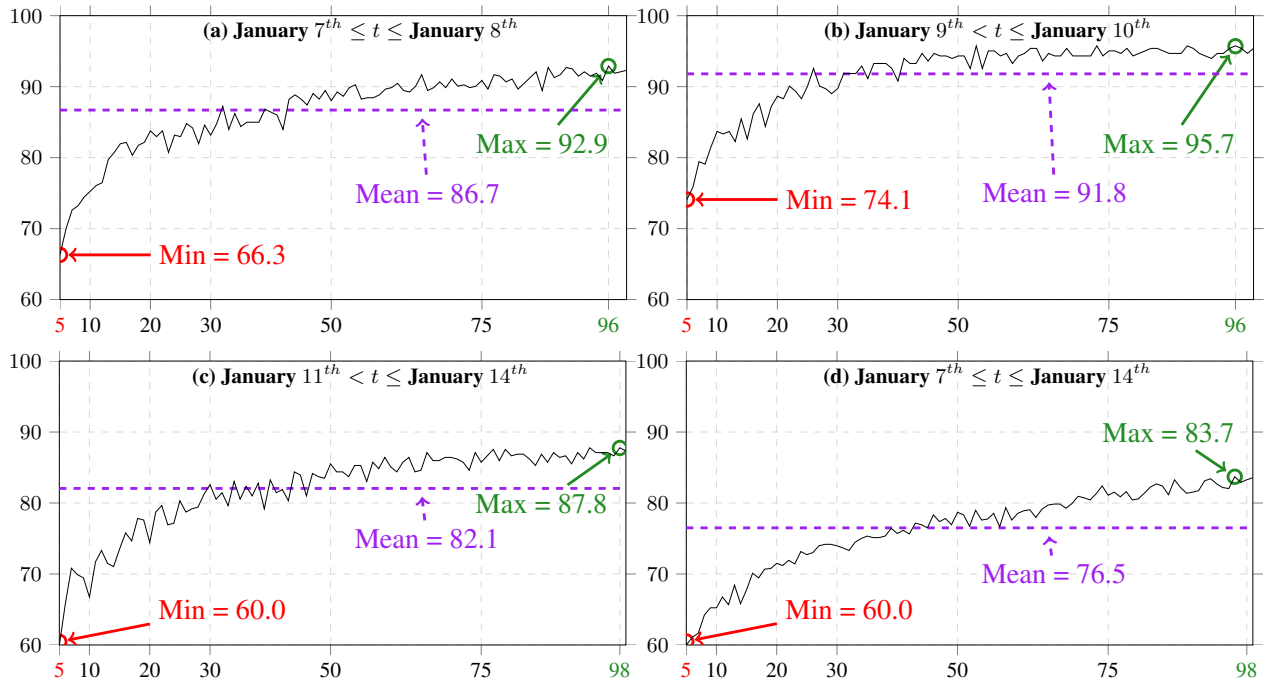

Figure 3: Country classification accuracies (%) using various author topic-based representations on the test sets with different epochs. X-axis represents the number n of classes contained into the topic space ($5 \leq n \leq 100$).

5. CONCLUSION

In this paper, we present an efficient way to deal with short text messages from Internet micro-blogging platforms which are highly error prone. The approach seeks to map a tweet into a high-level representation using the Author-topic (AT) model that takes into consideration all information contained in a tweet: the content itself (words), the label (country) and the relation between the distribution of words in the tweet and its location, considered as a latent relation. A high-level representation allows us to obtain very promising results during the identification of the country. Experiments conducted on a Twitter corpus showed the effectiveness of the proposed AT model with an accuracy reached of more than 95%. However, these results are based on the peculiar words of significantly different languages and it does not give any semantic information on the way the Charlie event was perceived by the population from these countries. This could be an additional interesting approach to develop, especially if it is linked to a map of tweets, to observe how such a worldwide event makes a buzz in space and time.

