Using Retweet Information as a Feature to Classify Messages Contents

David Burth Kurka
(1) Imperial College London
(2) University of Campinas
d.kurka@imperial.ac.uk

Alan Godoy
(1) CPqD Foundation
(2) University of Campinas
godoy@dca.fee.unicamp.br

Fernando J. Von Zuben
University of Campinas
vonzuben@dca.fee.unicamp.br

ABSTRACT

We investigate the use of machine learning algorithms to classify the topic of messages published in Online Social Networks using as input solely user interaction data, instead of the actual message content. During a period of six months, we monitored and gathered data from users interacting with news messages on Twitter, creating thousands of information diffusion processes. The data set presented regular patterns on how messages were spread over the network by users, depending on its content, so we could build classifiers to predict the topic of a message using as input only the information of which users shared such message. Thus, we demonstrate the explanatory power of user behavior data on identifying content present in Social Networks, proposing techniques for topic classification that can be used to assist traditional content identification strategies (such as natural language or image processing) in challenging contexts, or be applied in scenarios with limited information access.

Keywords

Online Social Networks; Twitter; Topic Classification; Network Feature Extraction

1. INTRODUCTION

An important challenge when working with data from Online social networks (OSNs) is the fact that content is usually not structured, composed mostly by written text and images. When dealing with text, the task of automatically evaluating posts is hampered by the abundant use of slang, abbreviations, non-verbal information (as emoticons or emojis) and irony [1, 11, 20]. Beyond that, as OSN users speak to people that share their social context, users are often too laconic, suppressing information that is common sense to a social group but is not obvious to outsiders [21]. These issues are intensified in microblogging services once they restrict posts length – Twitter, for instance, allow only 140 characters in each message –, reducing the efficiency of traditional text mining techniques, as topic detection and sentiment

analysis. Therefore, being able to use extra information, either to avoid the complexities of natural language processing or to supplement data obtained through such classical approaches, can be very useful.

In this work we investigated to which extent it is possible to understand and predict aspects of processes happening on a social network – virtual or not – from its users behaviour. More specifically, we used machine learning algorithms to classify messages in OSNs according to their subject not using any information about their content, but only data about which user shared each message. Twitter, a popular microblogging service, was chosen as source of data, as it provides rich information of user-produced content, user profiles and social connections, available to download. The achievements of this work indicate that even using simple classifiers, as k-NN and logistic regression, it is possible to obtain information about the content of the message without inspecting it, knowing solely the users that shared such message.

2. BACKGROUND

As OSNs are able to capture social dynamics, they can be used to the study of human collective behaviours, providing material and insights to areas like psychology, sociology and even economy. Thus, the popularization of OSNs enabled the development of a *computational social science* [13]. The computational analysis of social data brings new observations, methodologies and innovative results to fields that have been studied for many years.

A very common effect that has been observed and studied in this topic are information cascades, which are the phenomenon of creation, replication and transformation of content by OSN users [4, 8]. Twitter has a simple mechanism of creating content cascades, as users can repost messages posted by other users, in the so called *retweets* [15]. Cascades have been studied by many researchers, which explored themes as their characterization [4], how their dynamics are affected by network structures [7] and how cascades affect large political events [9].

Classifying the content that is being spread in OSNs, the focus of the present work, has also been an important topic of research. Knowing the subjects that are being diffused over the network can be useful, for instance, to identify events happening in real time in the world [3, 10] or to understand communities behaviour [14].

The literature present two main approaches to automatic content classification.

The first uses machine learning to analyse directly the content being shared. Therefore, in case of Twitter, techniques of natural language processing (NLP) are used to classify a corpora of tweets [12, 23]. Note that although the machine learning techniques used to classify OSN text are also applied to other generic NLP tasks, OSN data has some peculiarities that can be useful in the analysis. For example, Hu *et al.* [11] investigated using emoticons as features for sentiment classification and Reyes *et al.* [21] the role of *hashtags* to classify irony in tweets. However, OSN content can present serious and blocking challenges to text mining, as it often contains informal texts (such as abbreviations and slangs) and short size in terms of number of words [6].

