Tracking Dengue Epidemics using Twitter Content Classification and Topic Modelling

Paolo Missier[1], Alexander Romanovsky[1], Tudor Miu[1], Atinder Pal[1], Michael Daniilakis[1], Alessandro Garcia[2], Diego Cedrim[2], and Leonardo da Silva Sousa[2]

[1] School of Computing Science, Newcastle University, UK
[2] PUC-Rio, Rio de Janeiro, Brazil

Abstract. Detecting and preventing outbreaks of mosquito-borne diseases such as Dengue and Zika in Brasil and other tropical regions has long been a priority for governments in affected areas. Streaming social media content, such as Twitter, is increasingly being used for health vigilance applications such as flu detection. However, previous work has not addressed the complexity of drastic seasonal changes on Twitter content across multiple epidemic outbreaks. In order to address this gap, this paper contrasts two complementary approaches to detecting Twitter content that is relevant for Dengue outbreak detection, namely supervised classification and unsupervised clustering using topic modelling. Each approach has benefits and shortcomings. Our classifier achieves a prediction accuracy of about 80% based on a small training set of about 1,000 instances, but the need for manual annotation makes it hard to track seasonal changes in the nature of the epidemics, such as the emergence of new types of virus in certain geographical locations. In contrast, LDA-based topic modelling scales well, generating cohesive and well-separated clusters from larger samples. While clusters can be easily re-generated following changes in epidemics, however, this approach makes it hard to clearly segregate relevant tweets into well-defined clusters.
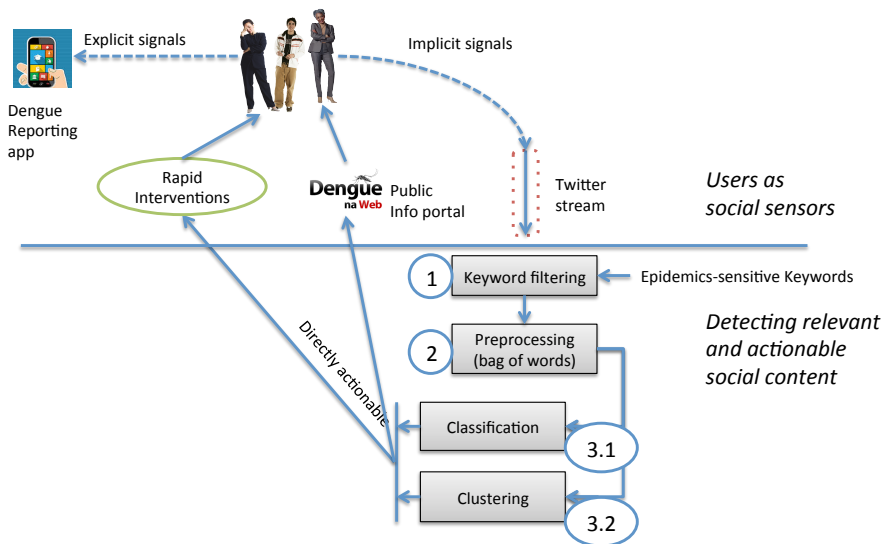
1 Introduction

Mosquito-borne disease epidemics are increasingly becoming more frequent and diverse around the globe and it is likely that this is only the early stage of epidemic waves that will continue for several decades. Rapidly spreading diseases to combat nowadays are those transmitted by the *Aedes* mosquitoes [CDC15], which carry not only *Dengue* virus, but also *Chikungunya* and *Zika* viruses [CDC15], which are responsible for thousands of deaths every year. Therefore, improved surveillance through rapid response measures against Aedes-borne diseases is a long-standing tenet to various health systems around the world. They are urgently required to mitigating the already heavy burden on those health systems and limiting further spread of mosquito-borne diseases within geographical locations, such as in Brazil. Control of Aedes-borne disease requires the vector control by identifying and reducing breeding sites.