ACKNOWLEDGEMENTS

This work has been partially funded by the ContNomina project supported by the French National Research Agency (ANR) under contract ANR-12-BS02-0009.

REFERENCES

Baayen, A. R. R. H., 1998. Aviating among the hapax legomena: Morphological grammaticalisation in current british newspaper english. Explorations in corpus linguistics (23), pp. 181.

Bellegarda, J., 1997. A latent semantic analysis framework for large-span language modeling. In: Fifth European Conference on Speech Communication and Technology.

Bellegarda, J. R., 2000. Exploiting latent semantic information in statistical language modeling. Proceedings of the IEEE 88(8), pp. 1279–1296.

Blei, D., Ng, A. and Jordan, M., 2003. Latent dirichlet allocation. The Journal of Machine Learning Research 3, pp. 993–1022.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K. P., 2010. Measuring user influence in twitter: The million follower fallacy. In: Intern. Conference on Weblogs and Social Media (ICWSM).

Choudhury, M., Saraf, R., Jain, V., Sarkar, S. and Basu, A., 2007. Investigation and modeling of the structure of texting language. In: IJCAI, pp. 63–70.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American society for information science 41(6), pp. 391–407.

Geman, S. and Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence (6), pp. 721–741.

Griffiths, T. L. and Steyvers, M., 2004. Finding scientific topics. Proceedings of the National academy of Sciences of the United States of America 101(Suppl 1), pp. 5228–5235.

Hofmann, T., 1999. Probabilistic latent semantic analysis. In: Proc. of Uncertainty in Artificial Intelligence, UAI'99, Citeseer, p. 21.

Huang, J., Thornton, K. M. and Efthimiadis, E. N., 2010. Conversational tagging in twitter. In: Proceedings of the 21st ACM conference on Hypertext and hypermedia, ACM, pp. 173–178.

Kwak, H., Lee, C., Park, H. and Moon, S., 2010. What is Twitter, a social network or a news media? In: WWW, pp. 591–600.