The second approach, in turn, uses metadata and features of the network to assist the classification. More than just the content produced by users, OSN services allow the collection of information about its users (such as profile, publication record) and about users' connections. This type of information was used, for example, by Cataldi *et al.* [5], that considered the 'authority' of users together with the message content, to classify and identify topics. Suh *et al.* [24] also investigate other properties from users such as number of connections, and number of messages to assist in the characterization of potential popularity of messages. Baba *et al.* [2] create groups of similar messages according to users behaviors, using a community detection algorithm to separate messages in clusters.

The work presented here goes in the same direction of this second approach, using users' sharing behavior to classify content, instead of exploring NLP techniques. The methods presented therefore can be used independently of language and applied even in contexts where the full text is not disclosed.

3. DATASET DESCRIPTION

In this work we evaluated whether the structure of a information diffusion process contains relevant information about the content of messages that are being exchanged in the network. We did this by using the set of users that shared a specific content as a feature vector for a supervised classifier, trained to infer the subject of such content. Although other works also explore the same vein of research (as showed above), we are unaware of works that similarly explore message topic classification using solely sharing data.

3.1 Data Acquisition

We used Twitter as source of data for the experiments, due to the simplicity of its basic content – the *tweet*, a short text message –,the amount of public content available through APIs[1] and its suitability as research object, as demonstrated in diverse studies [13].

One restriction imposed to access Twitter's data, however, is a significantly low limit of requests for messages already published, making impossible the posterior analysis of known diffusions. The alternative is to use the Stream API, that allows the extraction of a vast number of messages that are being transmitted in real-time on the service. Although the amount of data extracted is not an issue when using the Stream API, the necessity of using real-time data imposes restrictions on the cascades that can be analyzed. The main

challenge is therefore to determine *a priori* where a diffusion process will begin, in order to collect its information.

The solution found was to focus the analysis on cascades created and started by popular users on the network, as they predictably obtain great repercussion on most content posted by them, due to their big number of followers. Therefore, by filtering messages authored by famous users (such as actors, personalities, public figures or news agencies) it was possible to witness, in real-time, several processes of information diffusion. From the observation of the occurrence of such processes, it is possible to map the users that joined the diffusion, their relationship network and the time instant in which the messages were sent.

3.2 Data Description

Using the methodology described above, the Twitter account of the largest Brazilian newspaper, Folha de São Paulo[2] – which had over three million followers when the collection was started –, was chosen as content source and monitored. From March 19, 2014 to September 21, 2014, all messages (tweets) posted by the source, as well as any share of this content by other users (retweets) were collected.

It was possible to track, from the data collected, a large amount of information diffusion processes triggered by the observed source. During the observed period, 13463 distinct and original messages posted by the source account were collected. From this set, a series of filters was applied, forming a more appropriate data set for the work, as described below:

- Filter 1 - In a first step, only messages which had received at least 20 retweets were selected, resulting in a group of 4671 distinct messages. This increased the significance of the events observed and decreased the sparcity of our data;

- Filter 2 - Messages not belonging to one of the six main categories ("everyday news", "sports", "world", "politics", "entertainment" and "market") were removed from the above set, leaving 3185 messages;

- Filter 3 - Automated scripts (bots) were removed. A user was considered as a bot, if it retweeted more than 60% of the messages collected in the database. Only one user was identified as a bot (which was confirmed through a close inspection of the user's profile) and therefore removed from the database and from the retweets count;

- Filter 4 - As in 2014 Brazil held the FIFA World Cup and presidential elections, there were a high number of messages in categories "politics" and "sports". In order to balance the proportion of messages in each topic, a maximum limit of 450 messages was set for each category. The general proportion of each topic can be seen in Figure 1.