Our approach to addressing this problem involves the automatic detection of relevant content in Twitter, in order to determine its relevance as actionable information. Paraphrasing [SOM10], we note that social media users are increasingly viewed as *informative social sensors*, who spontaneously communicate valuable information, which

in this case may help in detecting the location and extent of mosquito foci. However, as the signal produced by these sensors is very noisy, our realistic goal is to automatically categorise Twitter messages into a few classes, segregating recognisable highly informative from less informative and noisy content.

Previous work, e.g. [GVM+11,LC10,ALG+11], has identified the potential of social media channels, such as Twitter, on offering continuous source of epidemic information, arming public health systems with the ability to perform real-time surveillance. However, previously proposed approaches are often limited or insufficient for rapid combat of epidemic waves for several reasons. Firstly, previous work has mainly explored the use of social media channels to predict Dengue cases and outbreak patterns by exploring disease-related posts from previous outbreaks. However, the combination of socio-economic, environmental and ecological factors dramatically changes the characteristics governing each epidemic wave. As a consequence, exploring disease-related posts from previous outbreaks tends to be ineffective to identify breeding sites in the outset of each outbreak. Secondly, previous work is not aimed at identifying map breeding sites of the mosquito within a region.

The role of Twitter content relevance detection is depicted in Fig. 1. Social sensors, the people in the upper half of the figure, contribute information either implicitly, i.e., by spontaneously carrying out public conversations on social media channels, or explicitly, i.e., by interacting with dedicate public Web portals and mobile apps. As an example, our group in Brazil has been developing both such a Dengue mapping portal, and a mobile app that members of the public may use to report cases of Dengue in their local areas [PR15].

Fig. 1. Role of automated Twitter relevance detection for health vigilance against Dengue

As shown in the figure, we monitor the Twitter feed, pre-select tweets according to a broad description of the Dengue topics using keywords, then classify the selected tweets, aiming to segregate relevant signal from the noise. We distinguish between relevant signal that is *directly* and *indirectly* actionable. Directly actionable tweets, which we classify as *mosquito focus*, are those that contain sufficient information regarding a breeding site (including geo-location), to inform immediate interventions by the health authorities. For instance:

> @Ligue1746 Atenção! Foco no mosquito da dengue. Av Sta Cruz, Bangu. Em frente ao hospital São Lourenço! (*@Ligue1746 Attention! Mosquito focus found in Santa Cruz avenue, Bangu. In front of the So Loureno hospital!*)

These posts are relatively scarce within the overall stream, however, accounting for about 16% of the ground truth class assignments. Indirectly actionable tweets carry more generic information about members of the public complaining about being affected by Dengue (the *Sickness* class), or *News* about the current Dengue epidemics. For example:

> Eu To com dengue (*I have dengue fever*)
> ES tem mais de 21 mil casos de dengue em 2015 (*ES has more than 21 thousands cases of dengue in 2015*)

The rest of the tweets are all considered noise. In particular, these include messages where people joke about Dengue in a sarcastic tone, which is commonly used in online conversation in Brazil, for example:

> Meu WhatsApp ta tão parado que vai criar mosquito da dengue (*My WhatsAp is so still that it'll create dengue mosquito*)

In this paper we report our experiments on automatically classifying directly and indirectly actionable tweets. In the Figure, the classifier plays the role of a filter to limit the amount of noise on the pages displayed on our Web portal.

One problem faced in our classification scenario is that *epidemic waves* differ from season to season. For instance, new symptoms caused by the *Zika* virus have been observed in the epidemic wave, which started in October 2015. Such types of epidemic changes drastically change the nature of Twitter content, requiring different keyword settings and filtering from the Twitter feed, in order to accurately track an epidemic. Examples of keywords for tracking different virus and new emerging symptoms include *Dengue*, *Chikunguya*, and, more recently, *Zika*. Simply taking the union of all three would just add to the noise. What is required instead is the ability to rapidly reconfigure the classifier following a drift in topic. This flexibility requirement naturally suggests an unsupervised approach to learning the classifier. At the same time, a supervised classifer that is trained using manually labelled content is likely to be more accurate.