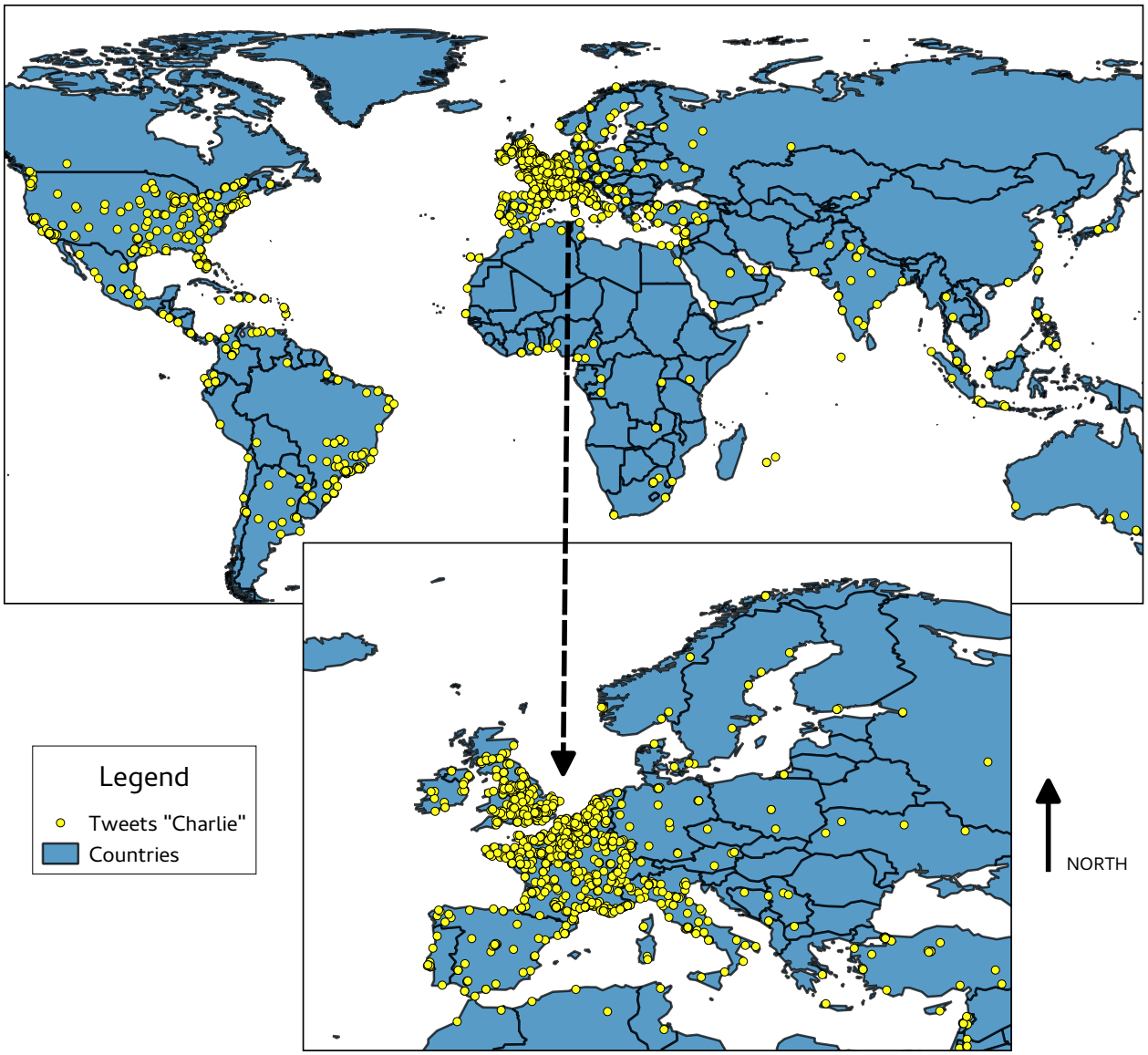

Figure 4: Locations of the tweets about the event ”Je Suis Charlie” (January, 7-14, 2015).

This contribution has been peer-reviewed.
Editors: C. Mallet, N. Paparoditis, and I. Dowman
doi:10.5194/isprsarchives-XL-3-W3-629-2015

Minka, T. and Lafferty, J., 2002. Expectation-propagation for the generative aspect model. In: Conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., pp. 352–359.

Morchid, M., Dufour, R. and Linarès, G., 2014a. A LDA-based topic classification approach from highly imperfect automatic transcriptions. In: LREC'14.

Morchid, M., Dufour, R., Bouallegue, M. and Linarès, G., 2014b. Author-topic based representation of call-center conversations. In: International Spoken Language Technology Workshop (SLT).

Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P., 2004. The author-topic model for authors and documents. In: UAI'04, pp. 487–494.

Salton, G., 1989. Automatic text processing: the transformation. Analysis and Retrieval of Information by Computer.

Salton, G. and Buckley, C., 1988. Term-weighting approaches in automatic text retrieval* 1. Information processing & management 24(5), pp. 513–523.

Suzuki, Y., Fukumoto, F. and Sekiguchi, Y., 1998. Keyword extraction using term-domain interdependence for dictation of radio news. In: 17th international conference on Computational linguistics, Vol. 2, ACL, pp. 1272–1276.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welpe, I. M., 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In: ICWSM, pp. 178–185.

Vapnik, V. and Lerner, A., 1963. Pattern recognition using generalized portrait method. Automation and Remote Control 24, pp. 774–780.

Yuan, G.-X., Ho, C.-H. and Lin, C.-J., 2012. Recent advances of large-scale linear classification. Proceedings of the IEEE 100(9), pp. 2584–2603.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. and Li, X., 2011. Comparing twitter and traditional media using topic models. In: Advances in Information Retrieval.