The filtered data resulted in a collection of 2444 messages (M) retweeted by 44627 distinct users (U). The data was organized in a binary matrix of retweets (T) of dimensions $M \times U$, with the element $T_{i,j}$ made equal to 1 in case the message i was *retweeted* by user j, and 0 otherwise (see Figure ??). Thus, each row of T (denoted by T_i) corresponds to

the sharing behaviour relative to a tweet published by Folha de São Paulo and can also be interpreted as a "signature" (feature vector) of a message for the algorithms used in the classification step.

Table 1 presents a comprehensive characterization of the collected data and the network formed between users (excluding the source user, Folha de São Paulo). It is worth pointing, also, that most users do not participate actively on the most diffusions, implying in high diversity of users participating in the processes, but low recurrence: during the observation period, each user retweeted in average about two messages from the source.

3.3 Topic Classification

The definition of a ground-truth to be used to train and evaluate the classifier was a major challenge in this experiment. Even though the data set used was not as large, manual classification of the tweets proved too costly, specially considering that each item should be classified by multiple individuals in order to obtain consistent labels.

Considering that practically all the messages published by Folha de São Paulo account are headlines of news followed by a link to the newspaper's website[3] with the news' full content, a second possibility would be use classical techniques for topic detection – as LSA or LDA – on the full text of the news to define the labels for each message. This approach, however, would present the drawback of limiting the results achievable to the quality of the topic detection.

Therefore, we decided to use as labels the division of news in thematic sections made by the newspaper editorial staff. From the URL present in the messages, it was possible to attribute a class to each tweet using an automated script. This procedure was carried out to all tweets and six predominant topics were verified, among them: "everyday news", "sports", "world", "politics", "entertainment" and "market" (in Portuguese: "cotidiano", "esporte", "mundo", "política", "entretenimento" and "mercado", respectively). Those categories were used on the filtering process (as described above) and were considered as classes on supervised machine learning algorithms, described in the Results sections. Figure 1 display the proportion of each category in the database, after the filter process.

This procedure configures a practical and reliable way to classify a large number of tweets. However, as the definition of each topic's class is made by the newspaper's staff, with a specific purpose of organization and separation, other schemes of categorization can be proposed and advocated. Also, the fact that each message is attributed to only one class can be seen as an issue when there are cases where news could belong to overlapping topics (e.g.: news reporting a protest before a soccer match could belong simultaneously to "politics" and "sports" classes). This issue affects the presented results and is further discussed in the conclusion. Another consequence of this approach is that it does not allow the analysis of diffusion processes initiated by multiple sources, as each newspaper has its own taxonomy and separation scheme – reason why we used only Folha de São Paulo's account as source.
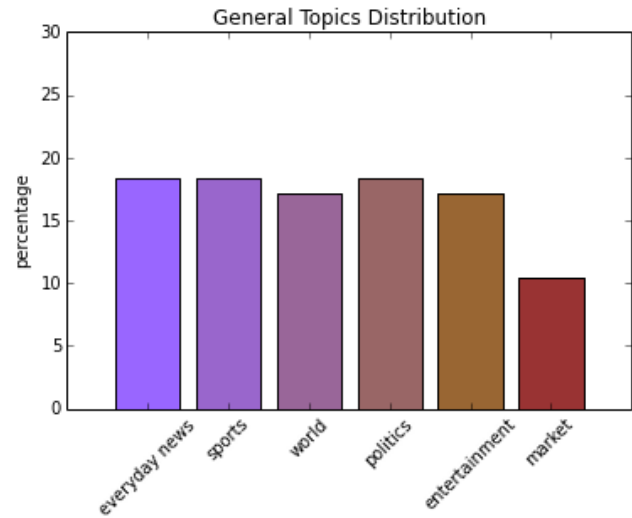

Figure 1: Topics distribution in database after filters.