1.1 Contributions

In this work we explore the trade-offs between accuracy and flexibility, by comparing and contrasting a supervised classifier learning approach (3.1 in Fig. 1) with an unsupervised content clustering, using Topic Models (3.2) and specifically on LDA [BNJ03],

a popular algorithm that has been previously shown to apply well to clustering Twitter data [RDL10,REC12]. We expect supervised classification to provide good accuracy, as well as give an obvious way to select actionable content from the most informative classes (*mosquito focus*, *sickness*, and *News* in this order). On the other hand, this model suffers from known limitations in the size of the training set, which may lead to disappointing performance on content in the wild, and it is expensive to re-train following changes in the filtering keywords.

In contrast, topic modelling is a form of semantic clustering where a clustering scheme can be easily periodically re-generated from large samples. While the clear characterisation of clusters using ranked lists of terms from the content's vocabulary (topics) makes this a popular approach, a topic may include heterogeneous content that cuts across expert-defined classes, such as those above, making it harder to associate them with a clear focus. This problem is particularly acute in our setting, where we already have a topic defined (through keywords), and we are essentially asking LDA to further refine it in terms of well-separated sub-topics.

We assessed the potential of our approach on large cities of Brazil, such as Rio de Janeiro, by analyzing two cycles of Aedes-related epidemic waves. Our specific contributions in this paper are: (i) a pipeline that implements both methods, including a dedicated pre-processing phase that accounts for idiosincratic use of the Brazilian Portuguese language in tweets, and (ii) an experimental evaluation of their effectiveness. The supervised classifier is currently in operation as part of the experimental Dengue Web portal developed at PUC-Rio [PR15].

1.2   Related work

This paper makes original contributions to an already existing landscape of research on monitoring social media for health vigilance purposes. Similarly to our work, Twitter data is used by [GVM[+]11] to track the Twitter stream and filter relevant signals from it. Because they only use supervised classification for content filtering, their approach is limited by the amount of labels made available by expert annotators. Moreover, this limitation does enable to easily reveal new information in the outset of each epidemic wave. In our work, we use not only supervised classification, but also unsupervised clustering as means of identifying relevant social signals. Finally, we contrast the results from both methods in order to: (i) reflect on the different use cases the methods require (i.e. in terms of annotation effort), and (ii) observe how unsupervised classification helps to better achieve the purpose of revealing new information in each epidemic wave.

[ALG[+]11] and [LC10] show that the frequency of tweets containing simple search keywords can be a good indicator of a trend for a flu epidemic. The authors show that there is a strong correlation between the number of medically registered visits to a GP concerning flu and the number of tweets mentioning flu. This approach to tracking epidemics is complementary to ours because, while the previously mentioned authors measure tweet activity on an entire corpus of tweets, we use machine learning to further discover sub-signals in the corpus in specific epidemic waves. Our approach enables one to further measure and study tweet activity within relevant sub-signals.

Similar methods of monitoring Twitter data have been applied for general event detection, as done, for example, by [CW14] or [BNG11]. However, obtaining ground truth

is recognised to be a serious bottleneck in a supervised learning pipeline and efforts to reduce the annotation effort have been attempted. For instance, [GBH09] automatically identify ground truth from emoticons for sentiment classification. However, as previously mentioned, even if ground truths are somehow identified, the use of supervised learning may not suffice to cope with tracking the changing characteristics of different epidemic waves.

2   Twitter content acquisition and processing

Our experimental dataset consists of three sets of Twitter content, harvested over two periods of time, during the first and second semester of 2015. These periods corresponded to two cycles of epidemic waves. The first two sets, of about 1,000 and 1,600 instances, respectively, were manually annotated by our group at PUC-Rio, which also included the participation of a medical doctor and an epidemiologist. They were used in supervised classification as our training and test set (using standard k-fold validation), and for further testing (no training), respectively. A larger third set of about 100,000 tweets was used for topic modelling.