4. EXPLORATION AND RESULTS

This section details the experiments conducted where classification algorithms were applied on the database.

Considering that the focus of this work is to investigate if sharing behavior contains sufficient information to predict the topic of a message and not to produce the best classifier, we opted to use k-NN and logistic regression, classical machine learning algorithms that are simple and have few parameters to optimize. Considering the small data set used, the choice of less powerful classifiers also aims at reducing the possibility of overfitting.

The classification task consisted in determining the category of a message i from its binary vector of retweets T_i. As every message in T was classified using the automatic procedure previously described, it was possible to perform supervised learning, where a subset of T was used to train models and the remaining subset to test the performance of the models.

4.1 k-NN

The first classification strategy applied was to use the well-known *k-nearest neighbours* (k-NN) algorithm, that is widely used for classification and regression. In k-NN, for each message i to be classified, we search the training set (for which classification is known) for the k messages that are closest to i in the feature space. The new message's class is, thus, defined to be the most common class present among its neighbors.

As previously stated, the behaviour vector T_i is used as the feature vector that represents the message i. Diverse metrics were considered to compute the similarity between two messages, where the Jaccard distance[4] was the one that produced the best results. Different values of k (neighbours) were tested and the performance achieved is registered in Figure 2. The results presented are the classifier's accuracy for a test set of 300 random tweets selected from the

---

[3]http://www.folha.uol.com.br/

[4]Metric used to compare Boolean vectors, defined by: $d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$, where A and B are Boolean vectors.

Table 1: Characterization of the collected data

|	
Number of users	44,627
Connections between users	686,326
Average degree	30.76 ($\langle k_{in} \rangle = \langle k_{out} \rangle = 15.21$)
Clustering coefficient	3.82% (C for a randomized network is 0.07%)
User with highest in-degree	@UOL (Online service and news provider) – 5793 followers
Diameter	16
Total number of messages	2,444
Total number of retweets	111,402 (2.49 per user / 45.58 per message)
Most popular message	739 retweets
Density of retweets	0.10%
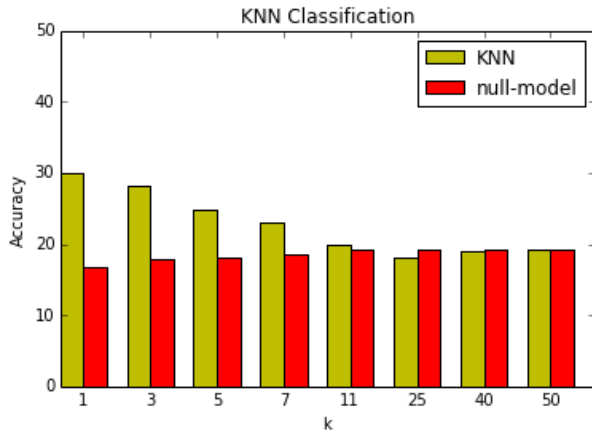
Figure 2: Classification results of the k-NN algorithm and of a random classifier (null model).

database. The obtained accuracy is the average of 30 different executions, using different samples.

It is noticeable that, for lower values of k ($k \in 1, 3$), the k-NN classifier shows results way above random, with almost 30% of accuracy, showing a relationship between type of message (topic) and sharing behaviour. For higher values of k, however, the results tend to be equal to the null model. A possible explanation to this is that each class is formed by small subclasses, spread along the space. Therefore, the first few closest neighbors tend to belong to the same class, but when k increases, elements of other classes appear as neighbors, undermining the classification performance.

Although there are six classes being evaluated, the random classifier is able to predict correctly slightly more than 17% (1/6), having an accuracy value around 20%. This happens due to the fact that the topics are not equally distributed in terms of number of messages, as Figure 1 shows.

4.2 Logistic Regression

Subsequently, a more elaborate classifier, the *logistic regression* was used. Although being more tunable than a k-NN, logistic regression is a linear model and, thus, it is still a quite simple model that does not take into account eventual nonlinear interactions between the input variables.

For the training process, 80% of the data set was used for training, producing an input feature vector of dimension

44627×1955, and the remaining 20% were separated for testing. The logistic regression implementation used was the one available on Python's library scikit-learn [18], using liblinear library to solve the regression. In the presented results, we use L2 regularization with parameter C (inverse regularization strength) set to 1.

From 10 distinct executions of the algorithm, using different partitions between training and test samples, the accuracy of $\mathbf{48.75 \pm 1.74}$% was achieved. This result shows that the classifier was able to identify patterns on the training data set, making the prediction task something feasible, for new samples presented to the model.

Table 2 depicts the average confusion matrix of the classification, showing the algorithm's performance for each class. "Market", a class with few samples, had the lowest precision and recall. The most precise classification involved messages of "sports" and "politics", with more than 60% recall. Messages from "everyday" were many times classified as "politics", what is also expected, as this category is usually related to political issues, with news about protests and low-quality public services. Some "politics" messages were mistakenly classified as "sports", what is not very surprising, given that the 2014 FIFA World Cup, held in Brazil, involved massive investments by Brazilian federal and state governments.

In Table 3 we show randomly selected tweets from "politics" that were wrongly classified as "sports" by the logistic regression. It is possible to see that three of the eight tweets were linked to the World Cup, while one of the remaining messages (about a viaduct that collapsed) referred to an infrastructure project built for the World Cup. So, even indirectly, those tweets are also related to "sports".

In Table 4, among tweets from "everyday news" that were classified as "politics", almost all were related to political issues: public services, new laws or protests. It is worth noticing that there is even a tweet about the Brazilian presidential election, wrongly classified by the newspaper editors as "everyday news", that our technique rightly identified as "politics".

It also must be noted that the algorithm reached a rate of 100% of accuracy on the training stage, that was not achieved within the test data set. Attempts to increase the model's generalization capability were conducted, by increasing the regularization. However, although the accuracy of the training stage decreased, the test accuracy rate did

Table 2: Confusion matrix for classification task using logistic regression.
The results reported refer to the execution with $\lfloor N/2 \rfloor$-th lowest classification error among all N repetitions.
Classes: (a) everyday news; (b) sports; (c) world; (d) politics; (e) entertainment; (f) market

		Observed							
		(a)	(b)	(c)	(d)	(e)	(f)	Total	Recall
Expected	(a)	**35**	9	12	21	9	4	90	38.9%
	(b)	9	**62**	12	8	6	2	99	62.6%
	(c)	14	7	**42**	10	6	1	80	52.5%
	(d)	11	10	5	**54**	1	4	85	63.5%
	(e)	10	15	13	3	**39**	2	82	47.6%
	(f)	13	8	12	11	3	**6**	53	11.3%
	Total	92	111	96	107	64	19		
	Precision	38.0%	55.9%	43.8%	50.5%	60.9	31.6%		

Table 3: Randomly selected misclassified tweets. All tweets were translated from Portuguese and notes were added between brackets when necessary.

Expected: politics, found: sports
Palácio do Planalto's [seat of Brazilian government] computer was used to edit Temer [vice-president] and Ideli [senator] pages. `http://t.co/wuqblBe2lD`
Marco Feliciano [congressman] says he sees contradiction in criticism from PT [political party] to Marina Silva [candidate to presidency]. `http://t.co/IZOEGLnAZL`
#FolhaninTheWorldCup protesters against the World Cup invade federal government seminar in Recife. `http://t.co/iP3CLCC7Nn`
#FolhaElections Eduardo Jorge [candidate to presidency] speaks live to 'TV Folha' at 4pm. `http://t.co/GHB3VbYdgW`
Viaduct collapses and leaves at least one dead in Belo Horizonte, in addition to several injured people. `http://t.co/BMTZuuCUN2`
Parents of player Khedira are robbed in Recife, says German newspaper. `http://t.co/52FnDzKAW2`
#FolhaInTheWorldCup Protester plays with ball during police siege to a protest in Porto Alegre. `http://t.co/6W4U4tqDbY`
Lula [former Brazilian president] talks about increasing political action if Dilma [current Brazilian president] is reelected. `http://t.co/xSXFIV6uHl`

Table 4: Randomly selected misclassified tweets. All tweets were translated from Portuguese and notes were added between brackets when necessary.