A technique similar to that described in [NGS$^+$09] was used to determine a set of filtering keywords for harvesting the tweets. Namely, we started with the single #dengue hashtag "seed" for an initial collection. Upon manual inspection of about 250 initial tweets, our local experts then extended the set to include the most relevant hashtags. These hashtags were those that all local experts agreed to be relevant after discussion amongst them. The final search set contains the following elements (including their common minor variations): { #Dengue, #suspeita, #Aedes, #Epidemia, #aegypti, #foco, #governo, #cuidado, #febreChikungunya, #morte, #parado, #todoscontradengue, #aedesaegypti }.[3]

Content pre-processing includes a series of normalisation steps, followed by POS tagging and lemmatisation.[4] We normalised the content by removing 38 kinds of "twitter lingo" abbreviations, some of which are regional to Brazil ("abs" for "abraço", "blz" for "beleza", etc.), as well as all emoticons and non-verbal forms of expressions. While those are crucial to understanding the *sentiment* expressed in a tweet, we found that they are not good class predictors, including the *Jokes* class. We also replaced links, images, numbers, and idiomatic expressions using conventional terms (*url*, *image*, *funny*,...).

3   Supervised classification

Our classification goal has been to achieve a finer granularity of tweet relevance than just a binary classification into actionable and noise. The following set of four classes, of decreasing relevance, gave us at the same time a good accuracy and granularity:

Mosquito-focus: this is the most *directly actionable* class, including tweets that report sites that are or may be foci for Dengue mosquito, or sites that provide conducive

---

[3] Only tweets in the Portuguese language were considered in this study.

[4] We used the tagger from Apache OpenNLP 1.5 series (`http://opennlp.sourceforge.net/models-1.5/`), and the LemPORT Lemmatizer customised for Portuguese language vocabulary.

environments to mosquito breeding. This class accounts for about 16% of tweets in our test set.

Sickness: This is the second most informative class. These tweets represent cases of: (i) users suspecting or confirming they are sick or they are aware of somebody else who is sick, and (ii) users discussing disease symptoms. Note that previous work (Sect. 1.7) on tracking Aedes-related epidemic waves make no distinction between this and the previous class.

News: This class represents general news about Dengue, ie tweets that spread awareness, report on available preventive measures, inform about health campaigns, and report the number of Dengue cases in certain locations. These are stilll *indirectly actionable* and useful eg. to show emerging outbreak patterns in specific areas.

Joke: Finally, about 20% of the tweets in our sample contain a combination of jokes or sarcastic comments about Dengue epidemic. While we regard these as noise, their detection requires an understanding of sarcastic tone in short text, which is challenging as it uses the same general terms as those found in more informative content.

The training set of about 1,000 messages was annotated by three local experts independently, by taking the majority class for each instance, requiring about 100 hours over three refinement steps to resolve inconsistencies and ambiguities. The classes are fairly balanced: *News*: 333 (31%), *Joke*: 148 (14%), *Mosquito focus*: 257 (24%), and *Sickness*: 338 (31%). Classification performance, measured using standard cross-validation, was similar across different classifier models, namely SVM, Naive Bayes, and MaxEntropy. We chose Naive Bayes as having probabilities associated to each class assignment helped identify the weak assignments, and thus the potential ambiguities in the manual annotations.

The classifier reported an overall 84.4% accuracy and .83 F-measure. In order to further validate these results, we then sampled an additional set of 1,600 tweets, none of them used for training, and performed both automated classification and manual annotation on this set. On this new set, the distribution of instances in each class, taken from the ground truth annotations, is not substantially different from that in the training set, except for the more abundant *Mosquito focus* class: *News*: 404 (25%), *Joke*: 289 (18%), *Mosquito focus*: 253 (16%), and *Sickness*: 649 (41%). Performance results for this classifier are reported in Table 1.

Class	Precision	Recall	F	Accuracy
News	.79	.74	.76	.74
Joke	.63	.85	.72	.85
Mosquito focus	.79	.85	.83	.86
Sickness	.91	.78	.84	.78

Table 1. Classifier performance on independent test set

4 Unsupervised content clustering using LDA

As discussed earlier, supervised classification does not fully meet our requirements, as manual annotation limits the size of training set and makes it difficult to update the model when the characteristics of the epidemics changes. Also, finding a crisp, unambiguous classification has been problematic.