Expected: everyday news, found: politics
Senator Aloysio Nunes does not dismiss the idea of being vice candidate of Aécio Neves [pre candidate to presidency]. `http://t.co/qyCQMbhrKG`
Subway workers reject counter offer and should block trains this Thursday in SP [Brazilian state]. `http://t.co/emu63AYHfx`
After fire in CTG [community center], gay wedding is celebrated at RS [Brazilian state] court: `http://t.co/hWIovrnDoW`
Alckmin [São Paulo state governor] will sanction this week new law to forbid masks in protests. `http://t.co/EkMmwM7jHk`
Franca [city] assembles a catalog of museum collection to be available on the Internet. `http://t.co/2gg8bFTLkd`
Doctor 'expelled' from SUS [public health system] writes in a book his daily routine in the system: `http://t.co/IOpMk3xxbJ`
House of Representatives approves project that equalize pharmacies and health facilities. `http://t.co/ZP3bxhDCMk`
Artifacts found with activist arrested in protest against the World Cup are not explosives, says report. `http://t.co/xEKqz4mwhA`

not increased (apparently, ruling out the hypothesis of over-fitting occurrence).

5. DISCUSSION

The results presented give insights to whether it is possible to obtain information about the subject of a message only knowing the users that shared it. It is noticeable that, even with a very sparse feature space (99.98% of the elements of the matrix T are zeros!) and with relatively few messages from which to learn (there was less than 2000 messages in each training set), it is possible to build classifiers that predict a message's class over six possibilities with an accuracy rate of near 50%. Also, the comparison to random models

shows that the positive results obtained are not simply fruit of some classes being more common than others.

The use of classes manually-defined by the editors of the newspaper was very practical and avoided limitations from automatic topic classification. However, it brings some limitations, mainly due to the use of only a single class per tweet. Most of the classification errors happened in categories that had significant overlap with other categories, as evidenced in Table 2. Further work may explore such issue, by using classification schemes that allow multiple classes for each message.

Furthermore, as we used the labels defined by the creators of the messages analyzed as ground-truth, we focused on cascades initiated on a single source, avoiding the effects of different classification schemes and criteria that would

arise if we used multiple sources. We point that this option, however, may limit the generality of the results here presented.

Considering the simplicity of the classifiers used in this work, it is possible that even better results may be achieved by using models that can analyze the patterns of interaction between input variables and that can treat more adequately missing values, as the content of a message is not the only factor determining if a user will share it or not (e.g.: one of the most common reasons for a user not to retweet a message is that he/she was not online when the message was first published).

The most remarkable fact, however, is that the algorithms presented are able to classify messages without using any direct information of the message's text. These techniques can be even more promising if combined to traditional topic detection algorithms that use textual features (natural language processing methods). It is expected that the classifiers based on network features are capable of detecting features not present on the textual content, thus enhancing the traditional classifiers. Also, this kind of technique can be applied on the development of minimally invasive classifiers, able to organize even encrypted data.

Additionally, the results obtained lead to matters of privacy, as it reveals aspects of the messages content that can be drawn solely from OSN meta-data. In the same manner that this information can be helpful to recommendation systems or text classifiers, it can also configure as a threat to individual privacy and should be cared for.