LDA-based clustering [BNJ03] has been used before for Twitter content analysis and topic discovery, for example by [MPLC13,DSL$^+$11,WLJH10]. What we investigate is an application of LDA that shows the potential for scalability and flexibility, i.e., by periodically rebuilding the clusters to track drift in Twitter search keywords.

For this experiment, our sample dataset consists of $107,376$ tweets, harvested in summer 2015 using standard keyword filtering from the Twitter feed, and containing a total of $17,191$ unique words. Raw tweets were pre-processed just like for classification (phase 2 in Fig. 1), producing a bag-of-words representation of each tweet. Additionally, as a further curation step we removed the 20 most frequent words in the dataset, as well as all words that do not recur in at least two tweets. This last step is needed to prevent very frequent terms from appearing in all topics, which reduces the effect of our cluster quality metrices and cluster intelligibility.

4.1 Evaluation of clustering quality

We explored a space of clustering schemes ranging from 2 to 8 clusters.[5] In the absence of an accepted gold standard, a number of evaluation methods have been proposed in the literature. For instance, [MPLC13] proposes to measure cluster quality by quantifying the differences caused in topic mining using two different stream sampling mechanisms. The method is based on the differences between the distribution of words across topics and between the two sampling mechanisms. However, it cannot be used in our setting, because our corpus of tweets is fixed, rather than a sample. Also, while any two individual words may have different frequency distributions, the approach does not necessarily take into account the importance, measured by relative frequency, of the words within the entire corpus. In an alternative approach, [DSL$^+$11] use ground truth in the form of pre-established hashtags. This is not applicable in our scenario, either, because by the way our topic filtering is done, most of the tweets in our corpus will already include a high number of hashtags, including for instance the #dengue hashtag.
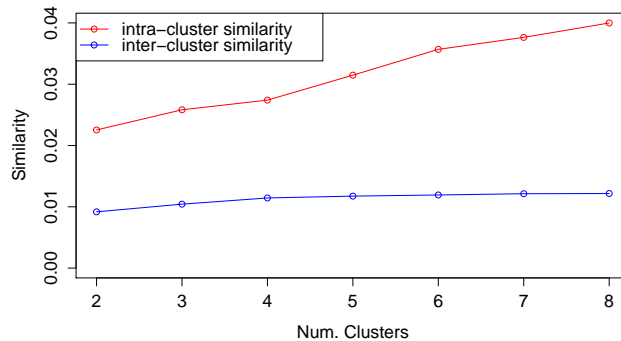
Instead, we propose to use *intra-* and *inter-* cluster similarity as our main evaluation criteria. This is inspired by *silhouettes* [Rou87], and based on the contrast between *tightness* (how similar data are to each other in a cluster) and *separation* (how dissimilar data are across clusters). Specifically, we define the similarity between two clusters C_a, C_b in terms of the cosine TF-IDF similarity of each pair of tweets they contain, i.e., $t_i \in C_a$ and $t_j \in C_b$, as follows:

$$sim(C_a, C_b) = \frac{1}{|C_a|\,|C_b|} \sum_{t_i \in C_a, t_j \in C_b} \frac{\mathbf{v}(t_i) \cdot \mathbf{v}(t_j)}{||\mathbf{v}(t_i)||\,||\mathbf{v}(t_j)||} \qquad (1)$$

---

[5] All experiments carried out using the Apache Spark LDA package `https://spark.apache.org`

where $\mathbf{v}(t_i)$ is the TF-IDF vector representation of a tweet. That is, the kth element of the vector, $t_i[k]$, is the TF-IDF score of the kth term. As a reminder, the TF-IDF score of a term quantifies the relative importance of a term within a corpus of documents [AZ12]. Eq. (1) defines the *inter-cluster similarity* between two clusters $C_a \neq C_b$, while the *intra-cluster similarity* of a cluster C is obtained by setting $C_a = C_b = C$.