Finally, beyond the machine learning possibilities it is intriguing to think about the analogy inspired by network science between brain and society – networks that share many features, as topological [25] and dynamic aspects [16]. As many studies indicate, in animal brains some regions fire in response to specific sensory input [22]. This response can be so particular that, apparently, some neurons only fire in response to images of specific celebrities [19]. In a broad sense, this is similar to the observed phenomena on which specific individuals "fire" in response to specific themes in a social network. If we lead these analogies one step further – to a functional similarity –, it is exciting to wonder whether someday we will be able to use information gathered from online social networks in order to see what is happening inside our "collective brain", similarly to how current research shows that it is possible to reconstruct images seen by individuals only from neural activation images obtained through an fMRI [17].

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of Twitter data. *LSM '11 Proceedings of the Workshop on Languages in Social Media*, pages 30–38, jun 2011.

[2] S. Baba, F. Toriumi, T. Sakaki, K. Shinoda, S. Kurihara, K. Kazama, and I. Noda. Classification Method for Shared Information on Twitter Without Text Data. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*, pages 1173–1178, New York, New York, USA, 2015. ACM Press.

[3] H. Becker, M. Naaman, and L. Gravano. Beyond Trending Topics: Real-World Event Identification on Twitter. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 1–17, 2011.

[4] J. Borge-Holthoefer, R. a. Baños, S. González-Bailón, and Y. Moreno. Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks*, 1(1):3–24, apr 2013.

[5] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining - MDMKDD '10*, pages 1–10, New York, New York, USA, 2010. ACM Press.

[6] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report*, page 12, feb 2007.

[7] S. Goel, A. Anderson, J. Hofman, and D. Watts. The structural virality of online diffusion. *Preprint*, 2013.

[8] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce - EC '12*, volume 1, page 623, New York, New York, USA, 2012. ACM Press.

[9] S. Gonzalez-Bailon, J. Borge-Holthoefer, and Y. Moreno. Broadcasters and Hidden Influentials in Online Protest Diffusion. *American Behavioral Scientist*, 57(7):943–965, mar 2013.

[10] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. K. L. Ma. Breaking news on twitter. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, CHI '12, page 2751, New York, New York, USA, 2012. ACM, ACM Press.

[11] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised Sentiment Analysis with Emotional Signals. In *International Conference on World Wide Web*, pages 607–617, Rio de Janeiro, Brazil, may 2013. International World Wide Web Conferences Steering Committee.

[12] S. Kataria and A. Agarwal. Supervised Topic Models for Microblog Classification. In *2015 IEEE International Conference on Data Mining*, pages 793–798. IEEE, nov 2015.

[13] D. Kurka, A. Godoy, and F. Von Zuben. Online social network analysis: A survey of research applications in computer science. *arXiv:0707.3168 [cs.SI]*, 2015.

[14] T. Lansdall-Welfare, V. Lampos, and N. Cristianini. Effects of the recession on public mood in the UK. In *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, page 1221, New York, New York, USA, 2012. ACM Press.

[15] K. Lerman and R. Ghosh. Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 90–97, 2010.

[16] I. N. Lymperopoulos and G. D. Ioannou. Online social contagion modeling through the dynamics of integrate-and-fire neurons. *Inf. Sci.*, 320(C):26–61, Nov. 2015.

[17] S. Nishimoto, A. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.

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

[19] R. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

[20] D. Ramage, S. Dumais, and D. Liebling. Characterizing Microblogs with Topic Models. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 1–8, 2010.

[21] A. Reyes, P. Rosso, and T. Veale. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, 47(1):239–268, jul 2012.

[22] S. Seung. *Connectome: How the brain's wiring makes us who we are*. A Mariner Book. Houghton Mifflin Harcourt, 2012.

[23] D. a. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*, pages 355–358, New York, New York, USA, mar 2011. ACM Press.

[24] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *2010 IEEE Second International Conference on Social Computing*, pages 177–184. IEEE, aug 2010.

[25] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.