Fig. 2 reports the inter- and intra-cluster similarity scores for each choice of clustering scheme. The absolute similarity numbers are small, due to the sparse nature of tweets and the overall little linguistic overlap within clusters. However, we can see that the intra-cluster similarity is more than twice the inter-cluster similarity, indicating good separation amongst the clusters across all configurations. This seems to confirm that the LDA approach is sufficiently sensitive to discover sub-topics of interest within an already focused general topic, defined by a set of keywords.
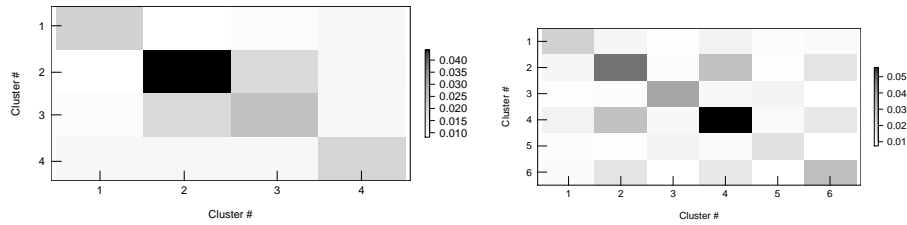

Fig. 2. Intra- and Inter-cluster similarities

The plots in Fig. 3 provide more detailed indication of the contrast between intra- and inter-cluster similarity at the level of detail of individual clusters. For example, in the 4-clusters case, the average of the diagonal values of the raster plot is the intra-cluster similarity reported in Fig. 2, whereas the average of the off-diagonal values represent the inter-cluster similarity. In these plots, darker boxes indicates higher (average) similarity. Thus, a plot where the diagonals are darker than the off-diagonal elements is an indication of a high quality clustering scheme.

Although the similarity metrics are objective and seems to confirm the good quality of the clustering, the plots in Fig. 2 and Fig. 3 do not provide much insight into the optimal number of clusters, or indeed their semantic interpretation. We therefore relied on our domain experts for the empirical selection of the clustering scheme (2,4,6,8 clusters) that would most closely lend itself to an intuitive semantic interpretation of the topics. Their assessment is reported below. In Sec. 4.3 we present a comparison of topics content using our four classes model as a frame of reference.

Fig. 3. Inter- and intra-similarity for 4 and 6 clusters topic models

4.2 Empirical topics interpretation

Expert inspection, carried out by native Brazilian Portuguese speakers, considered both the list of words within each topic, and a sample of the tweets for that topic. In this case, the most intelligible clustering scheme had 4 topics. The following is a list of most relevant topics for this scheme:

Topic 1: parado, água, fazer, vacina, até, meu, tão
Topic 2: combate, morte, sáude , confirma, ação , homem, chegar, queda, confirmado, agente
Topic 3: contra, suspeito, sáude , doença, bairro, morrer, combater, cidade, dizer, mutiro
Topic 4: mosquito, epidemia, pegar, foco, casa, hoje, mesmo, estado, igual

The importance of the words is given by LDA as a measure of how well they are represented in the topics.[6] Unsurprisingly, topic inspection suggests an interpretation that only partially overlaps with the a priori classification we have seen in the supervised case. Specifically, **Topic 1** is closely related to *Jokes*. Most of the tweets for this topic either make an analogy between Dengue and the users lives, or they use the words related to Dengue as a pun. A typical pattern is the following:

meu [algo como: wpp - WhatsApp, timeline, Facebook, twitter etc] está mais parado do que agua com dengue.
My [something like: wpp - WhatsApp, timeline, Facebook, twitter etc] is more still than standing water with dengue mosquito.

Specific examples include:

Aitizapi ta com dengue de tão parado (*Aitizapi is so still that it has been infected by dengue*)
Concessionária tá dando dengue de tão parada que tá (*Car dealership is so still that it has dengue*)

In the first example, the user was playing with the words when referring to the standing status and inactivity in his Whatsapp account. Breeding sites of the Aedes mosquito are mostly found in containers with standing water. In the second, the user is joking about significant decreases in car purchases due to the emerging economic crisis in Brazil. Many of the jokes in the last epidemic wave have been related to Zika, which in Braziliasn Portuguese, has been used as a new slang word for failure or any kind of personal problem.

---

[6] Some of the words are just noise. This is due to occasional imperfect lemmatisation during the preprocessing stage.

Topic 2 is interpreted as *news* about increase or decrease of Aedes-borne disease cases as well as specific cases of people who died because of the Aedes-borne diseases, i.e. Dengue, Chikungunya and Zika. It also contains news about the combat of the mosquito in certain locations as well. Examples:

> Rio Preto registra mais de 11 mil casos de dengue e 10 mortes no ano #SP
> *Rio Preto reports more than 11 thousand cases of dengue in the year #SP*
>
> 543 casos estão em análise - Londrina confirma mais de 2,5 mil casos de dengue em 2015 - [URL removed]
> *543 cases of dengue are under analysis - Londrina confirms more than 2.5 cases of dengue in 2015 - [URL removed]*

Topic 3 appears to contain mostly *news about campaigns* or actions to combat or to prevent Aedes-borne diseases, for instance:

> Curcuma contra dengue [URL removed] (*Curcuma against dengue*)
>
> Prefeitura de Carapicuba realiza nova campanha contra dengue e CHIKUNGUNYA[URL removed]
> *Carapicuba City Hall launches new campaign against dengue and CHIKUNGUNYA[URL removed]*

The difference between the news in topics 2 and 3 concerns the type of news, which for topic 2 is mostly about the increase or decrease of Aedes-borne diseases, whereas in topic 3 is about campaigns or actions to combat the propagation of the Aedes mosquito.

Finally, **Topic 4** contains mostly *sickness* tweets, with some instances of *jokes*:

> Será que eu to com dengue ? (*I wonder: do I have dengue?*)

4.3 Classes vs clusters

The point to note in the assessment above is that the most relevant tweets, those corresponding to the *Mosquito Focus* class, are not easily spotted, in particular they do no seem characterise any of the topics. Intuitively, this can be explained in terms of the relative scarcity of these tweets within the stream, combined with the balancing across topics that occurs within LDA.

In order to quantify this intuition, we have analysed the topics content using our predefined four classes as a frame of reference. In this analysis, we have used our trained classifier to predict the class labels of all the tweets in the corpus that we used to generate the topics (about 100,000). We then counted the proportion of class labels in each topic, as well as, for each class, the scattering of the class labels across the topics. The results are presented in Table 2 and Table 3, respectively, where the dominant entries for each column (resp row) are emphasised. It is worth remembering that these results are based on predicted class labels and are therefore inherently subject to the classifier's inaccuracy. Furthermore, the predicted class labels were *not* available to experts when they inspected topic content, thus they effectively performed a new manual classification on a content sample for each topic. Despite the inaccuracies introduced by these elements, Table 2 seems to corroborate the experts' assessment regarding topics 1 and 2, but less so for topics 3 and 4. This may be due to the sampling operated by the experts, which selected content towards the top of the topic (LDA ranks content by relevance within a topic) and may have

	Topic 1	Topic 2	Topic 3	Topic 4
News	13.9	**72.6**	27.2	**39.4**
Joke	**39.5**	0.1	2.8	4.1
Mosquito Focus	30	4.0	12.3	12.5
Sickness	16.6	23.3	**57.7**	**44.0**
Total	100	100	100	100

Table 2. Distribution (%) of predicted class labels within each cluster

	Topic 1	Topic 2	Topic 3	Topic 4	Total
News	29.1	28.5	8.9	33.5	100
Joke	**95.0**	0.03	1.05	4.0	100
Mosquito Focus	**79.5**	2.0	5.1	13.4	100
Sickness	34.8	9.1	18.8	**37.3**	100

Table 3. Scattering (%) of predicted class labels across clusters

come across joke entries which are otherwise scarce in topic 4. Although the heavy concentration on joke tweets in topic 1 from Table 2 seems promising (i.e., the other topics are relatively noise-free), Table 3 shows a problem, namely that topic 1 is also where the vast majority of *Mosquito Focus* tweets are found. Thus, although topic 1 segregates the most informative tweets well, it is also very noisy, as these tweets are relatively scarce within the entire corpus.

The analysis just described suggests that topic modelling offers less control over the content of topics, compared to a traditional classifier, especially on a naturally noisy media channel. Although relevant content can be ascribed to specific topics, these are polluted by noise. Despite this, LDA performs relatively well on creating sub-topics from a sample that is already focused on a specific topic, such as conversations on the Aedes-transmitted viruses. The main appeal of the classifier is that it makes it straightforward to select relevant content, with acceptable experimental accuracy. In our follow on research we are investigating ways to combine the benefits of the two approaches. Specifically, we are studying a unified semi-supervised model where topic modelling can be used to improve the accuracy of the classifier, i.e., by automatically expanding the training set, and to alleviate the cost of re-training at the same time.

5 Summary

In this paper we discussed methods for detecting relevant content in a Twitter stream that has been pre-filtered to focus on a specific topic, in this instance online discussions around Dengue and other Aedes-borne diseases in Brazil. Relevance is defined operationally in terms of four classes within the broad topic of Dengue. When reliably segregated from noise, relevant content can be used in multiple ways in the context of health vigilance to combat epidemics caused by the Aedes mosquito. We have compared two approaches for detecting relevance, supervised classification and clustering by topic modelling. Our experimental results indicate that the clusters produced using topic modelling tend to be noisy, perhaps because LDA is not very effective on text content that is pre-filtered for a specific set of keywords. Supervised classification, on the other hand, is costly as manual annotation requires multiple rounds due to ambiguities in the content, but is more appealing as a good proportion of actionable messages are segregated, i.e., in the two most relevant classes. We are currently exploring ways to combine the two approaches into one semi-supervised model, i.e., by exploiting the topics to enhance the training set and alleviate the cost of re-training.

Acknowledgments

References

[ALG$^+$11] H. Achrekar, R. Lazarus, A. Gandhe, S Yu, and B. Liu. Predicting Flu Trends using Twitter Data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707. IEEE, 2011.

[AZ12] Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. In *Mining Text Data*, pages 77–128. Springer, 2012.

[BNG11] Hila Becker, Mor Naaman, and Luis Gravano. Beyond Trending Topics: Real-World Event Identification on Twitter. *Procs. ICWSM*, pages 1–17, 2011.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, mar 2003.

[CDC15] CDC. Centers for Disease Control and Prevention. `http://www.cdc.gov/dengue/`, 2015. [Online; accessed 15-december-2015].

[CW14] Tao Cheng and Thomas Wicks. Event detection using Twitter: a spatio-temporal approach. *PloS one*, 9(6):e97807, jan 2014.

[DSL$^+$11] K. Dela Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking. Topical Clustering of Tweets. *SIGIR 3rd Workshop on Social Web Search and Mining*, 2011.

[GBH09] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12):12, 2009.

[GVM$^+$11] J. Gomide, Adriano Veloso, W. Meira, V. Almeida, F. Benevenuto, Fernanda Ferraz, and M. Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. *Proceedings of the ACM WebSci'11, June 14-17 2011, Koblenz, Germany.*, pages 1–8, 2011.

[LC10] Vasileios Lampos and Nello Cristianini. Tracking the flu pandemic by monitoring the social web. *2010 2nd International Workshop on Cognitive Information Processing, CIP2010*, pages 411–416, 2010.

[MPLC13] Fred Morstatter, J Pfeffer, H Liu, and Km Carley. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of ICWSM*, pages 400–408, 2013.

[NGS$^+$09] M. Nagarajan, K. Gomadam, Amit P. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. *LNCS*, 5802 LNCS:539–553, 2009.

[PR15] PUC-Rio. Efficient Monitoring of Aedes Mosquito in Brazil. vazadengue.inf.puc-rio.br/, 2015. [Online; accessed 15-december-2015].

[RDL10] Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing Microblogs with Topic Models. *ICWSM*, 10:1, 2010.

[REC12] Alan Ritter, Oren Etzioni, and Sam Clark. Open domain event extraction from twitter. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, page 1104, 2012.

[Rou87] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comp. Applied Math.*, 20:53 – 65, 1987.

[SOM10]   T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Procs. WWW '10*, page 851, 2010.

[WLJH10] J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Procs. WSDM '10*, pages 261–270. ACM, 2010.